State space modelling of extreme values with particle filters

David Peter Wyncoll, MMath

Submitted for the degree of Doctor of Philosophy

at Lancaster University

June 2009

State space modelling of extreme values with particle filters

David Peter Wyncoll, MMath

Submitted for the degree of Doctor of Philosophy

at Lancaster University, June 2009

Abstract

State space models are a flexible class of Bayesian model that can be used to smoothly capture non-stationarity. Observations are assumed independent given a latent state process so that their distribution can change gradually over time. Sequential Monte Carlo methods known as particle filters provide an approach to inference for such models whereby observations are added to the fit sequentially. Though originally developed for on-line inference, particle filters, along with related particle smoothers, often provide the best approach for off-line inference.

This thesis develops new results for particle filtering and in particular develops a new particle smoother that has a computational complexity that is linear in the number of Monte Carlo samples. This compares favourably with the quadratic complexity of most of its competitors resulting in greater accuracy within a given time frame.

The statistical analysis of extremes is important in many fields where the largest or smallest values have the biggest effect. Accurate assessments of the likelihood of extreme events are crucial to judging how severe they could be. While the extreme values of a stationary time series are well understood, datasets of extremes often contain varying degrees of non-stationarity. How best to extend standard extreme

value models to account for non-stationary series is a topic of ongoing research.

The thesis develops inference methods for extreme values of univariate and multivariate non-stationary processes using state space models fitted using particle methods. Though this approach has been considered previously in the univariate case, we identify problems with the existing method and provide solutions and extensions to it. The application of the methodology is illustrated through the analysis of a series of world class athletics running times, extreme temperatures at a site in the Antarctic, and sea-level extremes on the east coast of England.

Acknowledgements

I would first like to thank my supervisors Jonathan Tawn and Paul Fearnhead for their help and guidance throughout my PhD. I also acknowledge the assistance of staff and students at Lancaster University, in particular Granville Tunnicliffe-Wilson, Emma Eastoe and Chris Jewell.

I am grateful to the Natural Environment Research Council (NERC) for funding my studies at Lancaster.

I thank Dougal Goodman from the Foundation for Science and Technology for the Antarctic temperature dataset used in Section 4.2. I also acknowledge the British Oceanographic Data Centre, who are funded by the Environment Agency and NERC, for the sea-level data used in Section 5.2.

Finally, I would like to thank my wife Robyn for her support and encouragement over the last three and a half years.

Declaration

The work in this thesis is my own, as carried out under the supervision of Jonathan Tawn and Paul Fearnhead. It has not been submitted for the award of a higher degree elsewhere.

Section 3.2 has been submitted for publication as Fearnhead et al. (2009) along with the simplest EM algorithm in Subsection 3.3.1, the athletics analysis of Section 4.1 as well as corresponding material from Chapter 2 and Appendix A.

David Wyncoll

Contents

List of figures ix

List of tables xii

List of algorithms xiii

1 Introduction **1**

2 Literature review **5**

2.1 Particle Filtering 5

2.1.1 State space model 5

2.1.2 Kalman Filter 7

2.1.3 Basic particle filters 9

2.1.4 Auxiliary particle filter 14

2.1.5 Enhancements 16

2.1.6 Parameter estimation 18

2.2 Particle Smoothing 21

2.2.1 Kalman Smoothing 21

2.2.2 Smoothing while filtering 22

2.2.3 Forwards-backwards smoothers 24

2.2.4 Two-Filter Smoother 25

2.3 Univariate Extreme Value Theory 28

2.3.1 Maxima of IID random variables 28

2.3.2 Point process characterisation 34

2.3.3 Dependent processes 39

2.3.4 Non-stationary processes 43
2.4 Multivariate Extreme Value Theory 47
2.4.1 Multivariate extreme value distributions 47
2.4.2 Alternative approaches 51

3 New results in particle filtering 53
3.1 Choosing when to re-sample 53
3.1.1 Motivation 53
3.1.2 Theory 54
3.1.3 Simulation studies 60
3.1.4 Conclusion 74
3.2 A sequential smoothing algorithm with linear computational cost . 75
3.2.1 Weaknesses of current particle smoothers 75
3.2.2 New smoothing algorithm 77
3.2.3 Simulation studies 83
3.3 EM algorithm for static parameter estimates 93
3.3.1 EM algorithm 93
3.3.2 Estimating observed information 97

4 State space modelling of univariate extremes 98
4.1 Analysis of women's 3000m running event 98
4.1.1 Introduction 98
4.1.2 Dynamic r-smallest order statistics model 101
4.1.3 Parameter estimation 103
4.1.4 Results 104
4.2 Analysis of Antarctic temperature data 107
4.2.1 Introduction 107
4.2.2 Standardising the dataset 108
4.2.3 Dynamic point process model 112
4.2.4 Results 115

5 State space modelling of bivariate extremes 120

5.1 Pooling athletics data from two events 120

5.1.1 Introduction 120

5.1.2 Dynamic logistic model with correlated random walks 121

5.1.3 Parameter estimation 125

5.1.4 Results 127

5.2 Joint analysis of sea-level data 129

5.2.1 Introduction 129

5.2.2 Dynamic logistic model with variable dependence parameter 132

5.2.3 Parameter estimation 136

5.2.4 Results 138

A Implementation of particle methods 143

A.1 Linear-Gaussian model 143

A.2 Stochastic volatility 145

A.3 Bearings-only tracking 149

A.4 Analysis of women's 3000m running event 150

A.5 Analysis of Antarctic temperature data 154

A.6 Pooling athletics data from two events 156

A.7 Joint analysis of sea-level data 160

B Derivation of state models from SDEs 166

B.1 Integrated random walk 166

B.2 Correlated integrated random walks 169

C Derivation of bivariate logistic model 171

C.1 Joint distribution functions in Fréchet margins 171

C.2 Joint density functions in Fréchet margins 176

C.3 Transforming densities to GEV margins 177

Bibliography 179

Index **190**

List of Figures

2.1 General form of the state space model showing the conditional independences that exist between the observations Y_t and the hidden states X_t. 6

2.2 Diagram showing how the simple Filter-Smoother re-weights the filter particles. 23

2.3 Probability density functions of the Negative-Weibull, Gumbel and Fréchet distributions. 30

2.4 Example of a series declustered with the runs method. 41

2.5 Example of the selection of peaks over a constant threshold when the series is non-stationary. 44

3.1 Log variances of filter estimates of the state for different ESS thresholds. 54

3.2 Log variances of filter estimates of $\mathbf{E}(X_t|y_{1:t})$ in the AR(1) model for different ESS thresholds. 62

3.3 Log variances of filter estimates of the volatility in the stochastic volatility model for different ESS thresholds. 67

3.4 Simulated path of an object in the u-v plane over 24 time steps with the observed noisy bearings made from the origin. 70

3.5 Log variances of filter estimates of the range in the bearings-only tracking model for different ESS thresholds. 71

3.6 Average effective sample size for each of the 200 time steps using the filter, Forward-Backward and Two-Filter smoothers as well as the $\mathcal{O}(N^2)$ and $\mathcal{O}(N)$ versions of our new algorithm. 86

3.7 Average effective sample sizes as in Figure 3.6 with different ratios of the state noise ν^2 to the observation noise τ^2. 87

4.1 The μ_t and ξ_t components of 1000 particles from the filter at years 1973, 1979, 1983 and 1992 following the method of Gaetan and Grigoletto (2004). 100

4.2 Five fastest times for the women's 3000m race between 1972 and 2007 with Wang Junxia's time in 1993. 102

4.3 Raw Faraday temperature series showing annual maxima, mean and minima with linear least-squares fits. 107

4.4 Logged sum of squares of the discrepancy between the data removed for cross-validation and its smoothed estimate for a variety of kernel bandwidths. 110

4.5 Smoothed Faraday temperature series with confidence interval constructed from the smoothed standard deviation. 110

4.6 Standardised Faraday temperature series for 1958 with declustered threshold exceedances for the upper and lower tail. 111

4.7 Standardised Faraday temperature series with declustered threshold exceedances and fitted trend μ_t for the upper and lower extremes. . 116

5.1 Five fastest times for the women's 1500m and 3000m races between 1972 and 2007. 122

5.2 Contour plot of the model log likelihood for a range of dependence parameters ρ and α. 126

5.3 Map showing the locations of our sea-level data sources along the eastern coastline of England. 130

5.4 Monthly maximum sea-level surges for each of the three locations. 131

5.5 Sea-level surges at Immingham during the winter containing January 1989 with the five largest monthly cluster maximum obtained using blocks of radius $\kappa = 7$ days. 133

5.6 Contour plot of the model log likelihood for a range of dependence parameters ρ_μ and ρ_α with data from Immingham and Lowestoft. . 138

5.7 Smooth joint fit for Immingham and Lowestoft of the two largest cluster maxima during each of the winter months. 140

5.8 Smooth joint fit for Immingham and Sheerness as in Figure 5.7. . . 141

C.1 Decomposition of the bivariate point process event that generates $G(x_2, x_1, y_2, y_1)$ in terms of the possible relative positions of the componentwise maxima $M_{n,X}^{(1)}/n$ and $M_{n,Y}^{(1)}/n$ to the constants $x_2 \leq x_1$ and $y_2 \leq y_1$. 173

List of Tables

2.1 Properties of the Generalised Extreme Value distribution. 31

3.1 Average variances of filter estimates of the current state, upper 2.5% quantile and probability of exceeding the true state for 100 time steps using various ESS thresholds. 63

3.2 Average variances of filter estimates of the current state over 100 time steps with $\nu^2 = 100$ and $\nu^2 = 1/100$. 63

3.3 Average variances of filter estimates of the state and volatility over 100 time steps. 66

3.4 Average variances of filter estimates of the range over 24 time steps using various ESS thresholds. 73

3.5 Number of particles used and average run time of each algorithm. . 85

3.6 Comparison of the Filter-Smoother with two variations of our new $\mathcal{O}(N)$ algorithm. 90

3.7 Comparison of the Filter-Smoother with our new $\mathcal{O}(N)$ algorithm when different block sizes are used. 91

4.1 Model likelihood with σ and ξ estimates for different values of the smoothing parameter ν^2 and $r = 2$. 103

5.1 Marginal parameter estimates obtained for each site from an EM algorithm. 137

List of Algorithms

2.1 Kalman Filter. 8
2.2 Sampling Importance Re-sampling filter. 12
2.3 Auxiliary particle filter. 15

3.1 New $\mathcal{O}(N^2)$ smoothing algorithm. 79
3.2 New $\mathcal{O}(N)$ smoothing algorithm. 81
3.3 New $\mathcal{O}(N)$ block smoothing algorithm. 83
3.4 EM algorithm for fixed parameters in the observation density. 94
3.5 EM algorithm for fixed parameters in the state and observation densities. 96

Chapter 1

Introduction

While much of statistics is concerned with typical behaviour, it is often the most extreme values that have the biggest impact. Examples range from flooding to financial crashes, hurricanes and world records. Quantifying how unlikely these events are is key to predicting when they might happen again as well as how severe or remarkable they could be.

To answer these questions and more, extreme value theory provides a collection of models and modelling approaches for analysing the largest or smallest values of a dataset. The models are typically justified by asymptotic theory that considers the distribution of the most extreme values of an infinite sample. By assuming these to hold for a finite sample, approximate models for the extreme values of a series can be obtained while the distribution of the underlying sample remains unknown.

Since extreme values by definition occur infrequently, datasets of extreme values are often collected over a period of time so arise as a time series. While this gives an opportunity for the most extreme events to be observed, it often adds difficulties to modelling as the distribution of extreme values may change over time. The extremes of independent or stationary sequences are fairly well understood,

particularly of univariate series. However, the extreme values of non-stationary sequences are less understood with many recent papers proposing alternative models to capture the non-stationarity.

State space models have received much recent attention as a flexible class of non-linear models for general time series. Observations are assumed to be conditionally independent given a hidden state process which captures the non-linearity and non-stationarity present in the series. A Bayesian model is typically constructed where inference about the observations is obtained by integrating out the hidden process.

As with most complex Bayesian models, the integrals required to apply the model are intractable in general. Monte Carlo methods are typically used to overcome these intractabilities by sampling possible state values and parameters from the model to approximate the integrals with finite sums. For this, Markov Chain Monte Carlo (MCMC) is often used to sample from the joint distribution of the entire hidden state process given all the observations. However, MCMC often struggles with this as the states are frequently highly correlated and the dimension of the sample space grows with the number of observations.

Particle filters and related sequential Monte Carlo methods provide an alternative class of algorithms for fitting state space models. Originally developed to estimate the current value of the state on-line as observations arrive, particle methods are being increasingly applied off-line as an alternative to MCMC. By incorporating observations into the fit sequentially, particle methods sample one state at a time rather than sampling the whole process at once thus reducing the dimension of each sample.

In this thesis we use state space models to capture non-stationarity in extreme value time series. This allows smooth non-linear trends to be incorporated into an extreme value analysis through a Bayesian model that accounts naturally for the uncertainties involved. As well as models for the extremes of a univariate series, we also consider bivariate models which allow the dependencies between the extremes

of two variables to be studied.

We fit the models with sequential particle algorithms that we tailor to modelling extreme values. The use of sequential algorithms allows long datasets in particular to be fitted although this need not be the case. Through the course of improving particle methods for the modelling of extremes, we provide many new results that can be used with more general state space models.

The thesis is structured as follows. In Chapter 2 we review relevant literature within the fields of particle filtering and extreme value theory. We begin by defining the state space model before introducing basic particle filters that can be used for their inference. After listing their flaws we describe many enhancements to the basic filter that enable it to be applied more efficiently. We also review methods of estimating parameters in the model as well as extensions of the particle filter to perform smoothing, both of which will be useful when it comes to modelling extremes.

Following from Section 2.3 we present results from univariate extreme value theory. We start by considering the extremes of IID random variables before extending the results to dependent and non-stationary sequences. We then provide a similar review of multivariate extreme value theory which considers the extremes of components of multivariate variables as well as the dependencies between them.

In Chapter 3 we provide new results in particle filtering, focusing on aspects of particle methods that will aid our subsequent application to the analysis of extreme values. Section 3.1 begins with a study on re-sampling, an important step in most particle filters, with a new method for selecting when one should or should not re-sample. In Section 3.2 we present a new particle smoothing algorithm with linear computational complexity that compares favourably with the quadratic complexity of most particle smoothers. Finally, in Section 3.3 we construct an Expectation-Maximisation (EM) algorithm that uses our new smoothing algorithm to estimate fixed parameters in the model.

We begin our analysis of extreme values using state space models in Chapter 4. In this chapter we consider the extreme values of univariate time series, presenting our models through a couple of example analyses. Our first example looks at the women's 3000m running event with the aim of estimating the probability of an extreme world record. A state space model is used to smoothly account for the clear non-linear trend present in the series of historical annual records. For our second example we study a series of daily temperature measurements from the Antarctic peninsula made over a period of 44 years. Examining the upper and lower extremes of the series requires us to account for the dependence between neighbouring measurements as well as the separation of the trend in the bulk of the data to that in the extremes.

In our final chapter we jointly model the extreme values of a pair of related series. This allows us to study the dependence between the extreme values of one variable to the other as well as the connection between the non-stationary trends in each component. We begin by extending the women's 3000m analysis by drawing connections with the women's 1500m race. By exploiting these connections we can obtain a better estimate of our target tail probability. We conclude the thesis with a bivariate analysis of sea-levels at pairs of sites along the eastern coast of England. By making use of the state space approach, we allow the extremal dependence between two sites to vary smoothly over time to ask whether such a change is significant.

Chapter 2

Literature review

2.1 Particle Filtering

In this section we introduce the state space model used throughout this thesis and describe the filtering problem. We review the limitations of the Kalman Filter before describing particle filtering in detail. See Doucet et al. (2001) for a good introduction to particle filters and their applications.

2.1.1 State space model

State space models provide a flexible framework to handle non-linear time series. These models assume a time-series with observations Y_t that are conditionally independent given a hidden process X_t. We assume throughout this thesis that X_t is a real-valued Markov chain. Formally the model is given by a state equation and an observation equation, which can be represented in terms of conditional

distributions

$$X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} \sim f_t(\cdot|x_t), \tag{2.1}$$

$$Y_t|\{X_{1:t} = x_{1:t}, Y_{1:t-1} = y_{1:t-1}\} \sim g_t(\cdot|x_t), \tag{2.2}$$

where we use the notation that $x_{1:t} = (x_1, \ldots, x_t)$, and similarly for $y_{1:t}$. The state X_t and observations Y_t may both be multidimensional and the state and observation densities f_t and g_t are typically arbitrary.

We assume for simplicity that the observations are observed at times $1, 2, \ldots$ although this is really just shorthand for observation times $t_1, t_2, \ldots$. We will usually assume the state and observation densities are constant over time and remove t from the notation. We will be working with a Bayesian state space model which is completed through specifying a prior distribution $\pi(x_0)$ for X_0. Alternatively, some authors specify the prior for X_1. A graph representing the model is shown in Figure 2.1.

The Markov form of the state (2.1) means that if X_t is known at time t, neither the history of the state nor the sequence of currently available observations provide any additional information about the future of the series. Similarly, no additional information about Y_t is obtainable if X_t is known. Some sequential Monte Carlo algorithms introduced in later sections work with more general models than this but these are not considered here.

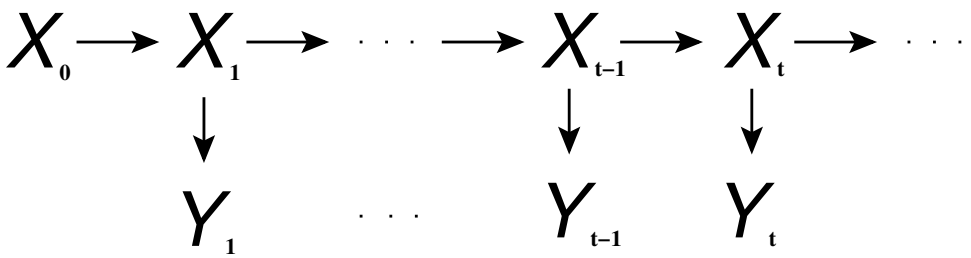


Figure 2.1: General form of the state space model showing the conditional independences that exist between the observations Y_t and the hidden states X_t.

2.1.2 Kalman Filter

When the observations are arriving sequentially we are often interested in the current value of the state X_t given all the available data. For this *filtering* problem, interest lies in estimating the posterior distribution $p(x_t|y_{1:t})$. This can, in principle, be calculated recursively using

$$\begin{aligned} p(x_t|y_{1:t}) &\propto g(y_t|x_t)\, p(x_t|y_{1:t-1}) \\ &= g(y_t|x_t) \int f(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1})\, \mathrm{d}x_{t-1}. \end{aligned} \tag{2.3}$$

Thus the filtering density is obtained, up to proportionality, by multiplying the likelihood $g(y_t|x_t)$ by the one-step *prediction density* $p(x_t|y_{1:t-1})$. However, for most models, solving (2.3) analytically is impossible.

An important exception of this is the Kalman Filter of Kalman (1960). Closed form expressions for the filtering density exist if both the state density $f(x_t|x_{t-1})$ and the observation density $g(y_t|x_t)$ are linear-Gaussian and the prior is also Gaussian. Put simply, the state space model can have the following form:

$$\begin{aligned} X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} &\sim \mathcal{N}(Fx_t, Q), \\ Y_t|\{X_{1:t} = x_{1:t}, Y_{1:t-1} = y_{1:t-1}\} &\sim \mathcal{N}(Gx_t, R), \\ X_0 &\sim \mathcal{N}(\mu_0, \Sigma_0), \end{aligned} \tag{2.4}$$

where both the dimensions d_f of the state vector X_t and d_g of the observation vector Y_t are arbitrary. The matrices F $(d_f \times d_f)$, Q $(d_f \times d_f)$, G $(d_g \times d_f)$, R $(d_g \times d_g)$, Σ_0 $(d_f \times d_f)$ and vector μ_0 (d_f) are all arbitrary except that, as covariance matrices, Q, R and Σ_0 must be positive semi-definite. When these conditions are met, the filter densities are all Gaussian with means and covariance matrices that can be calculated using only matrix operations.

The Kalman Filter equations for recursively calculating $p(x_t|y_{1:t}) \sim \mathcal{N}(\mu_t, \Sigma_t)$ are

Algorithm 2.1: Kalman Filter.

For $t = 1, 2, \ldots$, assume the filter at time $t-1$ is $\mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$. Then

1. **Predict:** Set $\mu_{t|t-1} = F\mu_{t-1}$ and $\Sigma_{t|t-1} = F\Sigma_{t-1}F' + Q$.
2. **Update:** Set $z_t = y_t - G\mu_{t|t-1}$, $S_t = G\Sigma_{t|t-1}G' + R$ and $K_t = \Sigma_{t|t-1}G'S_t^{-1}$ and then the filter at time t is $\mathcal{N}(\mu_t, \Sigma_t)$ with

$$\mu_t = \mu_{t|t-1} + K_t z_t \quad \text{and} \quad \Sigma_t = (I - K_t G)\Sigma_{t|t-1}.$$

given in Algorithm 2.1. The first step of each iteration produces a vector $\mu_{t|t-1}$ and matrix $\Sigma_{t|t-1}$ which are the mean and variance of the prediction density $p(x_t|y_{1:t-1})$. If the observation y_t is missing, this density is also the filter density since it uses all the information available at time t. It is therefore very easy to account for missing data with the Kalman Filter. It is also easy to make predictions of future states by repeating the prediction step k more times to give $p(x_{t+k}|y_{1:t})$.

Extensions

The biggest disadvantage of the Kalman Filter is its reliance on the linear-Gaussian forms of the state and observation densities. Many methods have been proposed to extend the Kalman Filter to non-linear and non-Gaussian cases.

The *Extended Kalman Filter* (EKF) is a commonly used alternative to the Kalman Filter which allows the state and observation equations to be non-linear while still assuming Gaussian distributions throughout. It works by linearising the model with first-order Taylor approximations so that the standard Kalman Filter equations can be used. If the model equations are highly non-linear the approximations made will lead to error which can cause the filter to diverge. See Jazwinski (1973) and Anderson and Moore (1979) for details.

Other extensions include the *Gaussian sum filter* of Alspach and Sorenson (1972) which allows non-Gaussian distributions by approximating them by Gaussian mixtures. The *Unscented Kalman Filter* (UKF) of Julier et al. (1995) aims to im-

prove on the EKF when the state and observation equations are highly non-linear. Finally, the *Ensemble Kalman Filter* (EnKF) of Evensen (1994) propagates an ensemble of state vectors to avoid calculating the covariance matrix. This is particularly useful when the dimension of X_t is large in which case the Kalman Filter equations can be slow.

While these extensions all aim to modify the Kalman Filter to allow non-linearity or non-Gaussianity, they all do so by making approximations which only hold true when the model is linear-Gaussian. They can therefore fail for models that are far removed from linear-Gaussian by giving bad estimates of the position as well as the covariances.

2.1.3 Basic particle filters

We would like to extend the Kalman Filter to arbitrary f and g without making linear-Gaussian approximations. Since arbitrary models no longer give Gaussian filtering densities, a more thorough approach is achieved by targeting the whole pdf rather than just the mean and covariances. Since the calculation of these densities is intractable in general, we resort to Monte Carlo methods. They have the potential of approximating the target densities with errors that can be made negligible by increasing the Monte Carlo sample size.

For general Bayesian models, *Markov Chain Monte Carlo* (MCMC) is widely used to draw approximate samples from complex distributions. Since the target density is typically required in closed form (up to proportionality), standard MCMC methods cannot be used to sample from $p(x_t|y_{1:t})$ whose form is unknown. However, MCMC methods may be applied to sample draws from the joint *smoothing* distribution

$$p(x_{0:t}|y_{1:t}) \propto \pi(x_0) \prod_{s=1}^{t} f(x_s|x_{s-1})\, g(y_s|x_s) \tag{2.5}$$

which then marginally give the filter distribution.

While in principle this provides a solution to extending the Kalman Filter to arbitrary densities, it is often inappropriate, especially when the observations are arriving sequentially. In this case we will typically wish to estimate $p(x_t|y_{1:t})$ as the data arrive but (2.5) provides no simple way to recursively update draws using MCMC alone. Berzuini et al. (1997) and others show how MCMC draws may be updated using alternative Monte Carlo methods such as importance sampling to approximate the filter densities for a few time steps. However, these methods often deteriorate with time and it becomes necessary to rejuvenate the sample by rerunning MCMC from scratch. To do this, the whole path $X_{0:t}$ must be sampled and so the complexity of this step increases with t making MCMC impractical for sequential inference.

Sampling Importance Re-sampling filter

Sequential Monte Carlo algorithms, known generically as *particle filters*, have been proposed to overcome the restrictions imposed by the Kalman Filter. While Handschin and Mayne (1969) and Handschin (1970) were the first to use Monte Carlo methods for non-linear filtering, their methods only estimate the mean and covariance of the filtering density $p(x_t|y_{1:t})$. Gordon et al. (1993) were the first to propose the following method which allows the state density $f(x_t|x_{t-1})$ and the likelihood $g(y_t|x_t)$ to be non-linear and non-Gaussian. Rather than approximating the filter distributions as Gaussian they use a swarm of possible draws to represent the intractable densities. These *particles* are then updated sequentially as the new observations arrive. The resulting algorithm, which was independently proposed by Kitagawa (1996), is known as the *Bayesian bootstrap* or the *Sampling Importance Re-sampling* (SIR) filter.

The basic idea is to represent all densities involving the state X_t by the discrete

distribution made of Monte Carlo samples $\{x_t^{(i)}\}_{i=1}^N$

$$p(x_t) \simeq \frac{1}{N} \sum_{i=1}^{N} \delta(x_t - x_t^{(i)}),$$

where $\delta(\cdot)$ is the Dirac delta function. Integrals made with respect to this density are then approximated by sample means of the form

$$\mathbf{E}_p(h(X_t)) = \int h(x_t)\, p(x_t)\, \mathrm{d}x_t \simeq \frac{1}{N} \sum_{i=1}^{N} h(x_t^{(i)}). \tag{2.6}$$

This means that properties of the distribution such as the expected value are approximated by that of the sample. It can be shown using the weak law of large numbers that these approximations become exact as $N \to \infty$.

The SIR filter has three stages for each time step t. If we assume we enter step t with a sample $\{x_{t-1}^{(i)}\}$ approximating $p(x_{t-1}|y_{1:t-1})$, we first create a new sample $\{\tilde{x}_t^{(i)}\}$ which approximates the one-step predictive distribution $p(x_t|y_{1:t-1})$. This is achieved by sampling a new particle $\tilde{x}_t^{(i)}$ from the state density $f(x_t|x_{t-1}^{(i)})$ once for each i. If the observation y_t is missing, these particles represent the filter distribution for this time step and we can move to step $t+1$.

If, however, the new observation y_t is available, we next use it to weight each of our new particles with the likelihood $g(y_t|\tilde{x}_t^{(i)})$. This can be shown to be the appropriate importance weight for approximating $p(x_t|y_{1:t})$ by a sample from $p(x_t|y_{1:t-1})$. We finally use these weights as probabilities with which to re-sample the predictive particles. The re-sampled draws $\{x_t^{(i)}\}$ then represent a sample approximating our target distribution $p(x_t|y_{1:t})$.

The algorithm is summarised in Algorithm 2.2. Its computational complexity is $\mathcal{O}(N)$ where N is the number of particles. The easiest way to initialise the method is to begin with a sample from the prior one step before the first observation as this can be thought of as a filter distribution too. Alternatively, we can begin with a sample from $p(x_1)$ if this is possible and go straight to the weight stage of the

Algorithm 2.2: Sampling Importance Re-sampling filter.

1. **Initialisation:** Sample $\{x_0^{(i)}\}$ from the prior $\pi(x_0)$.
2. For $t = 1, 2, \ldots$
   (a) **Predict:** Sample $\tilde{x}_t^{(i)} \sim f(.|x_{t-1}^{(i)})$ once for each i.
   (b) **Weight:** Assign each particle $\tilde{x}_t^{(i)}$ the weight
   $$w_t^{(i)} \propto g(y_t|\tilde{x}_t^{(i)})$$
   and normalise them to sum to 1.
   (c) **Re-sample:** Sample x_t N times from the discrete distribution with support $\{\tilde{x}_t^{(i)}\}$ and probability masses $\{w_t^{(i)}\}$.

algorithm.

Unlike the Kalman Filter, the SIR filter has very few requirements to satisfy for it to be applied. It requires only that

1. the prior $\pi(x_0)$ (or $p(x_1)$) can be sampled from,
2. the state distribution $f(x_t|x_{t-1})$ can be sampled from,
3. the form of the likelihood $g(y_t|x_t)$ is known up to a normalising constant.

While the algorithm is simple in design, it can perform poorly over time as Gordon et al. (1993) acknowledge. This is especially true when there is a big difference between the significant regions of the state space in $p(x_t|y_{1:t-1})$ and $g(y_t|x_t)$. When this is the case the weights $w_t^{(i)}$ will be very uneven and few of the predictive particles $\tilde{x}_t^{(i)}$ will be re-sampled, effectively wasting the others and reducing the effective sample size.

Weighted particles

While the SIR filter uses re-sampling to create a set of unweighted particles, many sequential Monte Carlo algorithms which output weighted samples have been pro-

posed. Each Monte Carlo sample $x_t^{(i)}$ is given a weight $w_t^{(i)}$ that becomes the probability mass placed on the sample in the discrete approximation of the density. Integrals made with respect to this density are then approximated analogously to (2.6) by weighted means of the form

$$\mathbf{E}_p(h(X_t)) = \int h(x_t)p(x_t)\,\mathrm{d}x_t \simeq \sum_{i=1}^{N} h(x_t^{(i)})w_t^{(i)}.$$

One early example of this is the *sequential imputation* method of Kong et al. (1994). When applied to our state space model[1], their method amounts to sequentially sampling $x_t^{(i)} \sim p(x_t|x_{t-1}^{(i)}, y_t)$ and weighting with $w_t^{(i)} \propto p(y_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)}$, where the weights are normalised to sum to 1. Thus, rather than re-sampling each time step, the weights are incrementally updated using the previous step's weighted particles. For sequential imputation to be applied, however, the densities $p(x_t|x_{t-1}, y_t)$ and $p(y_t|x_{t-1})$ must be known, which is often not the case.

More generically, the *Sequential Importance Sampling* (SIS) method of Doucet (1998) and Liu and Chen (1998) allows an arbitrary sampling distribution to be chosen and then accounted for by the importance weight. If particles are propagated through the state equation, the SIS differs from the SIR filter only in the lack of re-sampling.

Re-sampling has its advantages and disadvantages. By incrementally updating the weights rather than re-sampling, the weights become increasingly uneven; this ultimately leads to all the mass being placed upon a single particle, wasting the others. Re-sampling the particles using the current weights as probabilities produces multiple copies of useful particles while losing ones with little or no weight. This allows future particles to be sampled from the useful ones giving a greater sample overall. However, since re-sampling is a random process, it introduces noise so that estimates from re-sampled particles are initially worse than before

[1]Sequential imputation is one example of a sequential Monte Carlo algorithm that can be applied to a wider class of models than we consider here.

re-sampling.

Liu and Chen (1995) looked at the problem of deciding when to re-sample by using the *effective sample size* defined in Kong et al. (1994) as

$$\mathrm{ESS}(w_t) := \left(\sum_{i=1}^{N} w_t^{(i)2} \right)^{-1}. \tag{2.7}$$

As the name suggests, this gives a measure of the operational strength of the sample. It takes its maximum value of N when the weights are all even and its minimum value of 1 when all the mass is placed on a single point. We can therefore think of a weighted sample as roughly comparable with an unweighted one with sample size $\mathrm{ESS}(w_t)$. We note, however, that the ESS is only meaningful for independent samples and should therefore only be applied to importance weights before re-sampling; after re-sampling we have $\mathrm{ESS}(w_t) = N$ whereas the post re-sampling estimates are initially worse due to the extra Monte Carlo error induced.

Liu and Chen (1995) therefore suggest calculating $\mathrm{ESS}(w_t)$ every time step from the incremental weights and re-sampling if it falls below a predetermined threshold. However, while the threshold should clearly lie between 1 and N, it is unclear what exact value it should take for a given model.

2.1.4 Auxiliary particle filter

Throughout this thesis we will work with the *Auxiliary SIR* (ASIR) filter of Pitt and Shephard (1999a). If we assume that at time $t-1$ we have weighted particles $\{(x_{t-1}^{(i)}, w_{t-1}^{(i)})\}_{i=1}^{N}$ approximating $p(x_{t-1}|y_{1:t-1})$ we can use the filter recursion (2.3) to write our target at time t as

$$p(x_t|y_{1:t}) \simeq c\, g(y_t|x_t) \sum_{i=1}^{N} f(x_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)}, \tag{2.8}$$

where c is a normalising constant. In this approach we aim to approximate

$$c\, g(y_t|x_t)\, f(x_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)} \tag{2.9}$$

by

$$q(x_t|x_{t-1}^{(i)}, y_t)\, \beta_t^{(i)},$$

where $q(\cdot|x_{t-1}, y_t)$ is a distribution we can sample from and $\{\beta_t^{(i)}\}_{i=1}^N$ are normalised weights which sum to 1. We then use a combination of re-sampling and importance sampling to generate a weighted sample approximating (2.8).

The algorithm is shown in Algorithm 2.3. A key feature of the method is the augmentation of the state space by the *auxiliary variable* j_i. This allows the final weights $w_t^{(i)}$ to be even if q and $\beta_t^{(i)}$ are chosen so that (2.9) is well approximated. Since the auxiliary variable is nothing but a label selecting a particular particle $x_{t-1}^{(i)}$, sampling the j_is with the initial weights $\{\beta_t^{(i)}\}$ amounts to re-sampling the

---

Algorithm 2.3: Auxiliary particle filter.

---

1. **Initialisation:** Sample $\{x_0^{(i)}\}$ from the prior $\pi(x_0)$ and set $w_0^{(i)} = 1/N$ for all i.

2. For $t = 1, 2, \ldots$

   (a) **Optionally re-sample:** Calculate the re-sampling weights $\{\beta_t^{(i)}\}$ and compare $\mathrm{ESS}(\beta_t)$ with a predetermined threshold u_N:

      i. If $\mathrm{ESS}(\beta_t) < u_N$, **re-sample** by using the $\{\beta_t^{(i)}\}$ as probabilities to sample N indices $j_1, ..., j_N$ from $\{1, ..., N\}$.

      ii. If $\mathrm{ESS}(\beta_t) \geq u_N$, **do not re-sample** by resetting $\beta_t^{(i)} = 1$ and $j_i = i$ for all i.

   (b) **Propagate:** Sample the new particles $x_t^{(i)}$ independently from $q(\cdot|x_{t-1}^{(j_i)}, y_t)$.

   (c) **Re-weight:** Assign each particle $x_t^{(i)}$ the corresponding importance weight

   $$w_t^{(i)} \propto \frac{g(y_t|x_t^{(i)})\, f(x_t^{(i)}|x_{t-1}^{(j_i)})\, w_{t-1}^{(j_i)}}{q(x_t^{(i)}|x_{t-1}^{(j_i)}, y_t)\, \beta_t^{(j_i)}}$$

   and normalise them to sum to 1.

---

particles before they are propagated. By re-sampling first we propagate useful particles multiple times while leaving behind particles which would lead to small weights. This gives an evenly weighted sample of unique particles rather than the duplicates achieved by re-sampling at the end.

As before, the re-sampling step is optional and can be omitted by setting $j_i = i$ and removing $\beta_t^{(i)}$ from the weight (by setting $\beta_t^{(i)} = 1$) thus propagating each particle $x_{t-1}^{(i)}$ once. This eliminates the extra noise from re-sampling but gives uneven weights. In this form the algorithm is essentially the SIS algorithm applied to our state space model. As with SIS, the effective sample size may be used to decide when to re-sample but, since the initial weights $\{\beta_t^{(i)}\}$ are used to re-sample, the decision should be based upon ESS(β_t) rather than ESS(w_t).

To obtain evenly weighted particles after re-sampling, (2.9) must be well approximated. Optimally this is achieved by setting $q(x_t|x_{t-1}, y_t) = p(x_t|x_{t-1}, y_t)$ and $\beta_t^{(i)} \propto p(y_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)}$ as in sequential imputation in which case we say the filter is *adapted*. However, these densities are often intractable although good approximations of them can give almost even weights.

Pitt and Shephard (1999a) offer some advice on the selection of q and $\beta_t^{(i)}$ including Taylor expanding to second-order $\log g(y_t|x_t)$, so long as it is concave in x_t, to produce a Gaussian proposal density when the state is Gaussian. If the log likelihood is not concave a first-order Taylor expansion may be of use. The re-sampling weights $\beta_t^{(i)}$ can then be approximated by $p(y_t|\hat{x}_t^{(i)})\, w_{t-1}^{(i)}$ where $\hat{x}_t^{(i)}$ is some likely value of $f(x_t|x_{t-1}^{(i)})$. The simplest choice of proposal is $q(x_t|x_{t-1}, y_t) = f(x_t|x_{t-1})$ and $\beta_t^{(i)} = w_{t-1}^{(i)}$ which gives the SIR filter.

2.1.5 Enhancements

Numerous modifications and enhancements have been proposed for the particle filter. We focus here on those that will be of most use to us in this thesis.

Improved initialisation

The standard practice in Bayesian statistics of representing prior uncertainty by prior densities with large variances can cause problems with particle filters. By initialising the algorithm by sampling from a sparse prior it is likely that most of the particles are so far away from the target distribution that they are given negligible weights or are lost by re-sampling. With very uncertain priors this often leaves just one particle with any mass after the first time step.

This problem may be overcome by sampling the first time step $p(x_1|y_1) \propto p(x_1)\, g(y_1|x_1)$ separately. This can be done with importance sampling but other authors have suggested MCMC. If the first observation does not provide enough information to restrict the distribution of all components of the state, it may be necessary to sample the first few time steps simultaneously to initialise the algorithm.

Stratified re-sampling

Re-sampling at every time step is not the best strategy because the re-sampling process introduces extra variation. Liu and Chen (1995) showed that particles need not be sampled independently (termed *multinomial sampling* as the number of re-sampled particles follow a multinomial distribution) and that producing a more stratified sample reduces the additional variance. In particular, they propose *residual sampling* of the particles $\{x^{(i)}\}$ with weights $\{\beta^{(i)}\}$ which involves deterministically picking $\lfloor N\beta^{(i)} \rfloor$ copies of $x^{(i)}$ (where $\lfloor y \rfloor$ is the integer part of y) and sampling the remaining r particles independently from $\{x^{(i)}\}$ using probabilities $\gamma^{(i)} := (N\beta^{(i)} - \lfloor N\beta^{(i)} \rfloor)/r$.

Carpenter et al. (1999) show that the re-sampling noise is minimised by producing a stratified sample of the indices and give an $\mathcal{O}(N)$ algorithm to achieve this. Bolic et al. (2004) review the complexity of re-sampling algorithms and propose an improved algorithm for stratified sampling.

Using stratified sampling in a particle filter makes re-sampling more preferable by reducing the additional variation accrued whilst increasing the number of useful particles to propagate. This therefore means that we can re-sample more often by using a larger ESS threshold.

Rao-Blackwellisation

A further enhancement which can often be used to improve the filter is *Rao-Blackwellisation* (also known simply as *marginalisation*). The idea is that for some models it is possible to integrate out part of the state analytically. This enables the integrable part of the state to be represented by a distribution rather than a specific value which, as the Rao-Blackwell theorem shows, will give less variable estimates. See Casella and Robert (2001) for an introduction to Rao-Blackwellisation in general and Liu and Chen (1998) and Doucet et al. (2000) for applications to particle filtering.

2.1.6 Parameter estimation

We have so far focused on estimating the dynamic state vector X_t given a series of observations $y_{1:t}$ but in practical applications the model often contains additional static parameters θ. These may appear in the state transition density f, the observation density g or even the prior π. When observations are arriving sequentially, we may want incremental estimates of θ or even the joint distribution $p(x_t, \theta|y_{1:t})$. Alternatively, we may simply require a final estimate of the parameters to use for off-line inference.

Augmenting the state vector

The simplest way to estimate static parameters is to augment the state vector with θ and proceed with any particle filter. This naturally provides sequential

parameter estimates through the filter distribution $p(x_t, \theta|y_{1:t})$ but has one major flaw. Since θ remains constant from one time step to the next, the parameter space is only explored in the initialisation of the algorithm. Re-sampling therefore permanently reduces the number of unique θ values which ultimately degrades to a single particle with which to represent $p(\theta|y_{1:t})$.

Many authors have suggested methods to rejuvenate re-sampled particles, most of which can also be applied when there are no static parameters. Gordon et al. (1993) and Liu and West (2001) add noise to the particles which effectively replaces the static parameters with slowly varying ones. This can lead to errors when too much noise is added while with too little noise the particles still struggle to explore the sample space, and so degrade on re-sampling.

Other methods include Fearnhead (1998) and Gilks and Berzuini (2001) who introduce MCMC moves to diversify the sample after re-sampling. Since the MCMC moves require the closed form joint density $p(x_{0:t}|y_{1:t})$, the complexity of the moves increases with t. This is overcome by Fearnhead (2002) with the use of sufficient statistics when they exist.

Other methods

An alternative method for estimating θ of Storvik (2002) is based on writing the joint target at time t as

$$p(x_{1:t}, \theta|y_{1:t}) \propto p(x_{1:t-1}|y_{1:t-1})\, p(\theta|x_{1:t-1}, y_{1:t-1})\, f(x_t|x_{t-1}, \theta)\, g(y_t|x_t, \theta).$$

This justifies sampling $\theta^{(i)}$ afresh from $p(\theta|x_{1:t-1}, y_{1:t-1})$ and updating particles $\{x_{t-1}^{(i)}\}$ with a particle filter conditional on $\{\theta^{(i)}\}$. For its complexity to remain constant over time, there must be a sufficient statistic $T^{(i)}$ of $(x_{1:t-1}^{(i)}, y_{1:t-1})$ for θ that can be easily updated recursively.

When we only require an off-line estimate of the parameters given a block of data $y_{1:T}$ we may consider maximising the overall model likelihood $L(\theta) := p(y_{1:T}|\theta)$. Kitagawa (1996) gave the following estimate for the model likelihood:

$$L(\theta) = p(y_{1:T}|\theta) \simeq \prod_{t=1}^{T} \sum_{i=1}^{N} g(y_t|x_{t|t-1}^{(i)}, \theta)\, w_{t-1}^{(i)}, \tag{2.10}$$

where $x_{t|t-1}^{(i)}$ is a predictive particle sampled from the state $f(x_t|x_{t-1}^{(i)}, \theta)$ and $\{(x_t^{(i)}, w_t^{(i)})\}$ are sampled from a particle filter given θ. In principle, $L(\theta)$ may be maximised on a grid of θ values although this is only feasible when θ has few dimensions as each evaluation requires a full run of the particle filter. See Hürzeler and Künsch (2001) and Pitt (2002) for alternative likelihood based methods.

Poyiadjis et al. (2005) provide a particle method for approximating the derivative of the filter density with respect to unknown parameters θ. This can then be used to maximise the likelihood via gradient-based methods. Briers et al. (2004) and Wills et al. (2008) alternatively propose to maximise the likelihood using an *Expectation-Maximisation* (EM) algorithm based upon the repeated use of a particle smoother (see Section 2.2). To avoid this step, Andrieu et al. (2005) propose an EM-type algorithm that makes use of the invariant state distribution, if it exists, to maximise a pseudo-likelihood to give a parameter estimate.

2.2 Particle Smoothing

This section introduces the smoothing problem and reviews current extensions of particle filters to smoothing.

2.2.1 Kalman Smoothing

While the filtering problem focuses on estimating the current state given a series of observations, the corresponding smoothing problem is interested in updating past values of the state given a block of data. Specifically, we wish to estimate the *smoothing* distribution $p(x_t|y_{1:T})$ for $t = 1, \ldots, T$. In some applications the *joint smoothing distribution* $p(x_{1:T}|y_{1:T})$ is of interest but we assume for now that we wish only to review past states individually.

We motivated the filtering problem with the presumption that data were arriving sequentially and inference was required on-line. To contrast with this, we now assume that a block of data $y_{1:T}$ has been observed and we need only the final state estimates given this data. We can in principle use MCMC to draw samples from the joint smoothing distribution using (2.5) but these often perform poorly when the state is highly correlated. Also, since the whole path $X_{0:T}$ must be sampled jointly, the dimension of the MCMC state will be very large if we have a long time series.

As an alternative to MCMC, we focus on efficient algorithms that are sequential over time as these will allow us to consider inference for long series with an overall complexity of $\mathcal{O}(T)$ where T is the number of observations. We also focus on methods that make use of the sequential algorithms that exist for filtering.

A recursive formula for calculating the marginal smoothing density is given by

$$p(x_t|y_{1:T}) = p(x_t|y_{1:t}) \int \frac{f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} p(x_{t+1}|y_{1:T}) \, \mathrm{d}x_{t+1}, \quad \text{for } t < T, \tag{2.11}$$

where

$$p(x_{t+1}|y_{1:t}) = \int f(x_{t+1}|x_t)\, p(x_t|y_{1:t})\, \mathrm{d}x_t.$$

This allows the target density $p(x_t|y_{1:T})$ to be calculated backwards in time with prior knowledge of the filter densities $p(x_t|y_{1:t})$. The recursion begins with $p(x_T|y_{1:T})$ which is of course both a filter and smoother density. Thus the smoothing densities may, in principle, be calculated by filtering forwards in time and then smoothing backwards with (2.11).

As was the case for the filter recursion (2.3), the formula above will be intractable for general state space models. The linear-Gaussian model of (2.4) is again an important exception with the *Kalman Smoother* providing algebraic formulae for the mean and covariance matrices of the Gaussian smoothing densities (see Anderson and Moore (1979) for details).

2.2.2 Smoothing while filtering

In its simplest form, particle smoothing can be achieved from a simple extension to the particle filter as shown by Kitagawa (1996), and we call the resulting algorithm the *Filter-Smoother*. As with the filter distribution $p(x_t|y_{1:t})$ in (2.3), we have a recursive solution for the joint smoothing distribution:

$$p(x_{1:t}|y_{1:t}) \propto g(y_t|x_t)\, f(x_t|x_{t-1})\, p(x_{1:t-1}|y_{1:t-1}). \tag{2.12}$$

By comparing (2.3) and (2.12) it is easy to show that the particle filter steps can be used to update weighted paths $\{(x_{1:t}^{(i)}, w_t^{(i)})\}_{i=1}^N$ approximating $p(x_{1:t}|y_{1:t})$. Doing so simply requires keeping track of the inheritance of the newly sampled particle $x_t^{(i)}$. This means that any filtering algorithm can be used and the method inherits the $\mathcal{O}(N)$ computational complexity of the filter making large numbers of particles feasible.

While this Filter-Smoother approach can produce an accurate approximation of the filtering distribution $p(x_t|y_{1:t})$ it gives a poor representation of previous states. To see this we note that whenever we re-sample the paths $\{x^{(i)}_{1:t-1}\}$ (by re-sampling $\{\tilde{x}^{(i)}_t\}$ in the SIR filter or sampling the auxiliary variables $\{j_i\}$ in the auxiliary filter) we end up with multiple copies of some paths but lose others altogether. Therefore the number of distinct particles at any given time decreases monotonically the more times we re-sample. Also, with multiple copies of some particles, their weights are effectively added together on a single point so that marginally the weights become more uneven as we look back in time.

This can be seen in Figure 2.2 which represents ten smoothed paths $x^{(i)}_{1:6}$ showing how they re-weight filter particles. As you can see, particles which are lost due to re-sampling receive no weight and particles with many offspring have large weights. While the filter approximation at time 6 is good, the weights become more uneven as the number of weighted particles decreases going back in time. This is not surprising since the particles at times $t < 6$ are drawn to approximate $p(x_t|y_{1:t})$ so must be unevenly weighted if they are to represent a different distribution.

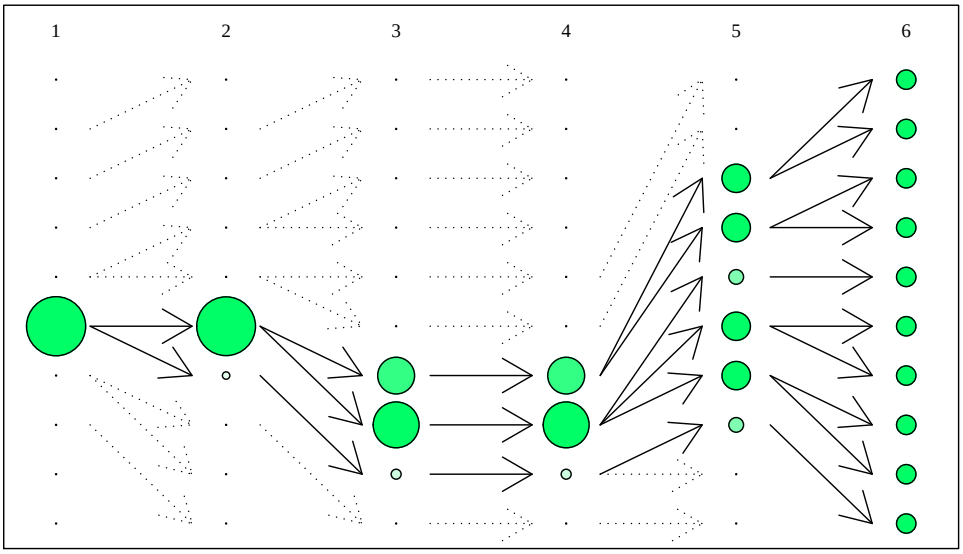


Figure 2.2: Diagram showing how the simple Filter-Smoother re-weights the filter particles. The arrows represent the dependencies between the particles at time t and $t-1$ due to re-sampling. The size of the particle represents its total weight as a draw from the smoothed distribution.

As a final point we note that re-sampling more infrequently can improve this method of smoothing although there is a limit to how much this can help. Even with no re-sampling, the approximation to $p(x_t|y_{1:T})$ will deteriorate as $T-t$ gets large: with the particle approximation tending to give non-negligible weight to all but a small subset of particles, and eventually only one particle having a non-negligible weight.

2.2.3 Forwards-backwards smoothers

The *Forward-Backward Smoother* of Doucet et al. (2000), as well as the related algorithms of Hürzeler and Künsch (1998) and Tanizaki (2001), is based around the backwards recursion (2.11). The unknown densities can be approximated using filter particles from the current time t and smoother particles from $t+1$ to obtain

$$p(x_t|y_{1:T}) \simeq \sum_{i=1}^{N} \delta(x_t - x_t^{(i)}) w_{t|T}^{(i)},$$

where

$$w_{t|T}^{(i)} := \sum_{j=1}^{N} \frac{f(x_{t+1|T}^{(j)}|x_t^{(i)}) w_t^{(i)}}{\sum_{k=1}^{N} f(x_{t+1|T}^{(j)}|x_t^{(k)}) w_t^{(k)}} w_{t+1|T}^{(j)} \tag{2.13}$$

and $\delta(\cdot)$ is the Dirac delta function.

This approximation can be used to sequentially re-weight the filter particles backwards in time so that they represent the marginal smoothing densities. Since the calculation of each weight is an $\mathcal{O}(N^2)$ operation, a crude application of the algorithm would be $\mathcal{O}(N^3)$. However, since the denominator of each summand in (2.13) does not depend on i, they may be each calculated once and stored to reduce the overall complexity to $\mathcal{O}(N^2)$.

While the Forward-Backward smoother only approximates the marginal smoothing densities, the related algorithms of Hürzeler and Künsch (1998) and Godsill et al. (2004) re-sample paths backwards in time to produce samples from the joint

smoothing density.

2.2.4 Two-Filter Smoother

The *Two-Filter Smoother* of Briers et al. (2004) combines samples from a particle filter with those from a *backwards information filter* to produce estimates of $p(x_t|y_{1:T})$.

Backwards information filter

The backwards information filter produces sequential approximations of the likelihood $p(y_{t:T}|x_t)$ backwards through time and is based on the following recursion:

$$p(y_{t:T}|x_t) = g(y_t|x_t) \int f(x_{t+1}|x_t) p(y_{t+1:T}|x_{t+1}) \, \mathrm{d}x_{t+1}, \quad \text{for } t < T. \tag{2.14}$$

Since $p(y_{t:T}|x_t)$ is not a probability density function in x_t it may not have a finite integral over x_t in which case a particle representation will not work. The smoothing algorithm in Kitagawa (1996) assumes implicitly that this is not the case but Briers et al. (2004) propose the following construction which will always give a finite measure.

They introduce artificial prior distributions $\gamma_t(x_t)$ to yield the backwards filter densities

$$\tilde{p}(x_t|y_{t:T}) :\propto \gamma_t(x_t) \, p(y_{t:T}|x_t). \tag{2.15}$$

Artificial priors are chosen so that $\gamma_t(x_t)$ is available in closed form. Briers et al. (2004) assume the recursive relationship $\gamma_t(x_t) = \int f(x_t|x_{t-1}) \gamma_{t-1}(x_{t-1}) \, \mathrm{d}x_{t-1}$ after initial specification of $\gamma_0(x_0)$ which, if selected to be the prior $\pi(x_0)$, yields $\gamma_t(x_t) = p(x_t)$ and $\tilde{p}(x_t|y_{t:T}) = p(x_t|y_{t:T})$ for all time steps t. This, however, restricts the applicability of the method to tractable state models such as the linear-Gaussian and is not necessary; any choice of $\gamma_t(x_t)$ will yield a valid back-

wards filter density to propagate. Also, if the likelihood $g(y_t|x_t)$ is integrable, we can instead propagate a particle representation of $p(y_{t:T}|x_t)$ by assuming $\gamma_t(x_t) \equiv 1$ throughout the following derivation.

Following on from (2.14) the backwards filter is derived via

$$\begin{aligned}\tilde{p}(x_t|y_{t:T}) &\propto \gamma_t(x_t)g(y_t|x_t)\int f(x_{t+1}|x_t)\frac{\tilde{p}(x_{t+1}|y_{t+1:T})}{\gamma_{t+1}(x_{t+1})}\,\mathrm{d}x_{t+1} \\ &\simeq \gamma_t(x_t)g(y_t|x_t)\sum_{k=1}^{N}\frac{f(\tilde{x}_{t+1}^{(k)}|x_t)}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})}\tilde{w}_{t+1}^{(k)},\end{aligned}$$

where the weighted particles $\{(\tilde{x}_{t+1}^{(k)}, \tilde{w}_{t+1}^{(k)})\}$ approximate $\tilde{p}(x_{t+1}|y_{t+1:T})$. This is very similar to the derivation of the forwards filter and as such many filtering algorithms and enhancements can be modified for this purpose.

For example, an auxiliary backwards filter in the style of Pitt and Shephard (1999a) can be made by finding a distribution $\tilde{q}(\cdot|y_t, \tilde{x}_{t+1}^{(k)})$ we can sample from such that

$$\tilde{q}(x_t|y_t, \tilde{x}_{t+1}^{(k)})\tilde{\beta}_t^{(k)} \simeq \gamma_t(x_t)g(y_t|x_t)f(\tilde{x}_{t+1}^{(k)}|x_t)\frac{\tilde{w}_{t+1}^{(k)}}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})}.$$

We then proceed analogously to Algorithm 2.3 for $t = T, ..., 1$ after initialising the algorithm with particles drawn from $\gamma_{T+1}(x_{T+1})$. An adapted backwards filter giving even weights $\tilde{w}_t^{(k)} = 1/N$ is achieved with $\tilde{q}(x_t|y_t, \tilde{x}_{t+1}^{(k)}) = p(x_t|y_t, \tilde{x}_{t+1}^{(k)})$ and $\tilde{\beta}_t^{(k)} \propto \tilde{p}(y_t|\tilde{x}_{t+1}^{(k)})\tilde{w}_{t+1}^{(k)}$ where we again use $\tilde{p}$ to denote a distribution which uses $\gamma_t(x_t)$ throughout instead of $p(x_t)$.

Two-Filter Smoother

Having run a forwards particle filter and a backwards information filter, it is possible to combine the two to estimate $p(x_t|y_{1:T})$. The Two-Filter Smoother is based

upon writing the target density as

$$\begin{aligned} p(x_t|y_{1:T}) &\propto p(x_t|y_{1:t-1}) \cdot p(y_{t:T}|x_t) \\ &\propto \int f(x_t|x_{t-1})\, p(x_{t-1}|y_{1:t-1})\, \mathrm{d}x_{t-1} \cdot \frac{\tilde{p}(x_t|y_{t:T})}{\gamma_t(x_t)}. \end{aligned}$$

Therefore filter particles $\{(x_{t-1}^{(j)}, w_{t-1}^{(j)})\}$ approximating $p(x_{t-1}|y_{1:t-1})$ and backwards filter particles $\{(\tilde{x}_t^{(k)}, \tilde{w}_t^{(k)})\}$ approximating $\tilde{p}(x_t|y_{t:T})$ are used to obtain

$$p(x_t|y_{1:T}) \simeq \sum_{k=1}^{N} \delta(x_t - \tilde{x}_t^{(k)}) \tilde{w}_{t|T}^{(k)},$$

where

$$\tilde{w}_{t|T}^{(k)} :\propto \frac{\tilde{w}_t^{(k)}}{\gamma_t(\tilde{x}_t^{(k)})} \sum_{j=1}^{N} f(\tilde{x}_t^{(k)}|x_{t-1}^{(j)}) w_{t-1}^{(j)}. \tag{2.16}$$

Thus particles from a forwards filter are used to re-weight those from a backwards filter so that they represent the target distribution.

2.3 Univariate Extreme Value Theory

This section gives an overview of univariate extreme value theory focusing on aspects that will be of use in this thesis. For a more general introduction to the theory and analysis of extreme values see Embrechts et al. (1997), Coles (2001) or de Haan and Ferreira (2006) amongst others.

2.3.1 Maxima of IID random variables

Extreme value distributions for maxima

We begin by studying the extremal properties of a sequence of independent and identically distributed (IID) univariate random variables $Y_1, \ldots, Y_n$. We present theory for the maxima $M_n := \max\{Y_1, \ldots, Y_n\}$ but can easily obtain corresponding results for the minima through the identity

$$\min\{Y_1, \ldots, Y_n\} = -\max\{-Y_1, \ldots, -Y_n\}.$$

If the Y_is have common distribution function $F(y) = \mathbf{P}\{Y_i \leq y\}$, we have

$$\mathbf{P}\{M_n \leq y\} = \mathbf{P}\{Y_1 \leq y, \ldots, Y_n \leq y\} = F(y)^n$$

so that the distribution function of M_n is known if F is known. However, since in practice we rarely know the exact distribution of a sample, we look for asymptotic arguments to motivate a distribution for M_n.

Taking the limit as $n \to \infty$, $M_n \to y^F$ where $y^F := \sup\{y|F(y) < 1\}$ is the upper end point of F. The limit distribution of M_n is therefore degenerate and so of no use for modelling with finite n. However, the same is true for $\bar{Y}_n := \text{mean}\{Y_{1:n}\}$ but, as the Central Limit Theorem shows, linear normalisation within the limit

can lead to a non-degenerate distribution (in the case of the Gaussian limit law under the additional constraint of $\mathrm{Var}(Y_i) < \infty$). With this in mind, Theorem 2.1, the *Extremal Types Theorem* (ETT) of Fisher and Tippett (1928), shows which distributions are obtainable by linear normalisation of M_n (see Leadbetter et al. (1983) for details).

Theorem 2.1 (Extremal Types Theorem)**.** *Given a sequence* $\{Y_n\}$ *of IID random variables, define* $M_n := \max\{Y_{1:n}\}$*. If there exist sequences* $a_n \in \Re$ *and* $b_n > 0$ *such that*

$$\mathbf{P}\left\{\frac{M_n - a_n}{b_n} \leq y\right\} \to G(y) \quad \text{as } n \to \infty,$$

for a non-degenerate distribution G*, then* G *is either:*

Negative-Weibull(μ, σ, α)**:**

$$G(x) = \exp\left(-\left[-\left(\frac{y-\mu}{\sigma}\right)\right]_+^{\alpha}\right),$$

Gumbel(μ, σ)**:**

$$G(y) = \exp\left(-\exp\left\{-\left(\frac{y-\mu}{\sigma}\right)\right\}\right),$$

Fréchet(μ, σ, α)**:**

$$G(y) = \exp\left(-\left[\frac{y-\mu}{\sigma}\right]_+^{-\alpha}\right),$$

for some $\mu \in \Re$*,* $\sigma > 0$ *and* $\alpha > 0$*, where* $[y]_+ := \max\{y, 0\}$*.*

Figure 2.3 shows the shape of the density functions for the three extremal distributions. Note that the Negative-Weibull distribution has an upper end point while the Fréchet has a lower end point.

Generalised Extreme Value distribution

Von Mises (1954) and Jenkinson (1955) identified the *Generalised Extreme Value*

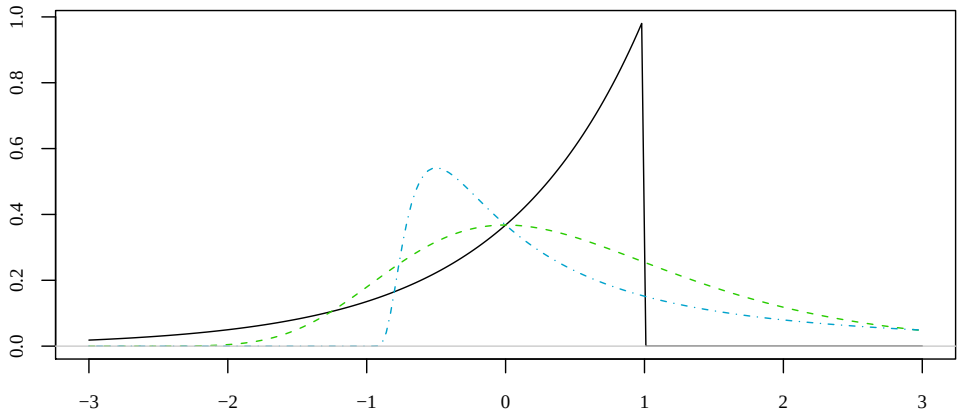


Figure 2.3: Probability density functions of the Negative-Weibull(1,1,1) = GEV(0,1,-1) (—), Gumbel(0,1) = GEV(0,1,0) (---) and Fréchet(-1,1,1) = GEV(0,1,1) (·–) distributions.

(GEV) distribution which unifies the three limit distributions for linearly normalised maxima. Its cumulative distribution function is given by

$$G(y) = \exp\left(-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right), \tag{2.17}$$

where $\mu \in \Re$, $\sigma > 0$ and $\xi \in \Re$. We interpret the $\xi = 0$ case to mean the limit as $\xi \to 0$.

The shape parameter ξ is key to relating the original three extremal distributions to the GEV form:

- If $\xi < 0$, GEV(μ, σ, ξ) = Negative-Weibull$(\mu - \frac{\sigma}{\xi}, -\frac{\sigma}{\xi}, -\frac{1}{\xi})$.
- If $\xi = 0$, GEV(μ, σ, ξ) = Gumbel(μ, σ).
- If $\xi > 0$, GEV(μ, σ, ξ) = Fréchet$(\mu - \frac{\sigma}{\xi}, \frac{\sigma}{\xi}, \frac{1}{\xi})$.

Basic properties of the GEV are summarised in Table 2.1. We note in particular that the Fréchet distribution ($\xi > 0$) has a heavy tail with $\mathbf{E}(X) = \infty$ if $\xi \geq 1$.

Another important property is *max-stability*: A distribution F is said to be *max-stable* if, for IID random variables $Y_1, \ldots, Y_n \sim F(\cdot)$, there exist constants A_n and $B_n > 0$ such that $A_n + B_n \max\{Y_{1:n}\} \sim F(\cdot)$ or equivalently $F^n((y - A_n)/B_n) =$

parameters	$\mu \in \Re$	$\sigma > 0$	$\xi \in \Re$
cdf	$G(y) = \exp\left(-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right)$		
pdf	$g(y) = \frac{1}{\sigma}\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)} G(y)$		
support	$(-\infty, \mu - \frac{\sigma}{\xi})$ if $\xi < 0$	$\Re$ if $\xi = 0$	$(\mu - \frac{\sigma}{\xi}, \infty)$ if $\xi > 0$
mean	$\mu + \frac{\sigma}{\xi}(\Gamma(1-\xi) - 1)$ if $\xi < 1$		∞ if $\xi \geq 1$
variance	$\frac{\sigma^2}{\xi^2}(\Gamma(1-2\xi) - \Gamma(1-\xi)^2)$ if $\xi < \frac{1}{2}$		∞ if $\xi \geq \frac{1}{2}$

Table 2.1: Properties of the Generalised Extreme Value distribution. Both the mean and the variance depend on the Gamma function $\Gamma(\cdot)$.

$F(y)$ for all y. It can be shown that the GEV satisfies this property for all ξ but also that it is the only class of distributions to do so. This mirrors the convolution property of the Gaussian distribution which says that the sum of n IID Gaussian random variables is also Gaussian.

Using the GEV distribution in Theorem 2.1 gives the *Unified Extremal Types Theorem* (UETT), Theorem 2.2 below. If, for IID random variables $Y_1, \ldots, Y_n \sim F(\cdot)$, the normalising sequences that give the GEV limit G exist, we say that F is in the *domain of attraction* of G and write $F \in \mathcal{D}_\xi$.

Theorem 2.2 (Unified Extremal Types Theorem)**.** *Given a sequence $\{Y_n\}$ of IID random variables, define $M_n := \max\{Y_{1:n}\}$. If there exist sequences $a_n \in \Re$ and $b_n > 0$ such that*

$$\mathbf{P}\left\{\frac{M_n - a_n}{b_n} \leq y\right\} \to G(y) \quad \text{as } n \to \infty, \tag{2.18}$$

for a non-degenerate distribution G, then G is the cumulative distribution function of the Generalised Extreme Value distribution for some parameters $\mu \in \Re$, $\sigma > 0$ and $\xi \in \Re$.

Modelling with the GEV distribution

Theorem 2.2 is used to motivate the GEV model for maxima by assuming the limit (2.18) holds for a finite n so that

$$\mathbf{P}\left\{\frac{M_n - a_n}{b_n} \leq y\right\} = G(y),$$

for some constants a_n and b_n. Setting $z = a_n + b_n y$ this gives

$$\mathbf{P}\{M_n \leq z\} = G\left(\frac{z - a_n}{b_n}\right)$$

which is a GEV distribution since a_n and b_n can be absorbed into the location and scale parameters μ and σ. This therefore justifies modelling the maxima of IID variables as GEV.

In order to produce multiple observations for a statistical analysis, a series of IID observations should be split into blocks and the block maxima modelled as GEV. The choice of block size is important as a larger block gives fewer observations but is better justified for the GEV limit.

Maximum likelihood is commonly used to estimate the GEV parameters from a sample as proposed by Prescott and Walden (1980, 1983). Smith (1985) showed that the regular asymptotic properties of the maximum likelihood estimator such as asymptotic normality only hold when $\xi > -1/2$. While this potentially causes difficulties with estimation, many authors have noted that $\xi \leq -1/2$ rarely occurs in practice. See Coles (2001) for a review of maximum likelihood estimation in extreme value models.

Coles and Powell (1996) review Bayesian estimation of the GEV parameters. No conjugate prior for the GEV distribution exists so numerical methods such as MCMC are commonly used for inference. One advantage of Bayesian estimation is that it overcomes the regularity constraints of the maximum likelihood estimates

on ξ. Other methods of parameter estimation include the moment based estimators of Hosking et al. (1985) and Dekkers et al. (1989).

Since the GEV model is based upon the assumption that a non-degenerate limit distribution exists, the GEV fit should be checked with methods such as probability-probability (PP) plots and quantile-quantile (QQ) plots. For cases where a non-degenerate limit does not exist, a GEV limit is often achieved after the variables are transformed by a non-linear function such as log. Therefore applying such a transformation to a dataset may give a better fit.

Extension to r-largest variables

Inference for IID variables with maxima alone is inefficient as most of the data are discarded. Keeping instead the r-largest values in a block gives more data which could give better estimates of the parameters. Using arguments similar to Theorem 2.2, it can be shown that if the r-largest order statistics $Y_r \leq \ldots \leq Y_1$ are linearly normalised towards a non-degenerate joint distribution, that distribution will have a joint density function of

$$g(y_1, \ldots, y_r) = G(y_r) \prod_{i=1}^{r} \frac{g(y_i)}{G(y_i)} \quad \text{for } y_r \leq \ldots \leq y_1, \tag{2.19}$$

where $g(y)$ and $G(y)$ are the pdf and cdf respectively of a GEV(μ, σ, ξ) random variable. For details see Weissman (1978), Smith (1986) and Tawn (1988b).

This result provides a model for the r-largest values that is consistent with the GEV model for maxima and requires no extra parameters. This therefore allows additional data values to be used for more precise estimates of μ, σ, and ξ. A choice must be made on the appropriate value of r before the data is fitted to balance the efficiency gains of more data with the asymptotic justification of the model. This choice is typically made by fitting a selection of r values and assessing the model fit.

2.3.2 Point process characterisation

Point process

We begin again by assuming $Y_1, \ldots, Y_n$ are IID random variables with common distribution function F and that $F \in \mathcal{D}_\xi$, that is F is in the domain of attraction of GEV$(0, 1, \xi)$. Define the sequence of point processes on $[0, 1] \times \Re$ by

$$P_n := \left\{ \left(\frac{i}{n+1}, \frac{Y_i - a_n}{b_n} \right) \middle| i = 1, \ldots, n \right\}, \tag{2.20}$$

where a_n and b_n are the same normalising sequences used in Theorem 2.2 to give a GEV$(0, 1, \xi)$ limit. We consider the limiting process as $n \to \infty$.

Clearly, the limiting process is non-degenerate since $(M_n - a_n)/b_n \to G(y)$ and the r-largest values also have a non-degenerate limit. However, the smallest values are all normalised towards the value

$$z_l := \lim_{n \to \infty} \frac{y_F - a_n}{b_n},$$

where $y_F := \inf\{y | F(y) > 0\}$ is the lower end point of F. Theorem 2.3, due to Pickands (1971), shows that the limiting process for all large values is a particular non-homogeneous Poisson process.

Theorem 2.3 (Point process limit)**.** *Given a sequence $\{Y_n\}$ of IID random variables with common distribution function $F \in \mathcal{D}_\xi$, define a sequence of point processes $\{P_n\}$ by (2.20). Then $P_n \to P$ as $n \to \infty$ on the set $[0, 1] \times (z_l, \infty)$ where P is a non-homogeneous Poisson process with intensity*

$$\lambda(s, z) = [1 + \xi z]_+^{-(1+\frac{1}{\xi})}.$$

We again interpret the $\xi = 0$ case as the limit as $\xi \to 0$ which gives the intensity $\lambda(s, z) = e^{-z}$.

The Poisson process limit says a lot about the asymptotic distribution of extreme values. Many probabilities of interest may be written in terms of the counting process $N(A)$ that counts the number of events in a set $A \subset [0,1] \times (z_l, \infty)$ and has a Poisson distribution with mean given by the integrated intensity

$$\Lambda(A) := \iint_A \lambda(s,z)\,\mathrm{d}s\,\mathrm{d}z.$$

For example, the GEV limit of the normalised maxima may be derived from

$$\begin{aligned}\mathbf{P}\left\{\frac{M_n - a_n}{b_n} \leq y\right\} &\to \mathbf{P}\big\{N\big([0,1]\times(y,\infty)\big) = 0\big\} \\ &= \exp\big(-\Lambda\big([0,1]\times(y,\infty)\big)\big) \\ &= \exp\Big(-[1+\xi y]_+^{-\frac{1}{\xi}}\Big).\end{aligned}$$

The distribution of the r-largest order statistics may be derived in a similar way.

Modelling with the point process limit

As before with the GEV limit, we can use this asymptotic result to motivate a model by assuming it to be true for a finite n, that is $P_n = P$. Since n is fixed, a_n and b_n again become location and scale parameters μ and σ respectively. We can then transform the current process with points $(i/(n+1), (Y_i - \mu)/\sigma)$ into the Poisson process $\{(i/(n+1), Y_i)\}$ on $[0,1] \times [u, \infty)$ with intensity

$$\lambda(s,y) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-(1+\frac{1}{\xi})},$$

where $u > \mu + \sigma z_l$ is a high threshold.

This suggests the following model: select a threshold u and assume that all $Y_i > u$ are samples from the above Poisson process. A dataset of n_u threshold exceedances

$y_1, \ldots, y_{n_u}$ can then be modelled with the likelihood

$$\begin{aligned} p(\mathbf{y}, n_u|\mu, \sigma, \xi) &= p(n_u|\mu, \sigma, \xi)\, p(\mathbf{y}|n_u, \mu, \sigma, \xi) \\ &\propto \exp\big(-\Lambda\big([0,1]\times[u,\infty)\big)\big) \prod_{i=1}^{n_u} \lambda\left(\frac{i}{n+1}, y_i\right), \end{aligned} \tag{2.21}$$

where

$$\Lambda\big([0,1]\times[u,\infty)\big) = \int_u^\infty \int_0^1 \lambda(s,y)\,\mathrm{d}s\,\mathrm{d}y = \left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}.$$

Note that n_u appears as data in the likelihood since the number of exceedances of u is random.

It should be noted that the previous strategy of modelling block maxima of a series as GEV(μ, σ, ξ) contradicts the current model which takes these parameters for the maximum of the whole dataset. This discrepancy may be removed by making better use of the time component in the point process. If $\{Y_i\}$ is a time series, the index i corresponds to time and so the current process transforms time from $1:n$ into $[0,1]$ by taking $s = i/(n+1)$. Assuming for simplicity we want annual maxima to be GEV(μ, σ, ξ), we merely need to transform time so that each year falls into a unit interval.

If we assume our series is observed between times t_b and t_e (measured in years), this gives a Poisson process $\{(t_i, Y_i)\}$ on $[t_b, t_e]\times[u,\infty)$ with intensity

$$\lambda(t,y) = \frac{1}{\sigma}\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-(1+\frac{1}{\xi})},$$

where t_i is the time corresponding to variable Y_i. A dataset $y_1, \ldots, y_{n_u}$ may then be modelled with the likelihood

$$p(\mathbf{y}, n_u|\mu, \sigma, \xi) \propto \exp\left(-(t_e - t_b)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right) \prod_{i=1}^{n_u} \lambda(t_i, y_i), \tag{2.22}$$

which only differs from (2.21) in the additional $(t_e - t_b) =$ *'number of years'* term. By exploiting the time component it is now easier to derive time related properties of the process. For example, the asymptotic distribution of the maximum over k years of observations, $M_{[0,k]}$, has cdf

$$\begin{aligned}\mathbf{P}\{M_{[0,k]} \leq y\} &= \mathbf{P}\big\{N\big([0,k] \times (y,\infty)\big) = 0\big\} \\ &= \exp\left(-k\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right) \\ &= G(y)^k,\end{aligned}$$

where G is the GEV(μ, σ, ξ) distribution function, so the distribution of the annual maximum is G.

By using all the data above a threshold rather than just block maxima to estimate the same three parameters we should expect gains in efficiency with tighter standard errors. However, the selection of u leads to a trade-off between variance and bias; by reducing u we include more data but the asymptotic justification for the model becomes weaker. It is common to fit a dataset with a selection of thresholds and pick the one with the best model fit.

For examples of modelling with the point process limit see Smith (1989), Embrechts et al. (1997) or Coles (2001) amongst others. See also Coles and Tawn (1996) for an example of a Bayesian analysis which incorporates expert information into the prior to supplement the data.

Generalised Pareto distribution

Davison and Smith (1990) propose a related approach which models the threshold exceedances and the number of them separately. As with the block maxima model, it may be derived from the point process.

We begin by showing the survivor function of the excess above the threshold u is

given by

$$\begin{aligned}\mathbf{P}\{Y-u>y|Y>u\} &= \frac{\mathbf{P}\{Y>u+y\}}{\mathbf{P}\{Y>u\}} \\ &= \frac{\Lambda\big([t_b,t_e]\times[u+y,\infty)\big)}{\Lambda\big([t_b,t_e]\times[u,\infty)\big)} \\ &= \left[1+\xi\frac{y}{\sigma_u}\right]_+^{-\frac{1}{\xi}},\end{aligned}$$

which is the survivor of the *Generalised Pareto Distribution* (GPD) with scale parameter $\sigma_u := \sigma + \xi(u-\mu)$ and shape parameter ξ. It is important to notice that the shape parameter is the same as in the GEV for block maxima and that the GPD limit holds for any threshold u, a property known as *threshold stability*.

The GPD shares many properties with the GEV distribution, most notably through the common shape parameter ξ. Pickands (1975) show that the GPD limit is obtained for threshold exceedances if and only if the maxima are in the domain of attraction of the GEV, with the same ξ. Like the GEV, the GPD is heavy tailed with $\xi > 0$ but has an upper bound if $\xi < 0$. If $\xi = 0$ the GPD becomes the exponential distribution with mean σ_u.

We have already shown that the number of exceedances of u is a Poisson distribution. A sample of threshold exceedances may therefore be fitted by modelling the number of exceedances, n_u, as Poisson(Λ) and the peaks over the threshold, $Y_i - u$, as IID GPD(σ_u, ξ). This is often referred to as the *Peaks Over Threshold* (POT) approach.

The equivalence of this approach to the point process model above is seen by writing down the joint likelihood

$$\begin{aligned}p(\mathbf{y}, n_u|\Lambda,\sigma_u,\xi) &= p(n_u|\Lambda)\,p(\mathbf{y}|n_u,\sigma_u,\xi) \\ &\propto \Lambda^{n_u}\exp(-\Lambda)\prod_{i=1}^{n_u}\frac{1}{\sigma_u}\left[1+\xi\left(\frac{y_i-u}{\sigma_u}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)},\end{aligned}$$

which is simply a reparametrisation of the likelihood (2.22). The advantage of the POT method is that inference about the expected number of observations Λ is separated from inference about the GPD parameters σ_u and ξ. One disadvantage, however, is that σ_u depends on the threshold u unlike any of the GEV parameters.

For further details and examples of the peaks over threshold method see Davison and Smith (1990) and Embrechts et al. (1997). See also Pickands (1994) for a Bayesian analysis with the GPD.

2.3.3 Dependent processes

So far we have considered the extremal properties of an IID sequence $Y_1, \ldots, Y_n$. We now relax this condition by assuming the sequence is stationary, that is

$$\mathbf{P}\{Y_{t_1} \leq y_1, \ldots, Y_{t_k} \leq y_k\} = \mathbf{P}\{Y_{t_1+\tau} \leq y_1, \ldots, Y_{t_k+\tau} \leq y_k\},$$

so that $Y_{t_1}, \ldots, Y_{t_k}$ and $Y_{t_1+\tau}, \ldots, Y_{t_k+\tau}$ have the same joint distribution for any k, $t_1, \ldots, t_k$ and τ.

Maxima of stationary sequences

We again begin by considering the maximum value of the sequence M_n. We know that if the sequence is independent, the only possible distribution for the linearly normalised maxima is GEV. However, we should not expect any stationary sequence to give a GEV limit. For example, if $Y_1 \sim F(\cdot)$ for an arbitrary distribution F and $Y_i = Y_1$ for all $i > 1$, the sequence is stationary but the maxima equals Y_1 which has distribution F. We therefore need to impose some extra conditions to obtain a smaller class of limit distributions.

One such condition is *Asymptotic Independence of Maxima* (AIM) of O'Brien (1987) which is one of many conditions that limit the amount of long-range de-

pendence in the series. A stationary sequence $\{Y_n\}$ is said to have Asymptotic Independence of Maxima relative to a sequence $\{c_n\}$ (written $\{Y_n\}$ has AIM(c_n)) if there exists an o(n) sequence $q_n > 0$, such that

$$\max |\mathbf{P}\{M_i \leq c_n, M_{i+q_n:i+q_n+j} \leq c_n\} - \mathbf{P}\{M_i \leq c_n\}\mathbf{P}\{M_j \leq c_n\}| \to 0 \quad \text{as } n \to \infty,$$

where the maximum is taken over all $i, j \geq q_n$ such that $i+q_n+j \leq n$ and $M_{n:m} := \max\{Y_n, \ldots, Y_m\}$ with $M_n := M_{1:n}$. This causes maxima over separated groups of points to become increasingly close to being independent as their separation increases at an appropriate rate.

The AIM condition provides enough long-range independence to ensure the limit distribution of linearly normalised maxima is GEV as is shown in Theorem 2.4. For details see Hsing (1987), Hsing et al. (1988) and Leadbetter et al. (1983).

Theorem 2.4 (Unified Extremal Types Theorem for stationary sequences)**.** *Given a stationary sequence* $\{Y_n\}$*, define* $M_n := \max\{Y_{1:n}\}$*. If there exist sequences* $a_n \in \Re$ *and* $b_n > 0$ *such that*

$$\mathbf{P}\left\{\frac{M_n - a_n}{b_n} \leq y\right\} \to H(y) \quad \text{as } n \to \infty,$$

for a non-degenerate distribution H *and if* $\{Y_n\}$ *has AIM*$(a_n + b_n y)$*, then* H *is the cumulative distribution function of the Generalised Extreme Value distribution for some parameters* $\mu \in \Re$*,* $\sigma > 0$ *and* $\xi \in \Re$*.*

Theorem 2.4 therefore justifies modelling maxima of separated blocks as GEV as long as the separation between blocks is sufficient for the AIM condition to hold.

It is useful to relate the stationary sequence limit H with the corresponding IID limit G to show how the dependence changes the limit distribution. Recall that Theorem 2.2 relates the IID limit to the marginal distribution F through $F^n(a_n +$

$b_n y) \to G(y)$ as $n \to \infty$. Leadbetter et al. (1983) show that

$$H(y) = G^{\theta}(y),$$

where $\theta \in [0, 1]$ is the *extremal index* defined by

$$\theta := \lim_{n\to\infty} \mathbf{P}\{M_{2:p_n} \leq a_n + b_n y | Y_1 > a_n + b_n y\}, \tag{2.23}$$

for an o(n) sequence $p_n > 0$ and any y. It can be shown that if G is GEV(μ, σ, ξ), then H is GEV($\mu + \sigma(\theta^{\xi} - 1)/\xi, \sigma\theta^{\xi}, \xi$) so that the presence of dependence in the series does not change the shape of the distribution.

Several estimators for the extremal index have been proposed. Perhaps the most intuitive is the *runs estimator* of Smith and Weissman (1994) which is motivated by (2.23). This involves first identifying clusters of threshold exceedances of the threshold $u = a_n + b_n y$ by assuming clusters are consecutive runs of points separated by at least $\kappa = p_n$ observations below u (see Figure 2.4 for an example). The extremal index θ is then estimated as the reciprocal of the average cluster size. For other estimators see Leadbetter (1983) and Ferro and Segers (2003).

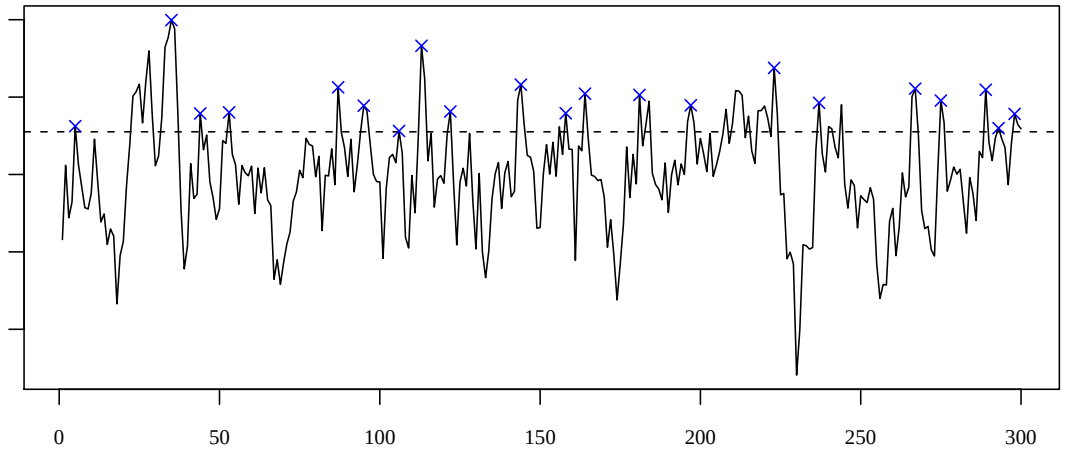


Figure 2.4: Example of a series declustered with the runs method. The clusters of threshold exceedances are each labelled with a cross (×) on the cluster maxima. Clusters were identified as consecutive points separated by at least $\kappa = 4$ observations below the threshold u (−−).

Point process

Having shown that stationary sequences with a long-range independence condition still give a GEV limit for maxima, we now look at how the point process characterisation as affected.

Letting $Y_1, \ldots, Y_n$ be a stationary sequence with marginal distribution function $F \in \mathcal{D}_\xi$, we again define the sequence of point processes P_n on $[0,1] \times \Re$ by

$$P_n := \left\{ \left(\frac{i}{n+1}, \frac{Y_i - a_n}{b_n} \right) \middle| i = 1, \ldots, n \right\},$$

where a_n and b_n are the normalising sequences used for IID variables from F in Theorem 2.2. We also assume that a long-range asymptotic independence condition similar to AIM($a_n + b_n y$) holds and consider the limit process P. It can be shown that, under the appropriate conditions, P is a particular clustered non-homogeneous Poisson process.

Unlike the limit process for IID variables, multiple points in this process may occur at the same time, forming a cluster. The cluster maxima themselves form a non-homogeneous Poisson process with intensity

$$\lambda(s,y) = \theta[1 + \xi y]_+^{-(1+\frac{1}{\xi})},$$

where θ is the extremal index. The distribution of values within a cluster is more complex but the expected number of exceedances of a threshold u per cluster is θ^{-1}, whatever the value of u.

This result shows that the cluster maxima may be modelled as Y_n was in the IID case in Subsection 2.3.2 with θ being absorbed into the location and scale parameters. This means that a dataset must first be declustered and then the point process or peaks over threshold likelihood may be applied. The presence of the extremal index in the intensity above shows that there are fewer cluster

maxima exceedances by a factor of θ than in the IID case.

See Smith (1989) and Davison and Smith (1990) for examples of modelling the peaks over the threshold after declustering.

2.3.4 Non-stationary processes

In many real world datasets, non-stationarity arises from many sources such as seasonal effects and long term trends. Attempts have been made to provide a general extreme value theory for non-stationary sequences (see for example Leadbetter et al. (1983)) but they tend to have limited statistical use. As such, a variety of statistical techniques may be used to account for non-stationarity by allowing variation of the parameters of the models that were introduced for extremes of stationary sequences.

Parametric

The standard approach as outlined in Coles (2001) is to model the parameters as functions of covariates. This can of course be done with any of the previously described models. A simple example would be to add a linear trend to the location parameter of the GEV by modelling annual maxima Y_t in year t as $\text{GEV}(\mu_t, \sigma, \xi)$ where $\mu_t := \beta_0 + \beta_1 t$. Maximum likelihood could then be used to estimate the parameters β_0, β_1, σ and ξ. For examples see Smith (1989) and Davison and Smith (1990).

An alternative approach of Eastoe and Tawn (2009) is to pre-process the whole dataset to remove the bulk of the non-stationarity and then model the residuals as before. This is especially useful with the peaks over threshold model where we are required to select a constant threshold to decide which data values to model; the non-stationarity of the dataset may cause locally extreme values to fall below a threshold while locally typical values are above as shown in Figure 2.5. Pre-

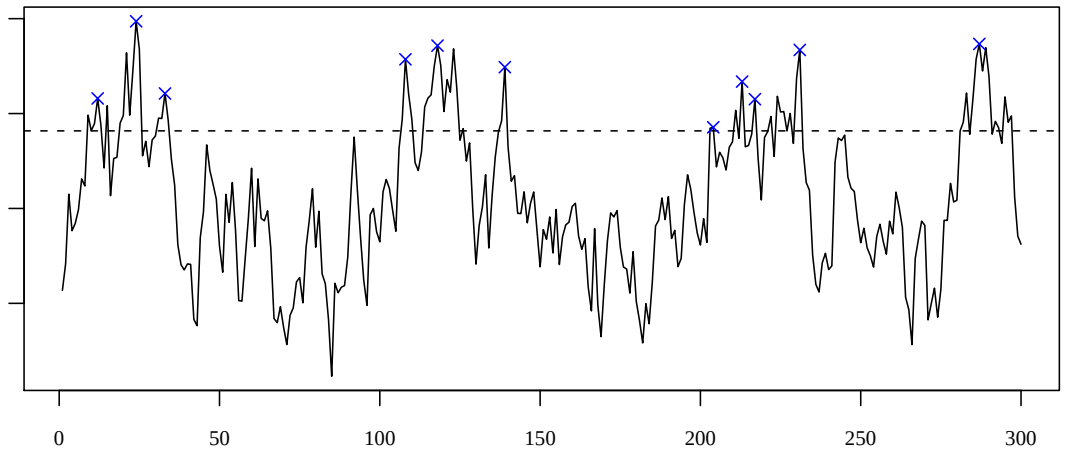


Figure 2.5: Example of the selection of peaks over a constant threshold when the series is non-stationary. The runs method was used with $\kappa = 4$ as in Figure 2.4

processing the data has the potential to remove enough of the non-stationarity to allow a constant threshold through the residuals to select only the data we would deem to be extreme.

Non-parametric

Hall and Tajvidi (2000) propose a non-parametric method for fitting extreme value data with a temporal trend. They use a kernel $K_i(t) := K((t - t_i)/h)$ to weight the log likelihood of each observation (t_i, y_i) and maximise the overall likelihood to obtain parameter estimates for each t. The standard Gaussian density is typically chosen for the kernel K with the bandwidth h chosen with cross-validation.

This method produces parameter estimates which vary smoothly over time to fully account for many types of non-stationarity. It has the disadvantage, however, that estimates can only be made over the range of the data and so the method cannot be used for prediction. For examples of this approach see Davison and Ramesh (2000) and Pauli and Coles (2001).

Stochastic

Another approach to handling non-stationarity is to allow the parameters to move stochastically over time. It is this approach which we focus on extending in chapters 4 and 5 of this thesis.

Smith and Miller (1986) present a state space model where observations are transformed to exponential with a rate that varies stochastically over time. More specifically, they assume a model of the form

$$\begin{aligned} X_{t+1}|\{X_{1:t} = x_{1:t}, Z_{1:t} = z_{1:t}\} &= X_t \rho_{t+1} \phi_{t+1}, \\ Z_t|\{X_{1:t} = x_{1:t}, Z_{1:t-1} = y_{1:t-1}\} &\sim \text{Exponential}(x_t), \\ X_0 &\sim \text{Gamma}(\alpha_0, \beta_0), \end{aligned}$$

where ρ_{t+1} is a constant and ϕ_{t+1} is sampled from a specific Beta distribution. Y_t is observed at each time step t before being transformed with $Z_t = T(Y_t|\theta)$.

The model form is chosen so that the filter and one-step prediction distributions $p(x_t|z_{1:t})$ and $p(x_{t+1}|z_{1:t})$ are all Gamma with parameters that can be updated analytically. The predictive distribution $p(z_{t+1}|z_{1:t})$ has a Pareto form which is also tractable. Smith and Miller (1986) show how specific forms of transformation can be chosen so that extremal data Y_t follow a Gumbel or a Weibull distribution given X_t.

Their particular approach is, by their own admission, confined to a rather narrow class of models for reasons of tractability although their approach is far removed from Gaussian distributions and Markov state of the Kalman Filter. Their method also requires maximum likelihood to estimate the model's hyper-parameters as well as the transformation parameter θ.

Gaetan and Grigoletto (2004) overcome the issue of tractability by using particle filters and smoothers to model a dynamic trend. Their model assumes block

maxima Y_t follow GEV(μ_t, σ_t, ξ_t) where μ_t, $\log \sigma_t$ and ξ_t each follow independent random walks in the state space. While their model is attractive, their method is in our opinion let down by their choice of particle methods; see Section 4.1 for a discussion on this as well as our proposal for an improvement on their model.

2.4 Multivariate Extreme Value Theory

In this section we describe multivariate extreme value theory and methods. For a wider review see Resnick (1987), Kotz and Nadarajah (2000) or Beirlant (2004).

2.4.1 Multivariate extreme value distributions

We now study the extremal properties of a sequence $\{\mathbf{Y}_n = (Y_{n,1}, \ldots, Y_{n,d})'\}$ of IID multivariate random variables with dimension d and common distribution function $\mathbf{P}\{Y_1 \leq y_1, \ldots, Y_d \leq y_d\} = F(y_1, \ldots, y_d)$. To simplify the presentation, we assume our variables have marginal distributions that are Fréchet$(0, 1, 1)$ (commonly referred to simply as Fréchet). This gives $F(y_j) = \exp(-1/y_j)$ for $y_j > 0$ and $\mathbf{P}\{M_{n,j}/n \leq y_j\} = F(y_j)$ where $M_{n,j} := \max\{Y_{1,j}, \ldots, Y_{n,j}\}$. This is not restrictive since results for arbitrary margins may be obtained by applying the probability integral transform to each component: if $Y_j \sim F(y_j)$ then $Y_j := G^{-1}(F(Y_j))$ has distribution function $G(y_j)$.

Componentwise maxima

Unlike the univariate case, there is no natural ordering of a multivariate sample. We are therefore less certain of what constitutes an extreme value. In specific applications there may exist a scalar function of $\mathbf{Y}$ whose large or small values are of interest, in which case the univariate techniques of Section 2.3 may be applied to the transformed variable (see Coles and Tawn (1994)). However, in more general situations we may be interested in the relationship between large values of individual components.

We therefore begin by considering the joint distribution of scaled componentwise maxima. This prompts the following definition: If there exists a non-degenerate

d-dimensional distribution function G with non-degenerate margins such that

$$\mathbf{P}\left\{\frac{M_{n,1}}{n} \leq y_1, \ldots, \frac{M_{n,d}}{n} \leq y_d\right\} \to G(y_1, \ldots, y_d) \quad \text{as } n \to \infty,$$

then G is known as a *multivariate extreme value distribution*. Since $\mathbf{P}\{M_{n,j}/n \leq y_j\} = F(y_j)$, the margins of G are all Fréchet. Also G is max-stable since $G^n(ny_1, \ldots, ny_d) = G(y_1, \ldots, y_d)$.

Pickands (1981) shows that G takes the form

$$G(y_1, \ldots, y_d) = \exp\big(-V(y_1, \ldots, y_d)\big),$$

with

$$V(y_1, \ldots, y_d) := d \int_{\mathcal{S}_d} \max\left\{\frac{w_1}{y_1}, \ldots, \frac{w_d}{y_d}\right\} \mathrm{d}H(w_1, \ldots, w_d), \tag{2.24}$$

where $\mathcal{S}_d := \{\mathrm{w} \in \Re_+^d | \sum_{j=1}^d w_j = 1\}$ is the p-dimensional simplex. The *spectral distribution function* H is an arbitrary distribution on $\mathcal{S}_d$ except that it must satisfy

$$\int_{\mathcal{S}_d} w_j \, \mathrm{d}H(w_1, \ldots, w_d) = \frac{1}{d},$$

that is $\mathbf{E}_H(W_j) = 1/d$ for all $j = 1, \ldots, d$. Therefore, unlike the univariate case, there is no finite parametrisation of the multivariate extreme value distribution.

Some authors have proposed parametric forms for the spectral distribution or for V directly (see Kotz and Nadarajah (2000) for a review). One simple choice is the *multivariate exchangeable logistic* distribution of Gumbel (1960) which has

$$V(y_1, \ldots, y_d) = (y_1^{-\frac{1}{\alpha}} + \ldots + y_d^{-\frac{1}{\alpha}})^{\alpha}, \tag{2.25}$$

for dependence parameter $\alpha \in (0, 1]$. With $\alpha = 1$ the components are independent while a positive association increases as α falls towards 0.

Having assumed up to now the margins are Fréchet, in practice these must be

estimated from the data. Since we know the appropriate distribution for linearly normalised maxima is the GEV, we simply replace the Fréchet margins on G with $\text{GEV}(\mu_j, \sigma_j, \xi_j)$ using the probability integral transform.

For more information on multivariate extreme value distributions see de Haan and Resnick (1977), Pickands (1981) and Resnick (1987).

Point process

As in the univariate case, approaching extreme values via a point process produces a model that allows more data to be used while giving a multivariate extreme value distribution for componentwise maxima. Again working with Fréchet margins, define a sequence of point processes on $\Re_+^d$ by

$$P_n := \left\{ \left(\frac{Y_{i,1}}{n}, \ldots, \frac{Y_{i,d}}{n} \right) \middle| i = 1, \ldots, n \right\}.$$

It was shown by de Haan (1985) that $P_n \to P$ on $\Re_+^d \setminus \{0\}$ where P is a non-homogeneous Poisson process with intensity

$$\lambda(y_1, \ldots, y_d) = \frac{d}{(\sum y_j)^3} h\left(\frac{y_1}{\sum y_j}, \ldots, \frac{y_d}{\sum y_j} \right),$$

where $h(\mathbf{w}) := \mathrm{d}H(\mathbf{w})/\,\mathrm{d}\mathbf{w}$. Note that this is an abuse of notation since the density $h(\mathbf{w})$ may not exist. The intensity is commonly reparametrised in terms of pseudo-radial coordinates $r := \sum y_j$ and $w_j := y_j/r$ giving a more precise intensity measure

$$\lambda(r, w_1, \ldots, w_d)\,\mathrm{d}r\,\mathrm{d}\mathbf{w} = \frac{d}{r^2}\,\mathrm{d}r\,\mathrm{d}H(w_1, \ldots, w_d).$$

This parametrisation has the advantage that the range r is independent of the angular components $\mathbf{w}$.

As in the univariate case, the point process may be used to derive distributions

of interest when they can be written in terms of the counting process $N(A)$. For example, the spectral form of the multivariate extreme value distribution may be derived by noting that

$$\begin{aligned}\mathbf{P}\left\{\frac{M_{n,1}}{n} \leq y_1, \ldots, \frac{M_{n,d}}{n} \leq y_d\right\} &\xrightarrow{n\to\infty} \mathbf{P}\Big\{N\big(A(y_1,\ldots,y_d)\big) = 0\Big\} \\ &= \exp\Big(-\Lambda\big(A(y_1,\ldots,y_d)\big)\Big),\end{aligned}$$

where

$$\begin{aligned}A(y_1,\ldots,y_d) &= \big\{\mathbf{z} \in \Re_+^d \big| y_1 < z_1 \text{ or } \ldots \text{ or } y_d < z_d\big\} \\ &= \left\{\mathbf{z} \in \Re_+^d \middle| \min\left\{\frac{y_1}{z_1}, \ldots, \frac{y_d}{z_d}\right\} < 1\right\}.\end{aligned}$$

Switching to pseudo-radial coordinates,

$$A(y_1,\ldots,y_d) = \left\{(r,\mathbf{w}) \in \Re_+ \times \mathcal{S}_d \middle| \min\left\{\frac{y_1}{w_1}, \ldots, \frac{y_d}{w_d}\right\} < r\right\}$$

so that

$$\begin{aligned}\Lambda\big(A(y_1,\ldots,y_d)\big) &= \iint_{A(y_1,\ldots,y_d)} \lambda(r, w_1, \ldots, w_d)\,\mathrm{d}r\,\mathrm{d}\mathbf{w} \\ &= \int_{\mathcal{S}_d} \int_{\min\left\{\frac{y_1}{w_1},\ldots,\frac{y_d}{w_d}\right\}}^{\infty} \frac{d}{r^2}\,\mathrm{d}r\,\mathrm{d}H(w_1,\ldots,w_d) \\ &= d\int_{\mathcal{S}_d} \min\left\{\frac{y_1}{w_1}, \ldots, \frac{y_d}{w_d}\right\}^{-1} \mathrm{d}H(w_1,\ldots,w_d) \\ &= V(y_1,\ldots,y_d). \qquad (2.26)\end{aligned}$$

The point process limit may also be modelled directly by using all data with range r above a threshold u. Having chosen a particular model for the spectral

distribution H, inference may be performed with the joint likelihood

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}, n_u|H) = p(n_u|H)\, p(\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}|n_u, H)$$
$$\propto \exp\big(-\Lambda(A_u)\big) \prod_{i=1}^{n_u} \lambda(\mathbf{y}_i),$$

where $A_u = \{\mathbf{z} \in \Re_+^d | \sum z_j > u\}$. For further details as well as examples of modelling with the point process limit see Coles and Tawn (1991, 1994), Joe et al. (1992) and de Haan and de Ronde (1998).

2.4.2 Alternative approaches

Limitations of multivariate extreme value distributions

While multivariate extreme value distributions provide a natural generalisation of the univariate techniques to higher dimensions, they can fail to account for variables that are near independence. Ledford and Tawn (1996) discuss the limitations of the above theory for variables that are *asymptotically independent*.

Sibuya (1960) define variables Y_1 and Y_2 with common marginal distribution F to be asymptotically independent if

$$\mathbf{P}\{Y_1 > y | Y_2 > y\} \to 0 \quad \text{as } y \to y^F, \tag{2.27}$$

where y^F is the upper end point of F defined on page 28. If the limit in (2.27) is a constant greater than 0 we say the variables are *asymptotically dependent*. A multivariate analogue is given by de Haan and Resnick (1977).

When the components of a multivariate variable $\mathbf{Y}$ are asymptotically independent, the limiting extreme value distribution has a spectral distribution function H that is degenerate. This causes problems when models are fitted to H as finite samples from an asymptotically independent variable cannot represent this degen-

eracy so biases are introduced. Multivariate extreme value distributions also lack the ability to differentiate between exact and asymptotic independence.

Alternative models

To rectify the deficiencies of the current methods, Ledford and Tawn (1996, 1997, 1998) propose an alternative class of model which naturally incorporates asymptotic dependence and independence. Working again with Fréchet marginal distributions, they show that, under mild regularity conditions, the joint survivor function takes the form

$$\mathbf{P}\{Y_1 > t, Y_2 > t\} \simeq \mathcal{L}(t)\, t^{-\frac{1}{\eta}} \quad \text{for large } t,$$

where $\eta \in (0, 1]$ is the *coefficient of tail dependence* and $\mathcal{L}$ is a *slowly varying* function in that it satisfies

$$\frac{\mathcal{L}(ty)}{\mathcal{L}(t)} \to 1 \quad \text{as } t \to \infty,$$

for any fixed y.

Asymptotic dependence corresponds to $\eta = 1$ with various levels of asymptotic independence given by $\eta \in (0, 1)$. By assuming $\mathcal{L}(t)$ to be a constant and setting $T := \min\{Y_1, Y_2\}$ so that $\mathbf{P}\{T > t\} = \mathbf{P}\{Y_1 > t, Y_2 > t\}$, a sample of Ts above a threshold may be used to estimate η.

An alternative approach is given by Heffernan and Tawn (2004) who assume a single component Y_j is extreme and model the other components conditionally. We can then consider variables that are extreme in only some components rather than the previous methods which assume every component is large. As well as allowing us to consider properties of variables that are only partially extreme, this also allows more data to be used to provide information about the largest extremes.

Chapter 3

New results in particle filtering

In this chapter we present new research in particle filtering. Though the primary aim is to overcome difficulties identified with the algorithms when modelling extreme values, the results may all be used in wider applications.

3.1 Choosing when to re-sample

This section looks at the problem of choosing when to re-sample in a particle filter and proposes a new method that extends the effective sample size thresholding of Liu and Chen (1995).

3.1.1 Motivation

In Subsection 2.1.3 we reviewed basic particle filters commenting that the re-sampling step has advantages and disadvantages and is optional. A strategy for deciding whether to re-sample is given by Liu and Chen (1995) who propose re-sampling when the effective sample size defined in (2.7) falls below a predefined threshold. It is unclear, however, what value this threshold should take.

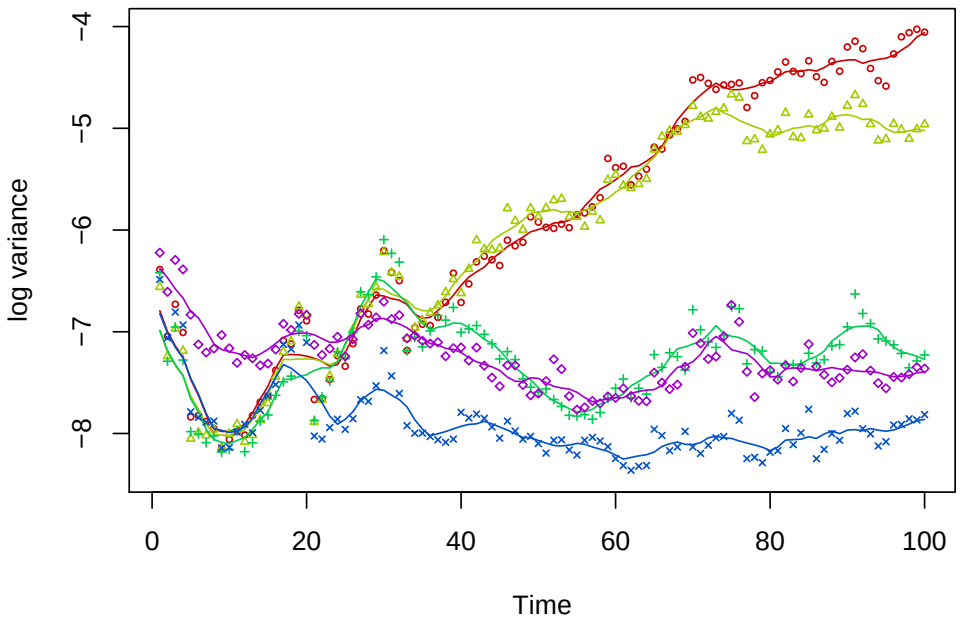


Figure 3.1: Log variances of filter estimates of the state for fixed ESS thresholds of 0 (○), 10 (△), 100 (+), 500 (×) and $1,000 = N$ (◊).

Liu and Chen (1995) concluded from a simulation study that the choice of threshold is unimportant as long as it is greater than 1. A small threshold was therefore preferred to save on computing time. However, this is not always the case as can be seen in Figure 3.1. This shows logged variances of the filters' estimates of the state in the stochastic volatility model (to be defined below in (3.7)) using $1,000$ particles and 100 time steps. The estimates are most variable when we never re-sample (using a threshold of 0) but they are little better when we use a threshold set at 10 (1% of N). While a threshold of 100 is much better here, the value of 500 would be preferred, so long as the extra re-sampling adds little to processing time.

3.1.2 Theory

Working with the auxiliary particle filter given in Algorithm 2.3, our approach is to choose whether to re-sample or not by looking at the effect this will have on the variance of a quantity of interest. Thus we will look at the variance of our estimate of $\mathbf{E}(h(X_t)|y_{1:t})$ given the weighted particles at time $t-1$ for a

known function h. This has been done before by Liu and Chen (1995) for the sequential imputation algorithm which is a special case of the auxiliary filter with proposal distribution $q(x_t|x_{t-1}^{(i)}, y_t) = p(x_t|x_{t-1}^{(i)}, y_t)$ and re-sampling weights $\beta_t^{(i)} \propto p(y_t|x_{t-1}^{(i)})w_{t-1}^{(i)}$. Despite remarking on the gains of residual sampling, they look only at multinomial sampling and compare the resulting variances only for extreme cases. We will calculate similar results for residual sampling and use them to suggest when to re-sample.

Estimating $\mathbf{E}(h(X_t)|y_{1:t})$

Given the weighted filter particles $\{(x_{t-1}^{(i)}, w_{t-1}^{(i)})\}$, our estimate for $\mathbf{E}(h(X_t)|y_{1:t})$ is given by

$$\mu := \sum_{i=1}^{N} h(x_t^{(i)})w_t^{(i)},$$

where $w_t^{(i)}$ are the normalised weights associated with the particles $x_t^{(i)}$ at time t. If we assume that q and $\beta_t^{(i)}$ are chosen so that the filter is fully adapted, we have

$$q(x_t|x_{t-1}^{(i)}, y_t)\, \beta_t^{(i)} \propto g(y_t|x_t)\, f(x_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)}$$

and so, if the particles are re-sampled at time t, the weights $w_t^{(i)}$ will equal $1/N$ after normalisation. If we do not re-sample, the weights are given by

$$w_t^{(i)} \propto \frac{g(y_t|x_t)\, f(x_t|x_{t-1}^{(i)})\, w_{t-1}^{(i)}}{q(x_t|x_{t-1}^{(i)}, y_t)}$$

and so $w_t^{(i)} = \beta_t^{(i)}$. Thus, if we do not re-sample the particles, our estimate becomes

$$\mu_0 := \sum_{i=1}^{N} h(x_t^{(i)})\beta_t^{(i)},$$

where the $x_t^{(i)}$ are sampled from $q(x_t|x_{t-1}^{(i)}, y_t)$. If we do re-sample, this simplifies further to

$$\mu_1 := \frac{1}{N}\sum_{i=1}^{N} h(x_t^{*(i)}),$$

where the $x_t^{*(i)}$ are sampled from $q(x_t|x_{t-1}^{(j_i)}, y_t)$ after the indices j_i are sampled using the $\beta_t^{(i)}$s.

Calculation of variances

To simplify notation we define $\mu^{(i)}$ and $\sigma^{2(i)}$ to be the mean and variance of $h(X_t)$ when X_t is sampled at time t from $q(x_t|x_{t-1}^{(i)}, y_t)$. Then, with no re-sampling, the variance of our estimate, given the weighted particles at $t-1$, is given by

$$\mathrm{Var}(\mu_0) = \sum_{i=1}^{N} \sigma^{2(i)}\beta_t^{(i)2},$$

as the $x_t^{(i)}$s are sampled independently.

For the re-sampled case, we separate the variance into

$$\mathrm{Var}(\mu_1) = \mathbf{E}(\mathrm{Var}(\mu_1|\mathbf{N})) + \mathrm{Var}(\mathbf{E}(\mu_1|\mathbf{N})),$$

where $\mathbf{N} = (N_1, \ldots, N_N)$ are the number of times the particles $(x_{t-1}^{(1)}, \ldots, x_{t-1}^{(N)})$ are re-sampled (via sampling the indices j_i). Then,

$$\begin{aligned}\mathbf{E}(\mu_1|\mathbf{N}) &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{E}_q(h(X_t)|x_{t-1}^{(j_i)}, y_t, \mathbf{N}) \\ &= \frac{1}{N}\sum_{k=1}^{N} \mu^{(k)} N_k,\end{aligned}$$

as N_k of the j_is are equal to k. Similarly,

$$\mathrm{Var}(\mu_1|\mathbf{N}) = \frac{1}{N^2}\sum_{k=1}^{N} \sigma^{2(k)} N_k,$$

since the $x_t^{*(i)}$s are sampled independently, given $\mathbf{N}$.

To complete the calculation we must use the distribution of $\mathbf{N}$ which depends on how we re-sample. We proceed with residual sampling as this is similar to the optimal stratified sampling but has a distribution that is easier to handle (see Subsection 2.1.5 where we review different methods of re-sampling). We therefore have $N_i = \lfloor N\beta_t^{(i)} \rfloor + M_i$, where $\mathbf{M} = (M_1, \ldots, M_N) \sim \text{Multinomial}(r, \gamma^{(1)}, \ldots, \gamma^{(N)})$ (where we recall $r := N - \sum \lfloor N\beta_t^{(i)} \rfloor$ and $\gamma^{(i)} := (N\beta_t^{(i)} - \lfloor N\beta_t^{(i)} \rfloor)/r$).

Defining μ_{r} to be our target estimate if we re-sample with residual sampling, we obtain

$$\begin{aligned}\mathbf{E}(\text{Var}(\mu_{\text{r}}|\mathbf{N})) &= \frac{1}{N^2}\sum_{i=1}^{N} \sigma^{2(i)}\, \mathbf{E}(N_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \sigma^{2(i)}\beta_t^{(i)},\end{aligned}$$

since $\mathbf{E}(N_i) = \lfloor N\beta_t^{(i)} \rfloor + r\gamma^{(i)} = N\beta_t^{(i)}$.

In a similar manner we have

$$\begin{aligned}\text{Var}(\mathbf{E}(\mu_{\text{r}}|\mathbf{N})) &= \frac{1}{N^2}\,\text{Var}\left(\sum_{i=1}^{N} \mu^{(i)} N_i\right) \\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N} \mu^{(i)2}\,\text{Var}(N_i) + 2\sum_{i<j} \mu^{(i)}\mu^{(j)}\,\text{Cov}(N_i, N_j)\right) \\ &= \frac{r}{N^2}\left(\sum_{i=1}^{N} \mu^{(i)2}\gamma^{(i)}(1-\gamma^{(i)}) - 2\sum_{i<j} \mu^{(i)}\mu^{(j)}\gamma^{(i)}\gamma^{(j)}\right) \\ &= \frac{r}{N^2}\left(\sum_{i=1}^{N} \mu^{(i)2}\gamma^{(i)} - \left(\sum_{i=1}^{N} \mu^{(i)}\gamma^{(i)}\right)^2\right) \\ &=: \frac{r}{N^2}\,\mathcal{V}ar_{\gamma}(\mu^{(i)}), \qquad (3.1)\end{aligned}$$

defining $\mathcal{V}ar_{\gamma}(\mu^{(i)})$ as the weighted sample variance of the $\mu^{(i)} = \mathbf{E}_q(h(X_t)|x_{t-1}^{(i)}, y_t)$ with weights $\gamma^{(i)}$.

Hence, putting these together we obtain

$$\mathrm{Var}(\mu_{\mathrm{r}}) = \frac{1}{N}\left(\sum_{i=1}^{N}\sigma^{2(i)}\beta_t^{(i)} + \frac{r}{N}\,\mathcal{V}ar_{\gamma}(\mu^{(k)})\right).$$

If we instead use multinomial sampling, our target estimate, denoted μ_{m}, satisfies the similar formula

$$\mathrm{Var}(\mu_{\mathrm{m}}) = \frac{1}{N}\left(\sum_{i=1}^{N}\sigma^{2(i)}\beta_t^{(i)} + \mathcal{V}ar_{\beta}(\mu^{(k)})\right).$$

Consequently, $\mathrm{Var}(\mu_{\mathrm{m}})$ is likely to be larger than $\mathrm{Var}(\mu_{\mathrm{r}})$ since $r < N$ and $\mathcal{V}ar_{\gamma}(\mu^{(k)})$ should be similar to $\mathcal{V}ar_{\beta}(\mu^{(k)})$.

Choosing when to re-sample

Using the above, we can choose to re-sample only when, for some function of interest h, we predict that the variance of our estimate will be less if we do re-sample. Assuming that relative biases will be minimal, this strategy ensures that the optimal re-sampling decision is made with respect to providing the best filter estimate of h at the following time step. It is hoped that by minimising only the variance at the next step, the estimates for all time steps will be optimal, though this is by no means guaranteed.

Following this strategy, for residual sampling, we should re-sample when

$$\sum_{i=1}^{N}\left(\beta_t^{(i)2} - \frac{\beta_t^{(i)}}{N}\right)\sigma^{2(i)} > \frac{r}{N^2}\,\mathcal{V}ar_{\gamma}(\mu^{(k)}). \tag{3.2}$$

As long as we know or can approximate $\mu^{(i)}$ and $\sigma^{2(i)}$ from the sampling distribution $q(x_t|x_{t-1}^{(i)}, y_t)$, we can calculate this at every step to make our decision. This, however, could require a lot of time to calculate at each step, time which could be better used by increasing the sample size, so a simplification would be useful.

In many situations (such as with our first example below, the AR(1) model), $\sigma^{2(i)} = \mathrm{Var}_q(h(X_t)|x_{t-1}^{(i)}, y_t)$ will equal a constant σ^2 for all i, or at least approximately so, which greatly simplifies the left hand side to give

$$\mathcal{V}ar(\beta_t^{(k)})\,\sigma^2 > \frac{r}{N^3}\,\mathcal{V}ar_\gamma(\mu^{(k)}), \tag{3.3}$$

where we define the sample variance of $\beta_t^{(i)}$ as

$$\mathcal{V}ar(\beta_t^{(k)}) := \frac{1}{N}\sum_{i=1}^{N}\left(\beta_t^{(i)} - \frac{1}{N}\right)^2$$

and $\mathcal{V}ar_\gamma(\mu^{(k)})$ is given in (3.1). Since the effective sample size is defined in (2.7) as $\mathrm{ESS}(\beta_t) = 1/\sum_{i=1}^{N}\beta_t^{(i)2}$, we can rearrange (3.3) to give an $\mathrm{ESS}(\beta_t)$ threshold, choosing to re-sample when

$$\mathrm{ESS}(\beta_t) < N\left(1 + \frac{r}{N}\frac{\mathcal{V}ar_\gamma(\mu^{(k)})}{\sigma^2}\right)^{-1}. \tag{3.4}$$

The formula for multinomial sampling is this with $r = N$ and $\gamma = \beta_t$, but since r is less than N we see that the $\mathrm{ESS}(\beta_t)$ threshold is higher for residual sampling. Thus re-sampling is more favourable when using a stratified re-sampling scheme since the extra variation it adds has been reduced. Also, as noted by Liu and Chen (1995), if $N\beta_t^{(i)}$ are all integers so that $r = 0$ or if $\mu^{(k)}$ are also constant, it is always favourable to re-sample.

Calculating this threshold still requires $\mathcal{V}ar_\gamma(\mu^{(k)})$ but there will often be ways this can be approximated which will still lead to the same re-sampling decision most of the time. For situations where $\sigma^{2(i)}$ varies greatly, the optimal threshold (3.2) may still be used.

3.1.3 Simulation studies

We will now show how this theory can be used in practice to improve parameter estimates from the filter.

AR(1) **model**

We first consider the simple one dimensional autoregressive model given by

$$X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} \sim \mathcal{N}(\phi x_t, \nu^2)$$
$$Y_t|\{X_{1:t} = x_{1:t}, Y_{1:t-1} = y_{1:t-1}\} \sim \mathcal{N}(x_t, \tau^2)$$

and a Gaussian prior for X_0. This is a special case of the linear-Gaussian model (2.4) with $F = \phi$, $Q = \nu^2$, $G = I$ and $R = \tau^2$.

The optimal propagation density for the auxiliary filter can be shown to be

$$q(x_t|x_{t-1}^{(i)}, y_t) = \mathcal{N}\left(x_t \middle| \frac{\phi x_{t-1}^{(i)}\tau^2 + y_t\nu^2}{\nu^2 + \tau^2}, \frac{\nu^2\tau^2}{\nu^2 + \tau^2}\right),$$

with optimal re-sampling weights

$$\beta_t^{(i)} \propto \mathcal{N}(y_t|\phi x_{t-1}^{(i)}, \nu^2 + \tau^2)\, w_{t-1}^{(i)},$$

where $\mathcal{N}(x|\mu, \sigma^2)$ is the density of $\mathcal{N}(\mu, \sigma^2)$ evaluated at x. This gives a filter that is fully adapted so our theory's assumptions are valid.

Assuming we are most interested in the value of the state itself, we will choose whether to re-sample by minimising the variance of $\mathbf{E}(X_t|y_{1:t})$. Therefore, noting $\sigma^{2(i)}$ is constant, we should re-sample using residual sampling when $\mathrm{ESS}(\beta_t)$ is below threshold (3.4) which simplifies to

$$N\left(1 + \frac{r}{N}\frac{\phi^2\tau^2}{\nu^2(\nu^2 + \tau^2)}\,\mathcal{V}ar_\gamma(x_{t-1}^{(k)})\right)^{-1}. \tag{3.5}$$

To gain greater insight into how this threshold is affected by the model parameters, we will approximate $\mathcal{V}ar_\gamma(x_{t-1}^{(k)})$. We first note that as long as $\phi^2 < 1$, the stationary distribution of the state is $X_\infty \sim \mathcal{N}(0, \nu^2/(1-\phi^2))$. Then, given an observation y, $X_\infty|\{Y = y\}$ has variance $\nu^2\tau^2/(\nu^2 + \tau^2(1-\phi^2))$. We use this as a rough approximation of $\mathcal{V}ar_\gamma(x_{t-1}^{(k)}) \simeq \mathrm{Var}(X_{t-1}|y_{1:t-1})$ assuming that the most recent observation accounts for most of the variation. This should be true when ϕ is small but will hopefully still give a reasonable re-sampling decision when it is not. This leads to the approximate threshold of

$$N\left(1 + \frac{r}{N}\frac{\phi^2\tau^2}{\nu^2(\nu^2+\tau^2)}\frac{\nu^2\tau^2}{\nu^2+\tau^2(1-\phi^2)}\right)^{-1}$$

$$= \frac{N(k-\phi^2)}{k-(1-\frac{r}{Nk})\phi^2}, \tag{3.6}$$

where $k = (\nu^2+\tau^2)/\tau^2$. From this we can see that the attractiveness of re-sampling depends largely on the ratio of system to observation noise and the dependence between states.

We now compare the performance of the adapted filter when these re-sampling thresholds and residual sampling are used. We simulated a dataset and ran the filter 200 times on the same dataset for each re-sampling scheme to measure the variance of its estimates. As well as estimating $\mathbf{E}(X_t|y_{1:t})$, we compare the filters' estimate of the upper 2.5% quantile and the posterior probability of the state being greater than its true value, $\mathbf{P}\{X_t > x_t|y_{1:t}\}$, for each time t. This allows us to see how the filters perform for a variety of inferences.

Figure 3.2 shows log variances of the posterior expectation estimates of the current state using $\nu^2 = \tau^2 = 1$, $\phi = 0.9$ and a $\mathcal{N}(0,1)$ prior for 25 time steps. $1,000$ particles were used and the exact threshold (3.5) and approximation (3.6) are compared to various fixed thresholds between 0 (never re-sampling) and $1,000$ (always re-sampling). Table 3.1 shows the average state estimate variances over 100 time steps and also those for the quantile and probability estimates.

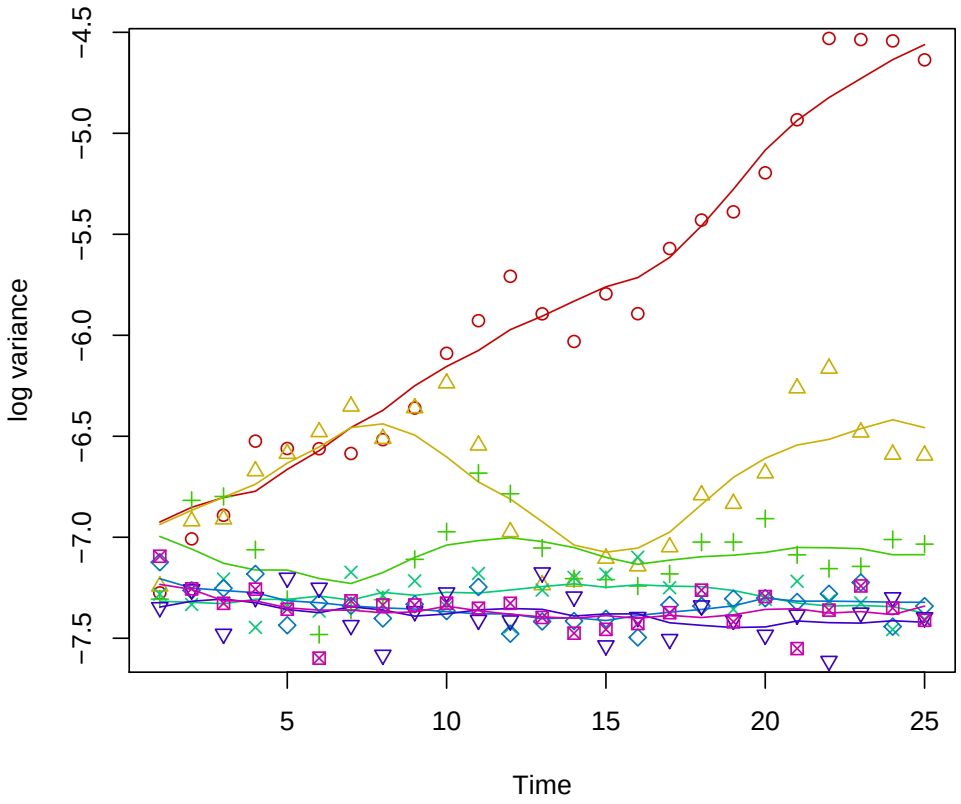


Figure 3.2: Log variances of filter estimates of $\mathbf{E}(X_t|y_{1:t})$ for fixed ESS thresholds of 0 ($\circ$), 250 ($\triangle$), 500 ($+$), 750 ($\times$) and $1,000 = N$ ($\Diamond$) as well as exact (∇) and approximate ($\boxtimes$) variable thresholds.

As we can see from the plot, the estimates' variance increases exponentially with time if we never re-sample, and re-sampling even occasionally can dramatically reduce this. Always re-sampling does much better here but using a threshold just below N seems to be the best fixed strategy. Our new exact threshold seems to do as well as any fixed threshold and the approximation which avoids calculating $\mathcal{V}ar_\gamma(x_{t-1}^{(k)})$ is very similar. From Table 3.1 we can see that, though our method was constructed around estimating the state, it gives good estimates of the upper 2.5% quantile and $\mathbf{P}\{X_t > x_t|y_{1:t}\}$ too, showing it gives an accurate filter for a variety of uses.

We now look at how the threshold required changes with the amount of dependence in the model. Table 3.2 compares the average state estimate variances over 100 time steps with the model using $\nu^2 = 100$ and $1/100$ rather than 1.

With $\nu^2 = 100$ there is little dependence between each state so X_{t-1} tells us little

ESS threshold	Var($\hat{\text{state}}$) ($\times 10^{-4}$)	Var($\hat{\text{quantile}}$) ($\times 10^{-3}$)	Var($\hat{\text{probability}}$) ($\times 10^{-4}$)
0	969.5	130.1	271.8
100	23.2	16.6	6.6
250	12.4	8.9	3.5
500	8.4	5.8	2.4
750	7.1	4.5	2.0
900	6.6	4.3	1.8
1,000	6.8	4.3	1.9
exact	6.7	4.3	1.8
approximate	6.7	4.4	1.9

Table 3.1: Average variances of filter estimates of the current state, upper 2.5% quantile and probability of exceeding the true state for 100 time steps using various ESS thresholds. Fixed ESS thresholds are compared to our exact variable threshold (3.5) and its approximation (3.6).

about the current state X_t. The optimal strategy is therefore to re-sample the $x_{t-1}^{(i)}$ all the time to sample more particles where the new observation y_t tells us the state is likely to be. As can be seen above, our new thresholds do as well as this optimal strategy. From the approximate threshold (3.6) we can see that decreasing ϕ^2 which also reduces the dependence would give the same effect.

Conversely, when $\nu^2 = 1/100$ the states are highly dependent so new particles are proposed near to their parent. This causes the filter to struggle to move about the state space so re-sampling is less preferable as we lose particles to propagate forward. The best threshold is therefore somewhere between 0 and N and we can

ESS threshold	$\nu^2 = 100$ Var($\hat{\text{state}}$) ($\times 10^{-4}$)	$\nu^2 = 1/100$ Var($\hat{\text{state}}$) ($\times 10^{-5}$)
0	13.9	39.1
100	13.9	14.6
250	14.0	9.8
500	13.9	6.8
750	11.7	7.5
900	10.3	8.7
1,000	9.9	13.8
exact	10.0	8.3
approximate	9.9	8.5

Table 3.2: Average variances of filter estimates of the current state over 100 time steps with $\nu^2 = 100$ and $\nu^2 = 1/100$.

see that 500 gave the smallest estimate variance here. Our new thresholds do not do as well as this but they are still better than most of the fixed thresholds which we could have chosen.

We have seen with this simple model that the amount of re-sampling required for the filter to perform well varies with the amount of dependence in the state. Our new re-sampling strategy follows this and can greatly improve the accuracy of a filter run with an arbitrarily selected threshold.

Stochastic volatility model

We now look at the stochastic volatility model given by

$$\begin{aligned} X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} &\sim \mathcal{N}(\phi x_t, \nu^2), \\ Y_t|\{X_{1:t} = x_{1:t}, Y_{1:t-1} = y_{1:t-1}\} &\sim \mathcal{N}(0, \beta^2 e^{x_t}), \end{aligned} \tag{3.7}$$

where $\phi \in (0, 1)$ and ν, $\beta > 0$. It provides a way of generalising the Black-Scholes option pricing model by allowing clustering of the volatility of returns on assets (see Hull and White (1987) for details).

The implementation of our particle filter is given in Appendix A.2. While Pitt and Shephard (1999a) show how rejection sampling may be used to produce an adapted filter, we choose not to do this to contrast with the AR(1) model above.

Since the purpose of fitting this model in practice would be to estimate the volatility $\beta e^{x_t/2}$, we could choose to re-sample by minimising the variance of this rather than of the state itself. However, to see what difference if any this will make to the final filter estimates, we will apply our theory with both $h(x) = x$ and $h(x) = \beta e^{x/2}$ to compare the two approaches.

Firstly, with the aim of minimising the variance of the state estimate, we can use the optimal inequality (3.2) with the mean and variance of our proposal density

to decide when to re-sample. Unlike the AR(1) model, this does not reduce to a threshold of ESS(β_t) since the proposal variance for each particle is different. However, approximating $x_{t-1}^{(i)}$ by $\mathbf{E}(X_{t-1}|y_{1:t-1})$ in the formula for the variance gives a constant value σ^2 which leads to the threshold (3.4). $\mathcal{V}ar_\gamma(\mu_t^{(k)})$ can be well approximated by using $\mu_t^{(i)} \simeq \phi x_{t-1}^{(i)}$ which leads to

$$\mathcal{V}ar_\gamma(\mu_t^{(k)}) \simeq \phi^2\, \mathcal{V}ar_\gamma(x_{t-1}^{(k)}) \simeq \phi^2 \operatorname{Var}(X_{t-1}|y_{1:t-1}).$$

This gives us the simple threshold of

$$N\left(1 + \frac{r}{N}\frac{\phi^2 \operatorname{Var}(X_{t-1}|y_{1:t-1})}{\sigma^2}\right)^{-1}$$

that depends only on the variance of the filtered state which would often have been calculated anyway.

Now, to create an alternative re-sampling rule based on the volatility, we need the mean and variance of $h(X_t) = \beta e^{X_t/2}$ under $q(x_t|x_{t-1}^{(i)}, y_t) = \mathcal{N}(x_t|\tilde{\mu}_t^{(i)}, \tilde{\sigma}_t^{2(i)})$. By noting $e^{X_t/2}$ has a log-normal distribution, these can be shown to be

$$\begin{aligned}
\mu_t^{(i)} &= \mathbf{E}_q(h(X_t)|x_{t-1}^{(i)}, y) = \beta \exp\left(\frac{\tilde{\mu}_t^{(i)}}{2} + \frac{\tilde{\sigma}_t^{2(i)}}{8}\right) \\
\sigma_t^{2(i)} &= \operatorname{Var}_q(h(X_t)|x_{t-1}^{(i)}, y) = \beta^2\left(\exp\left(\frac{\tilde{\sigma}_t^{2(i)}}{4}\right) - 1\right)\exp\left(\tilde{\mu}_t^{(i)} + \frac{\tilde{\sigma}_t^{2(i)}}{4}\right).
\end{aligned}$$

Now we can use the optimal inequality (3.2) to decide when to re-sample. This could, however, be a significant calculation for large N which may be better spent increasing the sample size so an approximation would be useful.

For this we will start by approximating $\sigma_t^{2(i)}$ by a constant σ_t^2 as we have seen that this gives us a simpler threshold for ESS(β_t). This we do by replacing each $x_{t-1}^{(i)}$ in the formulae for $\tilde{\mu}_t^{(i)}$ and $\tilde{\sigma}_t^{2(i)}$ by our filter estimate of $\mathbf{E}(X_{t-1}|y_{1:t-1})$. This leaves $\mathcal{V}ar_\gamma(\mu_t^{(k)})$ which we would like to write in terms of our filter estimate of the state

variance. For this we use $\tilde{\mu}_t^{(i)} \simeq \phi x_{t-1}^{(i)}$ and $\sigma_t^{2(i)} \simeq \sigma_t^2$ to write

$$\mu_t^{(i)} \simeq \beta \exp\left(\frac{\phi x_{t-1}^{(i)}}{2} + \frac{\tilde{\sigma}_t^2}{8}\right)$$

and with the delta method and $\mathcal{V}ar_\gamma(x_{t-1}^{(k)}) \simeq \mathrm{Var}(X_{t-1}|y_{1:t-1})$ we obtain

$$\mathcal{V}ar_\gamma(\mu_t^{(k)}) \simeq \mathrm{Var}(X_{t-1}|y_{1:t-1})\frac{\beta^2}{4}\exp\left(\phi\,\mathbf{E}(X_{t-1}|y_{1:t-1}) + \frac{\tilde{\sigma}_t^2}{4}\right).$$

Putting this together we get a simple threshold of

$$N\left(1 + \frac{r}{N}\frac{\mathrm{Var}(X_{t-1}|y_{1:t-1})}{(\exp(\sigma_t^2/4) - 1)}\right)^{-1}. \tag{3.8}$$

We will now compare the filter estimates of the volatility $\beta e^{x_t/2}$ when our new re-sampling rules are used. As with the AR(1) model we simulated a dataset and ran each filter 200 times to measure the variance of their estimates. We use $1,000$ particles with model parameters $\phi = 0.9720$, $\nu = 0.178$ and $\beta = 0.5992$ taken from Pitt and Shephard (1999a).

Table 3.3 shows the average variances of the state and volatility estimates over 100 time steps for our exact and approximate thresholds for the state, our exact in-

ESS threshold	$\mathrm{Var}(\hat{\text{state}})$ $(\times 10^{-4})$	$\mathrm{Var}(\hat{\text{volatility}})$ $(\times 10^{-5})$
0	51.8	87.5
100	7.5	9.0
250	4.5	5.2
500	4.0	4.4
750	3.7	4.1
900	4.8	5.3
1,000	7.5	8.8
exact (state)	4.0	4.4
approx. (state)	4.2	4.7
exact (volatility)	4.4	4.6
approx. (volatility)	6.0	6.9

Table 3.3: Average variances of filter estimates of the state and volatility over 100 time steps.

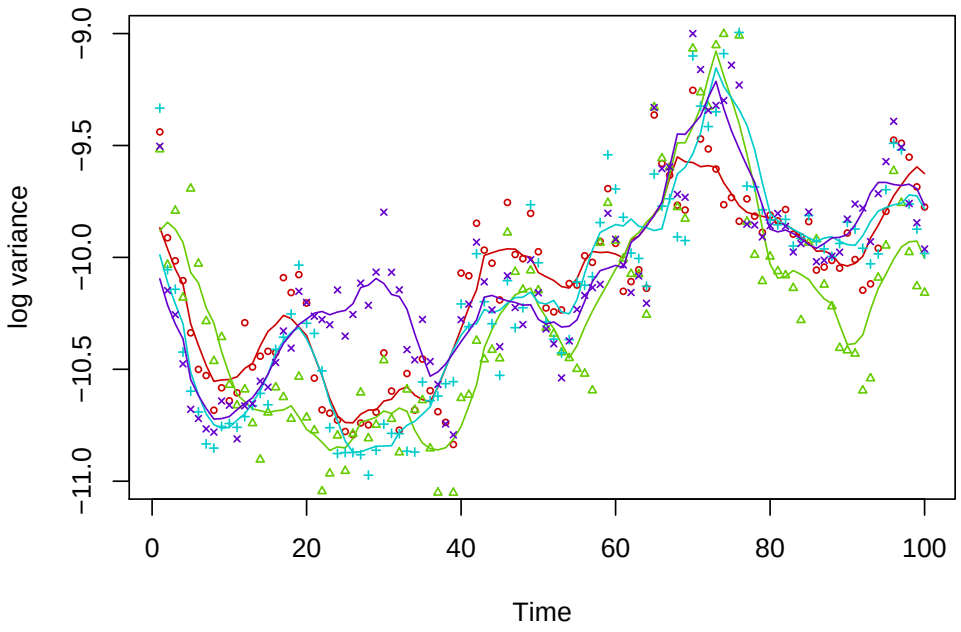


Figure 3.3: Log variances of filter estimates of the volatility for fixed ESS thresholds of 500 (○) and 750 (△) as well as exact variable thresholds for the state (+) and the volatility (×).

equality and approximate threshold for the volatility and various fixed thresholds. Figure 3.3 plots the volatility variances over time for a subset of these.

We first notice that none of our methods do as well as a fixed threshold at 75% of N though both of the exact methods are very close. Also it seems that too many simplifications were made for the volatility based rule to use only filter estimates since the variance here is significantly larger. We also note that the filters which aimed at producing an accurate state gave slightly better volatility estimates than those aimed at the volatility. While this seems contradictory we recall that our theory aims at reducing the variance of an estimate at the next time step only so this does not necessarily give the best estimates over all future times. It may be that using $h(x) = x$ is sufficient for the filter to run efficiently and thus produce accurate estimates for any property of interest.

Bearings-only tracking

Finally we look at the bearings-only tracking problem. Many writers including Gordon et al. (1993) and Pitt and Shephard (1999a) have implemented particle filters for simple tracking models but here we most closely follow the formulation of Fearnhead (1998). To avoid confusion between the x-y plane and the state and observation vectors, we take $x_t = (u_t, v_t, \dot{u}_t, \dot{v}_t)'$ to be the position and velocity in Cartesian coordinates of our target at time t. The state density assumes acceleration is due to white noise and is given by

$$X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} \sim Fx_t + \Gamma\,\mathcal{N}(0, \nu^2 \mathrm{I}_2), \tag{3.9}$$

where

$$F = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We observe only noisy measurements of the target's bearing from the origin and thus we take our likelihood to be

$$Y_t|\{X_{1:t} = x_{1:t}, Y_{1:t-1} = y_{1:t-1}\} \sim \mathcal{N}(\mathrm{atan2}(v_t, u_t), \tau^2), \tag{3.10}$$

where

$$\mathrm{atan2}(v, u) := \begin{cases} \tan^{-1}(v/u) & v \geq 0, u > 0 \\ \pi - \tan^{-1}(-v/u) & v \geq 0, u < 0 \\ \pi/2 & v > 0, u = 0 \\ -\mathrm{atan2}(-v, u) & v < 0 \end{cases}.$$

Our implementation of the particle filter is given in Appendix A.3. Our proposal

distribution is a modification of that of Fearnhead (1998) and is based on a change to polar coordinates (r, α).

Now that the particle filter is defined, we look at choosing when to re-sample. The obvious choice for a scalar quantity of interest is the range of the particle since the bearing is observed directly. To apply our theory we therefore need the mean and variance of the range r_t under q. Conditional on α_t it can be shown that

$$\mathbf{E}_q(R_t|\alpha_t, x_{t-1}^{(i)}, y_t) = m^{(i)}(\alpha_t) := \nu \left(s_t^{(i)} + \left(s_t^{(i)} + \frac{\Phi(s_t^{(i)})}{\phi(s_t^{(i)})} \right)^{-1} \right)$$

and

$$\mathrm{Var}_q(R_t|\alpha_t, x_{t-1}^{(i)}, y_t) = v^{(i)}(\alpha_t) := \nu^2 \frac{\left(\left(s_t^{(i)} + \frac{\Phi(s_t^{(i)})}{\phi(s_t^{(i)})} \right) \left(s_t^{(i)} + 2\frac{\Phi(s_t^{(i)})}{\phi(s_t^{(i)})} \right) - 1 \right)}{\left(s_t^{(i)} + \frac{\Phi(s_t^{(i)})}{\phi(s_t^{(i)})} \right)^2},$$

using $s_t^{(i)}$ defined in (A.4) and with Φ and ϕ here referring to the cdf and pdf of $\mathcal{N}(0,1)$ respectively.

To marginalise out α_t we use the approximations

$$\mathbf{E}_q(R_t|x_{t-1}^{(i)}, y_t) = \mathbf{E}_q(m^{(i)}(\alpha_t)) \simeq m^{(i)}(\mu_{\alpha_t}^{(i)})$$

and

$$\begin{aligned} \mathrm{Var}_q(R_t|x_{t-1}^{(i)}, y_t) &= \mathbf{E}_q(v^{(i)}(\alpha_t)) + \mathrm{Var}_q(m^{(i)}(\alpha_t)) \\ &\simeq v^{(i)}(\mu_{\alpha_t}^{(i)}) + \sigma_{\alpha_t}^{2(i)} \left(m^{(i)\prime}(\mu_{\alpha_t}^{(i)}) \right)^2, \end{aligned}$$

where $\mu_{\alpha_t}^{(i)}$ and $\sigma_{\alpha_t}^{2(i)}$ are the mean and variance of the bearing respectively. These are used with the optimal inequality (3.2) to decide whether to re-sample.

A simulated path of length 24 is shown in Figure 3.4. The target passes by the origin and moves away to the south. An independent normal prior was

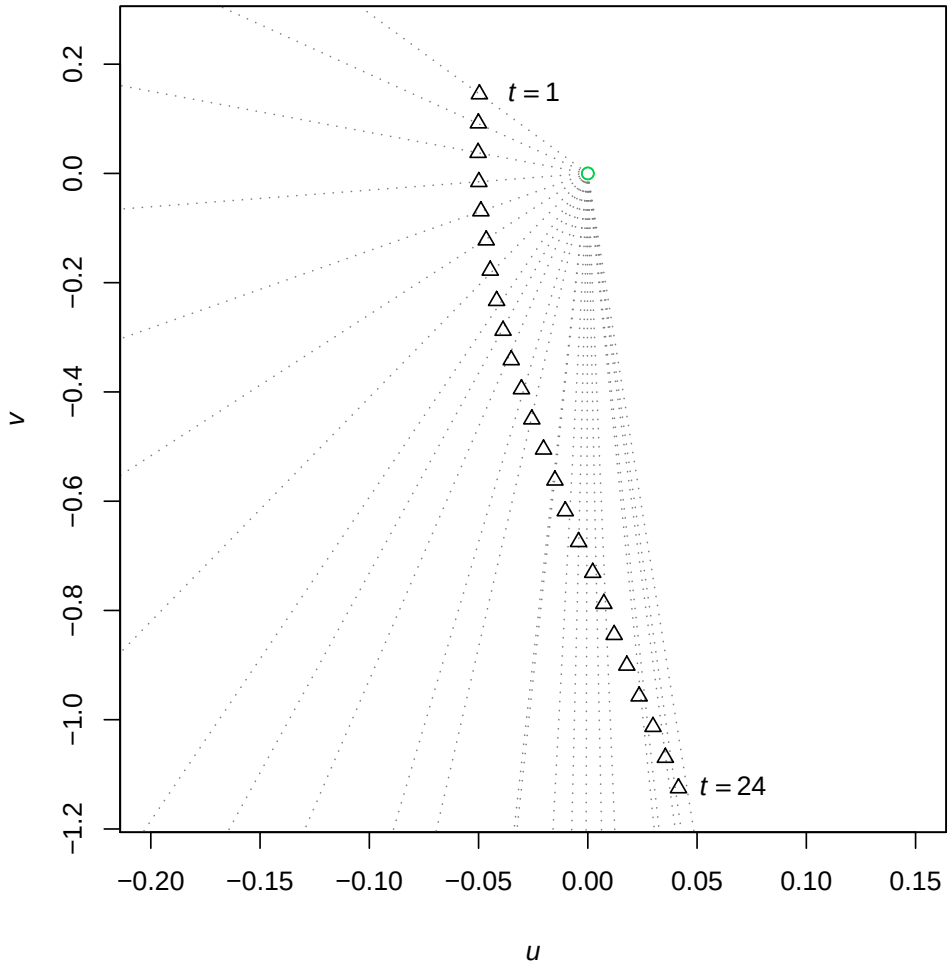


Figure 3.4: Simulated path of an object in the u-v plane over 24 time steps with the observed noisy bearings made from the origin.

used with means $(-0.05, 0.2, 0.001, -0.055)'$ and standard deviations $(0.05, 0.03, 0.0005, 0.001)'$ taken from Pitt and Shephard (1999a). The initial value of the state was taken to be the mean of the prior with $\nu = 0.001$ and $\tau = 0.005$, also from Pitt and Shephard (1999a).

Since the bearing is measured accurately, we will judge the performance of the filters by their estimate of the range. Figure 3.5 compares the variances of these estimates for filters using various fixed and variable thresholds. It shows how it is easy to estimate the range when the bearing changes rapidly but harder when the target moves away. We can see that choosing to minimise the variance of the range estimate for the next time step causes the filter to perform badly overall, worse than most fixed thresholds. This is because the range is unobserved so the

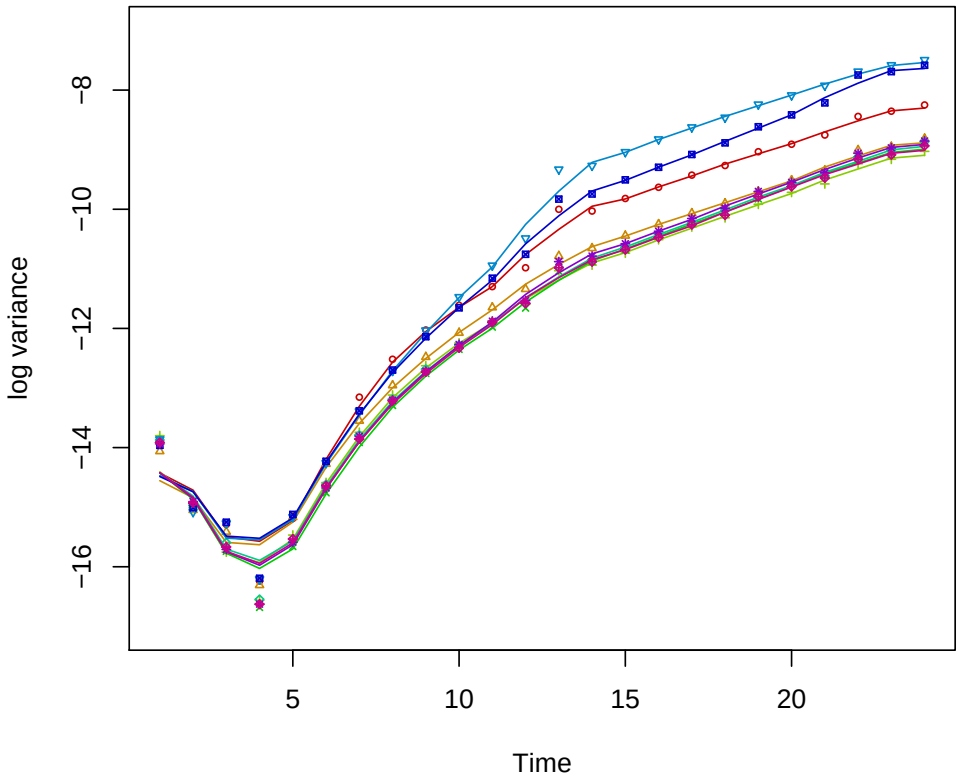


Figure 3.5: Log variances of filter estimates of the range for fixed ESS thresholds of 100 (○), 250 (△), 500 (+), 750 (×) and 900 (◇) as well as variable thresholds for the range (▽), Cartesian position (⊠), Cartesian velocity (✳) and bearing (⋄).

spread of the overall range density and hence the particles is large. Thus sampling between them adds a lot of variation to the overall estimate so we rarely want to re-sample, even if the weights are very uneven.

The SV model above showed that applying the theory to the state rather than a transformation of it could improve efficiency. However, since the state is now multidimensional, our theory no longer applies directly as we implicitly assumed that $h(x)$ was scalar. If we remove this constraint the variance formulae still hold where $\mu^{(i)}$ becomes a vector of expectations and we write $\Sigma^{(i)}$ as the variance matrix of $h(X_t)$ under q. Since we now have two variance matrices to compare and can no longer choose the smallest, we propose minimising the trace of the variance matrix. Analogously to (3.2), this means we should re-sample when

$$\sum_{i=1}^{N}\left(\beta_t^{(i)2}-\frac{\beta_t^{(i)}}{N}\right)\mathrm{tr}(\Sigma^{2(i)})>\frac{r}{N^2}\,\mathrm{tr}(\mathcal{V}ar_\gamma(\mu^{(k)})). \tag{3.11}$$

We now use this equation to minimise the variance of estimates of $h(x_t) = (u_t, v_t)'$, the position in Cartesian coordinates. The mean of $h(X_t)$ under q can be estimated using the means of the range and bearing as

$$\mu_t^{(i)} = \begin{pmatrix} \mu_{r_t}^{(i)} \cos(\mu_{\alpha_t}^{(i)}) \\ \mu_{r_t}^{(i)} \sin(\mu_{\alpha_t}^{(i)}) \end{pmatrix}.$$

The variance matrix $\Sigma_t^{(i)}$ can be estimated by calculating the variance of $(r_t, \alpha_t)'$ and transforming to $(u_t, v_t)'$ with the delta method as follows:

To first construct the variance matrix in polar coordinates we need the covariance between r_t and α_t. Dropping the dependence on $x_{t-1}^{(i)}$ and y_t from the notation, we write $\mathrm{Cov}_q(R, \alpha) = \mathbf{E}_q(f(\alpha))$ with $f(\alpha) := (\alpha - \mathbf{E}_q(\alpha))\,\mathbf{E}_q(R|\alpha)$ and making a Taylor approximation of $f(\alpha)$ about $\mathbf{E}_q(\alpha)$ we get

$$\mathrm{Cov}_q(R, \alpha) \simeq \frac{\mathrm{Var}_q(\alpha)}{2} f''(\mathbf{E}_q(\alpha)).$$

We can now approximate the variance in Cartesian coordinates using the delta method to get

$$\mathrm{Var}_q(U_t, V_t) \simeq \nabla g(\mathbf{E}_q(R_t, \alpha_t))' \,\mathrm{Var}_q(R_t, \alpha_t) \nabla g(\mathbf{E}_q(R_t, \alpha_t)),$$

where

$$\nabla g(r, \alpha) = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -r\sin(\alpha) & r\cos(\alpha) \end{pmatrix}.$$

The mean and variance can then be used in (3.11) above to select whether to re-sample.

Alternatively, we can try to minimise the velocity estimates' variance by using $h(x_t) = (\dot{u}_t, \dot{v}_t)'$. Since the velocity terms are sampled as the difference in position

at time $t-1$ and t, the variance remains the same and the mean is now given by

$$\mu_t^{(i)} = \begin{pmatrix} \mu_{r_t}^{(i)} \cos(\mu_{\alpha_t}^{(i)}) - u_{t-1}^{(i)} \\ \mu_{r_t}^{(i)} \sin(\mu_{\alpha_t}^{(i)}) - v_{t-1}^{(i)} \end{pmatrix}.$$

Finally, we will also compare the results for a filter minimising the variance of the bearing estimate. This is easiest to apply since the bearing is sampled from a normal distribution whose mean and variance are already known.

The results using these thresholds can also be seen in Figure 3.5. Table 3.4 below compares the average variances over the 24 time steps. We can see that using the Cartesian position gave a slight improvement over using the range but is still significantly worse than the other fixed thresholds.

Optimising for the velocity, however, gave range estimates comparable with the best fixed thresholds. This could be because we model acceleration as white noise which therefore acts first on the velocity of the target rather than its position. The variance of the velocity is often smaller after re-sampling while that of the position is larger simply because the spread of possible range values is large and we are sampling between them. Since the velocity drives the process it is better to ensure these estimates are most accurate.

ESS threshold	Var(r$\hat{\text{a}}$nge) ($\times 10^{-5}$)
0	88.1
100	6.5
250	3.6
500	2.9
750	3.1
900	3.2
1,000	3.4
range	14.0
Cartesian position	11.1
Cartesian velocity	3.4
bearing	3.1

Table 3.4: Average variances of filter estimates of the range over 24 time steps using various ESS thresholds.

As a final note, we can also see that optimising for the bearing estimate gives very good results. Since the bearing is observed accurately, there is little need to optimise these estimates for practical purposes but still the differences in the estimates' accuracy can be used to guide the re-sampling decision. This probably performs better than the range or position filters since they are both affected by the lack of knowledge about the range. Focusing only on variables unrelated to the unobserved range, we see the bearing gives similar results to the velocity and the best fixed thresholds confirming the observation made with the SV model that the choice of $h(x)$ makes little difference within this context.

3.1.4 Conclusion

We have proposed a new method of choosing when to re-sample using the auxiliary particle filter. Rather than re-sampling when the effective sample size is below a fixed threshold, we choose to re-sample by minimising the variance of a quantity of interest. This effectively leads to a threshold which varies with the amount of information supplied by each observation. We have shown that the quality of filter estimates is sensitive to the choice of threshold and our method avoids the problem of selecting this.

Our simulation studies have shown how the optimal threshold changes with the amount of dependence in the model and our method follows this. When the dimension of the state is greater than 1 we found that our method can be sensitive to the choice of quantity of interest. When this is chosen to be an unobserved component or a transformation of it the filter can re-sample too infrequently. However, when this is not the case the performance of the filter is comparable to the optimal fixed thresholds.

3.2 A sequential smoothing algorithm with linear computational cost

In this section we propose a new particle smoother for sequentially estimating $p(x_t|y_{1:T})$ that aims to improve upon those reviewed in Section 2.2.

3.2.1 Weaknesses of current particle smoothers

In Section 2.2 we reviewed current sequential algorithms that extend the particle filter to estimate the marginal smoothing densities $p(x_t|y_{1:T})$. The simplest of these is the Filter-Smoother which is a simple extension of the particle filter and therefore shares its $\mathcal{O}(N)$ complexity. However, as we show in Subsection 2.2.2, the smoother's estimates of $p(x_t|y_{1:T})$ are increasingly poor as t falls from T to 1.

Both the Forward-Backward and Two-Filter smoothers aim to improve on the simple Filter-Smoother by removing its dependence on the inheritance paths of the particle filter. Forward-Backward smoothing does this by re-weighting the filter particles while Two-Filter smoothing re-weights particles sampled from a backwards filter. However, both algorithms are $\mathcal{O}(N^2)$ as the calculation of each particle's weight is an $\mathcal{O}(N)$ operation. Thus, while variants of these particle smoothers produce better estimates for a fixed particle number N, far fewer particles can be used for these algorithms than can for the Filter-Smoother in a fixed amount of time.

Another advantage of the Filter-Smoother is that it gives draws of the joint smoothing distribution $p(x_{1:T}|y_{1:T})$ rather than only the marginal distributions. It is possible to adapt the Forward-Backward Smoother to also draw samples from the joint smoothing distribution as shown in Godsill et al. (2004). Their derivation is similar to that of the Forward-Backward Smoother above and as such share its complexity. They therefore achieve better samples of the joint distribution than

the Filter-Smoother for a fixed N but give a slightly worse representation of the marginal distributions than the Forward-Backward Smoother.

Since the Forward-Backward Smoother and the Filter-Smoother rely on the support of filter particles we may expect them to approximate $p(x_t|y_{1:T})$ best for t close to T where the target is most similar to $p(x_t|y_{1:t})$. Likewise the Two-Filter Smoother may do best for small t when the backwards filter distribution $\tilde{p}(x_t|y_{t:T})$ is likely to be closest to our target. However, when there is a large discrepancy between these distributions the particles will be weighted very unevenly as they will not be located in the right position to represent the smoothed distribution. Ideally we would like an algorithm which samples particles in the correct position for the smoothed distribution.

Degeneracy of the Forward-Backward and Two-Filter smoothers

As a final point we note that the Forward-Backward and Two-Filter smoothers' reliance on the form of the state density causes degeneracy problems with certain models and filters. Specifically, this happens whenever the state density $f(x_t|x_{t-1})$ is zero or approximately so for most combinations of possible x_t and x_{t-1}. As an example, consider the simple AR(2) process

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \epsilon_t$$

with $\epsilon_t \sim \mathcal{N}(0, \nu^2)$. The model can be written as a two-dimensional Markov process by defining the state as $x_t = (x_{t,1}, x_{t,2})'$ where $x_{t,1} = z_t$ and $x_{t,2} = z_{t-1}$.

This gives the state transition density

$$f(x_t|x_{t-1}) = \mathcal{N}(x_{t,1}|\phi_1 x_{t-1,1} + \phi_2 x_{t-1,2}, \nu^2)\, \delta(x_{t,2} - x_{t-1,1}),$$

where we write $\mathcal{N}(z|\mu, \nu^2)$ for the density of $\mathcal{N}(\mu, \nu^2)$ evaluated at z and $\delta(\cdot)$ for

the Dirac delta function. This density is zero whenever the second component of x_t does not equal the first component of x_{t-1}. This means that for two sets of particles $\{x_{t-1}^{(j)}\}$ and $\{x_t^{(i)}\}$, $f(x_t^{(i)}|x_{t-1}^{(j)})$ is likely to be zero unless $x_t^{(i)}$ was generated from $x_{t-1}^{(j)}$.

Since the Forward-Backward Smoother relies on comparing particles sampled from the filter at time t with those at time $t+1$, it can be shown that the weight (2.13) reduces to the effective weight given to each particle by the Filter-Smoother. However, the situation is worse for Two-Filter smoothing which fails completely as the forwards and backwards filter particles were sampled independently. With probability 1, no pairs of forwards and backwards filter particles match and so all the weights (2.16) will be zero.

The situation is similar whenever the state process is highly dependent as then there is a near-deterministic relationship between successive states; while the state transition density $f(x_t|x_{t-1})$ may be non-zero for every combination of x_{t-1} and x_t, it is likely that $f(x_t^{(i)}|x_{t-1}^{(j)})$ will be negligible unless $x_t^{(i)}$ was generated from $x_{t-1}^{(j)}$. Thus, while it is possible that the Forward-Backward and Two-Filter smoothers may be extended to avoid degeneracy with exact deterministic relationships, long range dependence in the state will often cause near-deterministic relationships that will hinder these algorithms.

3.2.2 New smoothing algorithm

We now describe our new smoothing algorithm which attempts to overcome the weaknesses of the current methods. Our primary aim is to draw new particles from the marginal smoothing densities directly rather than re-weight those drawn from another distribution. We describe the basic idea first, and then look at how the smoother can be implemented so that its computational cost is linear in the number of particles.

We start with a similar derivation to the Two-Filter Smoother given in Subsection 2.2.4 by writing the target density in terms of a forwards filter and a backwards information filter. Using the artificial priors $\gamma_t(x_t)$ and backwards filter densities $\tilde{p}(x_t|y_{t:T})$ of (2.15), we have

$$\begin{aligned} p(x_t|y_{1:T}) &\propto p(x_t|y_{1:t-1}) \cdot g(y_t|x_t) \cdot p(y_{t+1:T}|x_t) \\ &\propto \int f(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) \, \mathrm{d}x_{t-1} \cdot g(y_t|x_t) \cdot \\ &\qquad \int f(x_{t+1}|x_t) \frac{\tilde{p}(x_{t+1}|y_{t+1:T})}{\gamma_{t+1}(x_{t+1})} \, \mathrm{d}x_{t+1}. \end{aligned} \tag{3.12}$$

Thus the target smoothing density is the product of a filter term, a backwards filter term and the observation density $g(y_t|x_t)$.

These integrals can be approximated using weighted particles $\{(x_{t-1}^{(j)}, w_{t-1}^{(j)})\}$ from a particle filter at time $t-1$ and $\{(\tilde{x}_{t+1}^{(k)}, \tilde{w}_{t+1}^{(k)})\}$ from a backwards information filter at time $t+1$ to obtain

$$p(x_t|y_{1:T}) \simeq c \sum_{j=1}^{N} \sum_{k=1}^{N} f(x_t|x_{t-1}^{(j)}) w_{t-1}^{(j)} \cdot g(y_t|x_t) \cdot \frac{f(\tilde{x}_{t+1}^{(k)}|x_t)}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})} \tilde{w}_{t+1}^{(k)}, \tag{3.13}$$

where c is a normalising constant. Though this formula can be written as the product of two sums, we write it as a double sum to emphasise that there are N^2 (j,k) pairs. We also note that any filtering algorithm can be used to generate $\{x_{t-1}^{(j)}\}$ and $\{\tilde{x}_{t+1}^{(k)}\}$ as long as the artificial priors $\gamma_{t+1}(x_{t+1})$ here are the same ones used to sample $\{(\tilde{x}_{t+1}^{(k)}, \tilde{w}_{t+1}^{(k)})\}$ in the backwards information filter. As with the Two-Filter Smoother, we may assume $\gamma_{t+1}(x_{t+1}) \equiv 1$ throughout if the backwards filter is selected to approximate $p(y_{t+1:T}|x_{t+1})$ instead of $\tilde{p}(x_{t+1}|y_{t+1:T})$.

To sample from this approximation we start by mirroring the auxiliary particle filter of Pitt and Shephard (1999a) by finding a sampling distribution $\bar{q}$ and weights

$\bar{\beta}_t^{(j,k)}$ such that

$$\bar{q}(x_t|x_{t-1}^{(j)}, y_t, \tilde{x}_{t+1}^{(k)})\bar{\beta}_t^{(j,k)} \simeq f(x_t|x_{t-1}^{(j)})g(y_t|x_t)f(\tilde{x}_{t+1}^{(k)}|x_t)\frac{w_{t-1}^{(j)}\tilde{w}_{t+1}^{(k)}}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})}.$$

Algorithm 3.1 gives the algorithm that results from using the $\bar{\beta}_t^{(j,k)}$s to sample (j,k) pairs before using $\bar{q}$ to sample new particles $\bar{x}_t^{(i)}$.

Note that the output of Algorithm 3.1 is a set of triples, $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)}, \tilde{x}_{t+1}^{(k_i)})$, with associated weights, $\bar{w}_t^{(i)}$. These can be viewed as a particle approximation to $p(x_{t-1:t+1}|y_{1:T})$. If our interest solely lies in the marginal $p(x_t|y_{1:T})$ we just keep the particles, $\bar{x}_t^{(i)}$, and their associated weights, $\bar{w}_t^{(i)}$. Alternatively, we can use the weighted triple to provide the marginal for three time steps and iterate the smoothing stage of our algorithm at every third step only. We compare the efficiencies of both these methods with the stochastic volatility model in Subsection 3.2.3.

We note that the optimal choice of propagation density is $\bar{q}(x_t|x_{t-1}^{(j)}, y_t, \tilde{x}_{t+1}^{(k)}) =$

Algorithm 3.1: New $\mathcal{O}(N^2)$ smoothing algorithm.

---

1. **Filter forwards:** Run a particle filter to generate $\{(x_t^{(j)}, w_t^{(j)})\}$ approximating $p(x_t|y_{1:t})$ for $t = 0, \ldots, T$.
2. **Filter backwards:** Run a backwards information filter to generate $\{(\tilde{x}_t^{(k)}, \tilde{w}_t^{(k)})\}$ approximating $\tilde{p}(x_t|y_{t:T}) \propto \gamma_t(x_t)p(y_{t:T}|x_t)$ for $t = T, \ldots, 2$.
3. **Smooth:** For $t = 1, \ldots, T-1$
   (a) **Re-sample:** Calculate the $\bar{\beta}_t^{(j,k)}$s and use them as probabilities to sample N pairs $\{(j_i, k_i)\}_{i=1}^N$.
   (b) **Propagate:** Sample the new particles $\bar{x}_t^{(i)}$ independently from $\bar{q}(\cdot|x_{t-1}^{(j_i)}, y_t, \tilde{x}_{t+1}^{(k_i)})$.
   (c) **Re-weight:** Assign each particle $\bar{x}_t^{(i)}$ the weight

$$\bar{w}_t^{(i)} \propto \frac{f(\bar{x}_t^{(i)}|x_{t-1}^{(j_i)})\, g(y_t|\bar{x}_t^{(i)})\, f(\tilde{x}_{t+1}^{(k_i)}|\bar{x}_t^{(i)})\, w_{t-1}^{(j_i)}\, \tilde{w}_{t+1}^{(k_i)}}{\bar{q}(\bar{x}_t^{(i)}|x_{t-1}^{(j_i)}, y_t, \tilde{x}_{t+1}^{(k_i)})\, \bar{\beta}_t^{(j_i,k_i)}\, \gamma_{t+1}(\tilde{x}_{t+1}^{(k_i)})}$$

   and normalise them to sum to 1.

$p(x_t|x_{t-1}^{(j)}, y_t, \tilde{x}_{t+1}^{(k)})$ while the optimal re-sampling probabilities are given by

$$\bar{\beta}_t^{(j,k)} \propto \int f(x_t|x_{t-1}^{(j)})\, g(y_t|x_t)\, f(\tilde{x}_{t+1}^{(k)}|x_t)\, \mathrm{d}x_t \frac{w_{t-1}^{(j)}\, \tilde{w}_{t+1}^{(k)}}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})}. \tag{3.14}$$

We do not require our algorithm to generate samples for time T since these are available from the filter. Similarly, particles for time 1 are available from the backwards filter if we use $\gamma_t(x_t) = \pi(x_t)$ for the artificial priors.

Like the Two-Filter Smoother in Subsection 2.2.4, our smoothing step is not sequential and can be performed independently for each time t. Also, the computational complexity of each step is $\mathcal{O}(N^2)$ which is comparable with all but the simplest Filter-Smoother. However, as it stands we have N^2 $\bar{\beta}_t^{(j,k)}$s to calculate making it $\mathcal{O}(N^2)$ in memory also which could mean that it is impractical for even modest sample sizes.

Making Algorithm 3.1 $\mathcal{O}(N)$

The above smoothing algorithm has a computational cost that is $\mathcal{O}(N^2)$, that is quadratic in the number of particles, due to the need to calculate N^2 probabilities, $\bar{\beta}_t^{(j,k)}$. A simple approach to reduce the computational cost of the smoothing algorithm is to choose these probabilities so that they correspond to choosing particles at time $t-1$ and backward-filter particles at time $t+1$ independently of each other. Our algorithm will then be $\mathcal{O}(N)$ in computational complexity as well as memory and as such will be much faster for large N. A similar idea is used in Briers et al. (2005) for inference for graphical models, however we are unaware of any previous use of this approach within particle smoothing algorithms.

Now the optimal distribution from which to choose the particles at time $t-1$ will be the corresponding marginal distribution of the optimal probabilities for $\bar{\beta}_t^{(j,k)}$,

given in (3.14). Marginalising we get:

$$\begin{aligned}\sum_{k=1}^{N}\bar{\beta}_t^{(j,k)} &\propto \sum_{k=1}^{N}\int f(x_t|x_{t-1}^{(j)})\,g(y_t|x_t)\,f(\tilde{x}_{t+1}^{(k)}|x_t)\,\mathrm{d}x_t\frac{w_{t-1}^{(j)}\,\tilde{w}_{t+1}^{(k)}}{\gamma_{t+1}(\tilde{x}_{t+1}^{(k)})}\\ &\xrightarrow{N\to\infty}\iint f(x_t|x_{t-1}^{(j)})\,g(y_t|x_t)\,f(x_{t+1}|x_t)\,\mathrm{d}x_t\frac{w_{t-1}^{(j)}\,\tilde{p}(x_{t+1}|y_{t+1:T})}{\gamma_{t+1}(x_{t+1})}\,\mathrm{d}x_{t+1}\\ &\propto p(y_{t:T}|x_{t-1}^{(j)})\,w_{t-1}^{(j)}.\end{aligned}$$

Calculating this analytically will be impossible, but it suggests two simple approximations. The first is to sample particles at time $t-1$ according to their filtering weights $w_{t-1}^{(j)}$. However, a better approach will be to sample according to an approximation of $p(y_t|x_{t-1}^{(j)})w_{t-1}^{(j)}$, as it includes the information in the observation at time t. Now, in performing the particle filter we used the auxiliary filter which sampled particle $x_{t-1}^{(j)}$ with a probability $\beta_t^{(j)}$ which is chosen to be an approximation to $p(y_t|x_{t-1}^{(j)})w_{t-1}^{(j)}$. Thus we suggest using exactly the same probabilities to sample the particles within one iteration of our sampling algorithm.

By similar calculations, it can be shown that we should optimally choose the backward-filter particles at time $t+1$ with probability proportional to $\tilde{p}(y_{1:t}|\tilde{x}_{t+1}^{(k)})\tilde{w}_{t+1}^{(k)}$. Again, we cannot calculate these exactly, but a simple idea is to use probabilities that approximate $\tilde{p}(y_t|\tilde{x}_{t+1}^{(k)})\tilde{w}_{t+1}^{(k)}$. Thus we can simply use the probabilities $\tilde{\beta}_t^{(k)}$ that were used in the backward filter, as these were chosen as to be an approximation to $\tilde{p}(y_t|\tilde{x}_{t+1}^{(k)})\tilde{w}_{t+1}^{(k)}$.

Algorithm 3.2: New $\mathcal{O}(N)$ smoothing algorithm.

Proceed as Algorithm 3.1 but substitute steps 3(a) and 3(c) with

3. (a) **Re-sample:** Use $\{\beta_t^{(j)}\}$ from the filter to sample $j_1,...,j_N$ and $\{\tilde{\beta}_t^{(k)}\}$ from the backwards filter to sample $k_1,...,k_N$ from $\{1,...,N\}$

   (c) **Re-weight:** Assign each particle $\bar{x}_t^{(i)}$ the weight

$$\bar{w}_t^{(i)} \propto \frac{f(\bar{x}_t^{(i)}|x_{t-1}^{(j_i)})\,g(y_t|\bar{x}_t^{(i)})\,f(\tilde{x}_{t+1}^{(k_i)}|\bar{x}_t^{(i)})\,w_{t-1}^{(j_i)}\,\tilde{w}_{t+1}^{(k_i)}}{\bar{q}(\bar{x}_t^{(i)}|x_{t-1}^{(j_i)},y_t,\tilde{x}_{t+1}^{(k_i)})\,\beta_t^{(j_i)}\,\tilde{\beta}_t^{(k_i)}\,\gamma_{t+1}(\tilde{x}_{t+1}^{(k_i)})}$$

   and normalise them to sum to 1.

We thus obtain a similar algorithm to before, but with particles at time $t-1$ and $t+1$ sampled independently, and with $\bar{\beta}_t^{(j,k)}$ replaced by $\beta_t^{(j)}\tilde{\beta}_t^{(k)}$ in the calculation of the weight. Thus we have an $\mathcal{O}(N)$ version of our smoothing algorithm shown in Algorithm 3.2. We note that we can speed up the algorithm further as the probabilities $\beta_t^{(j)}$ and $\tilde{\beta}_t^{(k)}$ (or even the auxiliary variables $\{j_i\}$ and $\{k_i\}$) can be saved from the filters to reduce the number of calculations in the smoothing step.

Degeneracy and block smoothing

Algorithms 3.1 and 3.2 overcome the degeneracy problem of the Forward-Backward and Two-Filter smoothers when there is a deterministic or near-deterministic relationship between the states at successive time-points, as demonstrated in Subsection 3.2.1 with the AR(2) model. They will still have degeneracy problems where there is a deterministic relationship between components of states separated by two or more time-points. However, it is simple to extend our method so that we jointly sample a block $(x_t, \ldots, x_{t+n-1})$ so that this restriction may be removed (see Doucet et al. (2006) for an example of block sampling in particle filters).

The block version of our smoother may be derived by following similar arguments to the original algorithm starting from

$$p(x_{t:t+n-1}|y_{1:T}) \propto p(x_t|y_{1:t-1}) \left(\prod_{s=t+1}^{t+n-1} f(x_s|x_{s-1}) \right) \cdot \left(\prod_{s=t}^{t+n-1} g(y_s|x_s) \right) \cdot p(y_{t+n:T}|x_{t+n-1}).$$

in place of (3.12). This leads to an algorithm where we sample a block $(x_t, \ldots, x_{t+n-1})$ given filter particles $\{x_{t-1}^{(j)}\}$ and backwards filter particles $\{\tilde{x}_{t+n}^{(k)}\}$. The same marginalising argument may be used to give an $\mathcal{O}(N)$ version which suggests re-sampling $\{x_{t-1}^{(j)}\}$ with $\beta_t^{(j)}$ and $\{\tilde{x}_{t+n}^{(k)}\}$ with $\tilde{\beta}_{t+n-1}^{(k)}$. The resulting algorithm is given in Algorithm 3.3.

Algorithm 3.3: New $\mathcal{O}(N)$ block smoothing algorithm.

Run the filter and backwards filter as before in Algorithm 3.1.

3. **Smooth:** For $t = 1, n+1, 2n+1, \ldots, \lfloor \frac{T-2}{n} \rfloor n + 1$,

   (a) **Re-sample:** Use $\{\beta_t^{(j)}\}$ from the filter to sample $j_1, ..., j_N$ and $\{\tilde{\beta}_{t+n-1}^{(k)}\}$ from the backwards filter to sample $k_1, ..., k_N$ from $\{1, ..., N\}$.

   (b) **Propagate:** Sample new particle blocks $\bar{x}_{t:t+n-1}^{(i)}$ independently from $\bar{q}(\cdot|x_{t-1}^{(j_i)}, y_{t:t+n-1}, \tilde{x}_{t+n}^{(k_i)})$.

   (c) **Re-weight:** Assign each particle block $\bar{x}_{t:t+n-1}^{(i)}$ the weight

$$\bar{w}_t^{(i)} \propto \frac{\prod_{s=t}^{t+n} f(\bar{x}_s^{(i)}|\bar{x}_{s-1}^{(i)}) \cdot \prod_{s=t}^{t+n-1} g(y_s|\bar{x}_s^{(i)}) \cdot w_{t-1}^{(j_i)} \, \tilde{w}_{t+n}^{(k_i)}}{\bar{q}(\bar{x}_{t:t+n-1}^{(i)}|x_{t-1}^{(j_i)}, y_{t:t+n-1}, \tilde{x}_{t+n}^{(k_i)}) \, \beta_t^{(j_i)} \, \tilde{\beta}_{t+n-1}^{(k_i)} \, \gamma_{t+n}(\tilde{x}_{t+n}^{(k_i)})},$$

   (where for brevity we define $\bar{x}_{t-1}^{(i)} := x_{t-1}^{(j_i)}$ and $\bar{x}_{t+n}^{(i)} := \tilde{x}_{t+n}^{(k_i)}$) and normalise the weights to sum to 1.

By choosing n sufficiently large such that there is not a deterministic relationship between components of $x_{t-1}^{(j)}$ and $x_{t+n}^{(k)}$, our approach to smoothing can then be applied successfully. Sampling particles in a block may also be beneficial when there is no issue with degeneracy, as by using a larger block size we can reduce the dependence between the filter and the backwards filter particles which may improve the efficiency of the algorithm. We demonstrate this in Subsection 3.2.3 by applying the block version of our algorithm to the stochastic volatility model.

3.2.3 Simulation studies

We now compare the efficiency of our new algorithm against the currently available methods for the linear-Gaussian and the stochastic volatility models.

Linear-Gaussian model

Our first simulation study is based on a model with linear-Gaussian state and observation models. The specific state model we used is chosen to be the same as for

our athletics application in Section 4.1. We have a chosen a linear-Gaussian observation model so that we can compare results of different particle smoothers with the true smoothing distributions obtained from the Kalman filter and smoother (see Kalman (1960) and Anderson and Moore (1979)).

Specifically, we consider the model (2.4) on page 7 in two dimensions with

$$F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad Q = \nu^2 \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}, \tag{3.15}$$
$$G = (1, 0), \qquad R = \tau^2.$$

The state transition distribution, defined by F and Q, is derived in Appendix B.1 from the pair of stochastic differential equations (SDEs):

$$\begin{aligned} \mathrm{d}X_{t,1} &= X_{t,2}\,\mathrm{d}t, \\ \mathrm{d}X_{t,2} &= \nu\,\mathrm{d}B_t, \end{aligned} \tag{3.16}$$

and so the first component $X_{t,1}$ is the integrated path of the random walk $X_{t,2}$. A noisy observation of the first component is made at each time step. The parameter ν^2 determines the smoothness of the state over time. With a large value of ν^2 the state can move freely and thus follows the observations. When ν^2 is small, however, the model makes a linear fit to the observations.

We compare the two versions of our new algorithm with the simple Filter-Smoother of Subsection 2.2.2, the Forward-Backward Smoother of Subsection 2.2.3 and the Two-Filter Smoother of Subsection 2.2.4. We also look at how the relative performance of the algorithms is affected by the ratio of the state noise ν^2 to observation noise τ^2. The details of our particle filter, backwards filter and smoothing algorithms for this model are given in Appendix A.1.

To compare the accuracy of our smoothing algorithms' estimates of $X_{t,d}$ we esti-

mate the *effective sample size* $N_{\text{eff}}(X_{t,d})$. Motivated by the fact that

$$\mathbf{E}\left(\frac{(\bar{X}-\mu)^2}{\sigma^2}\right) = \frac{1}{N},$$

when $X^{(1)}, \ldots, X^{(N)}$ IID $\mathcal{N}(\mu, \sigma^2)$ and $\bar{X}$ is their sample mean, we take

$$N_{\text{eff}}(X_{t,d}) = \mathbf{E}\left(\frac{(\hat{x}_{t,d}-\mu_{t,d})^2}{\sigma^2_{t,d}}\right)^{-1}, \tag{3.17}$$

where $\mu_{t,d}$ and $\sigma^2_{t,d}$ are the true mean and variance of $X_{t,d}|y_{1:T}$ obtained from the Kalman smoother and $\hat{x}_{t,d}$ is the random estimate from a particle smoother. We can therefore crudely say that the weighted sample produced by our smoother is as accurate at estimating $X_{t,d}$ as an independent sample of size $N_{\text{eff}}(X_{t,d})$. To estimate the expectation in (3.17) we use the mean value from 100 repetitions of each algorithm.

We first compare the smoothing algorithms using model parameters of $\nu^2 = \tau^2 = 1$ with $\mu_0 = (0,0)'$ and $\Sigma_0 = I_2$ for the prior. We generated 20 datasets, each of length 200, and averaged the effective sample sizes to remove effects caused by a single dataset.

We chose different numbers of particles for each algorithm to try to reflect the varying complexities of each method. We started by choosing $10,000$ particles for the Filter-Smoother and $3,000$ for the $\mathcal{O}(N)$ version of our new algorithm since they then took approximately the same amount of time to run. We would have liked to scale the $\mathcal{O}(N^2)$ algorithms to take the same time to run but their speeds varied greatly. Part of this may be due to how the algorithms are implemented in R. We therefore fixed the number of particles for these three algorithms at 300.

Algorithm	Filter	Forward-Backward	Two-Filter	New $\mathcal{O}(N^2)$	New $\mathcal{O}(N)$
N	10,000	300	300	300	3,000
Run time (s)	224	688	358	40	255

Table 3.5: Number of particles used and average run time of each algorithm.

This made the $\mathcal{O}(N^2)$ version of our new algorithm a lot faster but the other two methods slower than the Filter-Smoother. The average time taken by each algorithm per run is shown in Table 3.5.

Figure 3.6 shows how the average effective number of particles for estimating $X_{t,1}$ varies through time for the five algorithms considered. The results for $X_{t,2}$ (not shown) are very similar.

We can see that the Filter-Smoother does very well for times close to $T = 200$ as this filter has by far the most particles and the filter and smoothing distributions are similar at this stage of the process. As predicted, however, this algorithm gets progressively worse as it goes backwards through time. This is not necessarily the case with the other algorithms whose efficiencies remain roughly constant over time when averaged over the 20 datasets. Of the two $\mathcal{O}(N)$ algorithms we see that our new method vastly outperforms the Filter-Smoother for all but the final few time steps, despite taking a similar amount of time to run.

From Figure 3.6 we can also see that the three $\mathcal{O}(N^2)$ algorithms have near identical efficiencies for this particular model. This could be because they all have similar forms in that filter particles are combined, via an $\mathcal{O}(N^2)$ approximation,

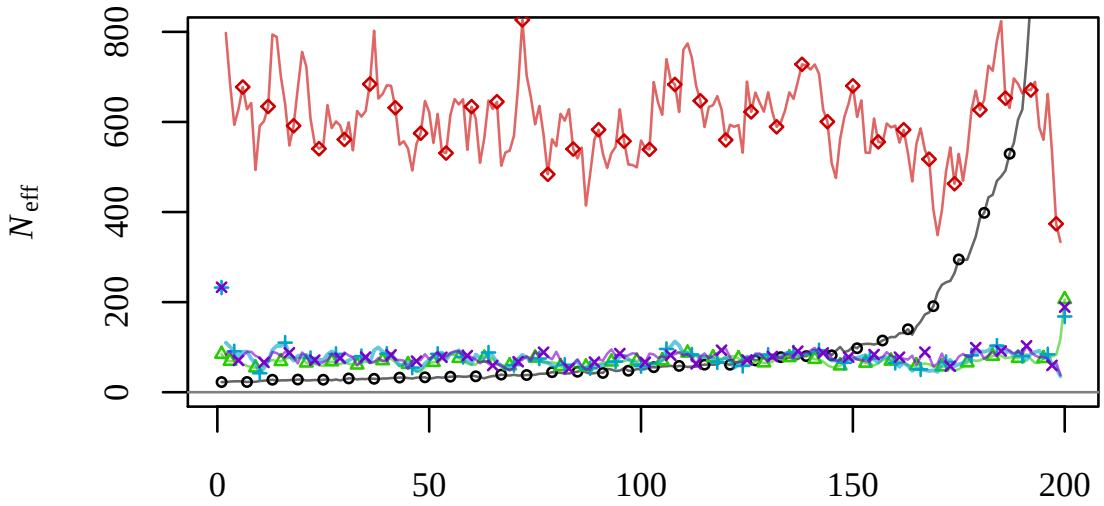
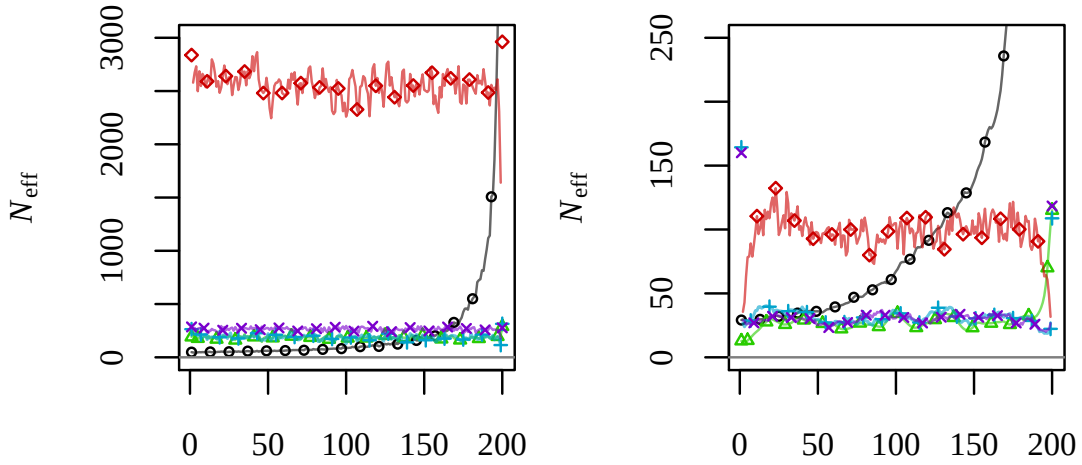


Figure 3.6: Average effective sample size for each of the 200 time steps using the filter ($\circ$), Forward-Backward ($\triangle$) and Two-Filter smoothers ($+$) as well as the $\mathcal{O}(N^2)$ ($\times$) and $\mathcal{O}(N)$ versions ($\Diamond$) of our new algorithm.

with particles from future time steps sampled backwards in time. We recall that these were run with the same number of particles N though in our implementation our new algorithm was faster than the other two here. However, even with this taken into account, the $\mathcal{O}(N)$ version is many times more efficient for even these modest sample sizes N.

To see how these results are affected by the ratio of the state noise ν^2 to the observation noise τ^2, we repeat the experiment first with $\nu^2 = 100$ while keeping $\tau^2 = 1$. This gives the state freedom to follow the observations which helps the algorithms to perform well. The results are shown in Figure 3.7a below. Those for $\nu^2 = 1$ and $\tau^2 = 1/100$ gave very similar results.

We see that the accuracy of the Filter-Smoother still diminishes as it progresses backwards through time but all the other methods are close to their optimal efficiency of an effective sample size equal to N. This is particularly the case with our new $\mathcal{O}(N^2)$ algorithm which outperforms the other $\mathcal{O}(N^2)$ methods at every time step. Our new $\mathcal{O}(N)$ algorithm, however, is by far the fastest allowing it to have 10 times as many particles as the slower methods. Its efficiency also suggests that our choice of re-sampling weights is reasonable.



(a) $\nu^2/\tau^2 = 100$ (b) $\nu^2/\tau^2 = 1/100$

Figure 3.7: Average effective sample sizes as in Figure 3.6 with different ratios of the state noise ν^2 to the observation noise τ^2.

We finally repeat the experiment with $\nu^2/\tau^2 = 1/100$ which makes the state highly dependent through time and causes all the particle methods to struggle. This can be seen from the low effective sample sizes in Figure 3.7b. Even though the Filter-Smoother diminishes at a slightly faster rate than before, it does better than the other algorithms for a large number of time steps. This is possibly due to the total accumulation of error in the filter, backwards filter and smoother, each of which performs badly in this case, which hinder the other methods. Also, because the state is so highly correlated, the filter weights change little going back in time so that the degradation of the Filter-Smoother is less, countering the poor performance of the filter. However, the Filter-Smoother eventually drops below the accuracy of our $\mathcal{O}(N)$ method showing that our $\mathcal{O}(N)$ algorithm can give stronger estimates of the earliest smoothing densities in even the toughest situations.

As a final point, we consider the surprisingly poor performance of our $\mathcal{O}(N)$ algorithm at the very earliest and latest time steps, as shown in Figure 3.7b. While, unfortunately, we can offer no explanation of this, we have verified that it is a genuine feature of our algorithm under extremely high state correlation, at least using this linear-Gaussian model. Given the good performance of the Filter-Smoother at the final time steps, it is likely that our $\mathcal{O}(N)$ algorithm may be improved by instead using the Filter-Smoother estimates for the last few steps. Also, if the artificial priors are chosen as $\gamma_t(x_t) = p(x_t)$, the backwards filter can replace the filter to produce a backwards Filter-Smoother algorithm that would be expected to provide better estimates for the earliest time steps. It is likely that this combination of the Filter-Smoother, backwards Filter-Smoother and our new $\mathcal{O}(N)$ algorithm may produce optimal estimates whatever the relative dependencies in the model.

Stochastic volatility

For our second simulation study we use the stochastic volatility model given by (3.7). We consider two issues: Firstly, we investigate whether our algorithm performs better when we use the triples $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)}, \tilde{x}_{t+1}^{(k_i)})$ as samples from $p(x_{t-1:t+1}|y_{1:T})$ iterating the smoother every three time steps instead of keeping only the sampled particle $\bar{x}_t^{(i)}$ and iterating at every step. By using the triples the smoothing stage takes a third of the time so that more particles can be used for higher accuracy. However, by using re-weighted filter and backwards filter particles we have fewer distinct samples than we have if we sample all the smoother particles afresh. We may therefore expect the efficiency to drop, especially when the filter's densities differ considerably from their smoother counterparts.

We also look at how sampling particles in blocks as described by Algorithm 3.3 affects the efficiency of our algorithm (see Shephard and Pitt (1997) for an example of block updating in MCMC with the stochastic volatility model). By choosing larger block sizes the dependence between the re-sampled filter and backwards filter particles may be reduced. This is beneficial for our $\mathcal{O}(N)$ algorithm since here the filter particles are matched up independently and so when there is high dependence most of the pairings will be unlikely and so given a negligible weight. However, increasing the block size increases the dimension over which the importance weight applies which could also lead to uneven weights.

For the model parameters we take $\phi = 0.9720$, $\nu = 0.178$ and $\beta = 0.5992$ from Pitt and Shephard (1999b) and use the stationary distribution $\mathcal{N}(0, \nu^2/(1-\phi^2))$ for the prior. This causes high dependence between the states which larger block sizes may help to overcome. We generate a dataset of length 300 which remains constant over the simulation. Since in practice this model would be used to estimate the volatility $\beta e^{x_t/2}$, we compare the performance of the smoothers by using the variance of the volatility estimates and the effective sample size (3.17) for the state over 300 repetitions. However, since the effective sample size requires the true smoother

mean and variance, we first estimate these using the Filter-Smoother with two million particles.

For the simulation itself, we compare the Filter-Smoother with multiple versions of our $\mathcal{O}(N)$ algorithm. We first run our algorithm as described by Algorithm 3.2 sampling $\bar{x}_t^{(i)}$ for $t = 2, \ldots, T-1$ while discarding the re-sampled filter and backwards filter particles. We compare this with a second version that keeps the triples $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)}, \tilde{x}_{t+1}^{(k_i)})$, uses these to estimate the smoothing distributions at times $t-1$, t and $t+1$, and then iterates the smoother every three time steps. We chose 10,000 particles for the Filter-Smoother and scaled N for both versions of our $\mathcal{O}(N)$ algorithm so that they all took the same time to run. Details of the forward and backwards filters as well as our smoother for this model are given in Appendix A.2.

The results are summarised in Table 3.6. We see that by iterating the smoothing step less often we are able to use nearly twice as many particles overall in the same amount of time. It is interesting to see that this gain extends to the effective sample size suggesting that little is lost using re-weighted filter particles. By using the full output of our algorithm we reduce the variance of the volatility estimates by almost a half.

To see whether, by using triples, we could have shown even greater results in the previous linear-Gaussian simulation study, we run the triples version of our $\mathcal{O}(N)$ algorithm on this model (with $\nu^2 = 1$). Using $N = 4,324$ particles, this ran at the

Algorithm	N	N_{eff}	$\widehat{\text{Var}}$(volatility) ($\times 10^{-5}$)
Filter-Smoother	10,000	786	5.240
New $\mathcal{O}(N)$ as Algorithm 3.2	3,780	710	2.790
New $\mathcal{O}(N)$ using triples	7,005	1,343	1.550

Table 3.6: Comparison of the Filter-Smoother with two variations of our new $\mathcal{O}(N)$ algorithm: one that samples $\bar{x}_t^{(i)}$ at every time step and another which uses the triples $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)}, \tilde{x}_{t+1}^{(k_i)})$ iterating every three steps. The number of particles is varied so that each algorithm took the same time to run. The final columns give the average effective sample size and the average variance of the volatility estimate over the 300 time steps.

same speed as the original $\mathcal{O}(N)$ algorithm with $3,000$ particles. While the original method had an overall average effective sample size of 621 (see Figure 3.6), the triples version obtained 930. This again shows that any loss in efficiency due to re-sampling filter particles for two of every three time steps is more than compensated for by the increase in particle numbers. It is likely that this is the best strategy for any model, at least whenever the marginal smoothing distributions are not too dissimilar to the forwards and backwards filters.

We now return to the stochastic volatility model and focus on the block sampling extension of our algorithm given in Algorithm 3.3. Adding to the results of Table 3.6 we run our smoother with a selection of larger block sizes, again scaling N so that they take the same time to run. Since we have just demonstrated the gains made by using the re-weighted filter and backwards filter particles produced by our algorithm, we now do the same for larger block sizes. This is done by sampling a block of size n and extending this with the re-weighted particles to give an overall block $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)}, \ldots, \bar{x}_{t+n-1}^{(i)}, \tilde{x}_{t+n}^{(k_i)})$ of size $n+2$. The smoothing step then iterates every $n+2$ steps rather than every n as reported in Algorithm 3.3.

The results are given in Table 3.7. We see that the number of particles varies with block size since larger blocks lead to fewer smoothing steps but each step takes

Algorithm	Block size	N	N_{eff}	$\widehat{\text{Var}}$(volatility) $(\times 10^{-5})$
Filter-Smoother	-	10,000	786	5.240
New $\mathcal{O}(N)$	3	7,005	1,343	1.550
	5	8,055	2,129	0.807
	10	9,181	3,488	0.458
	20	9,714	4,054	0.367
	30	9,547	4,140	0.349
	50	8,624	2,515	2.200
	100	6,755	334	9.110

Table 3.7: Comparison of the Filter-Smoother with our new $\mathcal{O}(N)$ algorithm when different block sizes are used. The block size reported includes the two re-sampled particles at either end of the block. The number of particles is varied so that each algorithm took the same time to run. The final columns give the average effective sample size and the average variance of the volatility estimate over the 300 time steps.

longer to run. Despite running two filters as well as a smoothing stage, some runs of our algorithm have almost as many particles as the Filter-Smoother which is possible since the Filter-Smoother has the extra overhead of keeping track of each particle's history.

Table 3.7 shows that the accuracy of our smoother increases with block sizes greater than 3, peaking here around 30. We see that this is a vast improvement over the Filter-Smoother. The accuracy of our smoother diminishes, however, with very large block sizes as the dimension over which the importance weight applies becomes too great.

3.3 EM algorithm for static parameter estimates

In this final section we present an *Expectation-Maximisation* (EM) algorithm that can be used for estimating static parameters in the model.

3.3.1 EM algorithm

We often require estimates of parameters θ which, to contrast with the state X_t, remain constant over time. We review current methods for this in Subsection 2.1.6. While augmenting the state vector and the method of Storvik (2002) produce sequential parameter estimates that are updated with the filter, they often require sufficient statistics of θ to exist for them to work efficiently. Alternatively, direct maximum likelihood methods require the filter to be run for each θ value in a grid which can be infeasible when there are many parameters to estimate.

As an alternative strategy, we intend to obtain an estimate of θ from the Expectation-Maximisation (EM) algorithm of Dempster et al. (1977). A similar method is proposed by Briers et al. (2004) and Wills et al. (2008). To do this we aim to maximise the likelihood $p(y_{1:T}|\theta)$ by iteratively maximising

$$Q(\theta|\theta^{(n-1)}) := \mathbf{E}\Big(\log(p(X_{0:T}, y_{1:T}|\theta)) \;\Big|\; y_{1:T}, \theta^{(n-1)}\Big)$$

to give $\theta^{(n)}$.

Estimating observation parameters

If we initially assume the parameters θ only appear in the observation density and not the prior or state densities, the joint log likelihood can be written as

$$\log(p(x_{0:T}, y_{1:T}|\theta)) = \log(\pi(x_0)) + \sum_{t=1}^{T} \log(f(x_t|x_{t-1})) + \sum_{t=1}^{T} \log(g(y_t|x_t, \theta)).$$

We therefore have

$$Q(\theta|\theta^{(n-1)}) = \text{const} + \sum_{t=1}^{T} \mathbf{E}\Big(\log(g(y_t|X_t,\theta)) \;\Big|\; y_{1:T}, \theta^{(n-1)}\Big)$$
$$\simeq \text{const} + \sum_{t=1}^{T}\sum_{i=1}^{N} \log(g(y_t|x_t^{(i)},\theta))\, w_t^{(i)},$$

where $(x_t^{(i)}, w_t^{(i)})$ are weighted particles approximating $p(x_t|y_{1:T}, \theta^{(n-1)})$. Thus we only require particles from the marginal smoothing densities to estimate the expectation so any smoothing algorithm can be used.

The EM algorithm therefore proceeds as follows. We start with an initial estimate of our parameters, $\theta^{(0)}$. Then, given our current estimate $\theta^{(n-1)}$, we use a smoothing algorithm such as Algorithm 3.2 to generate particles from each marginal smoothing density $p(x_t|y_{1:T}, \theta^{(n-1)})$. Then we use numerical optimisation (such as the `optim` function in R) to maximise $Q(\theta|\theta^{(n-1)})$ to give us a new estimate $\theta^{(n)}$. This is summarised in Algorithm 3.4.

The estimates should be monitored so that the algorithm can stop when $\theta^{(n)}$ differs very little from $\theta^{(n-1)}$. The model likelihood may also be estimated at every iteration so that the algorithm can stop when the likelihood reaches a maximum. This requires little additional work if the likelihood formula of Kitagawa (1996) is

Algorithm 3.4: EM algorithm for fixed parameters in the observation density.

1. **Initialisation:** Begin with an initial estimate $\theta^{(0)}$.
2. For $n = 1, 2, \ldots$
   (a) **Smooth:** Using the current parameter estimates $\theta^{(n-1)}$, run a particle smoother to generate smoothed particles $\{(x_t^{(i)}, w_t^{(i)})\}$ for $t = 1, \ldots, N$.
   (b) **Maximise:** Maximise
   $$Q(\theta|\theta^{(n-1)}) := \sum_{t=1}^{T}\sum_{i=1}^{N} \log(g(y_t|x_t^{(i)},\theta))\, w_t^{(i)}$$
   with a numerical optimiser to give new estimate $\theta^{(n)}$.

used since this uses filter particles which will most likely be calculated within the smooth step of the EM algorithm.

The EM algorithm in general is guaranteed to converge to a local maximum but this convergence can be slow. Since each iteration requires a complete run of a particle smoother algorithm and speed is likely to be more of an issue, it is therefore beneficial to use our $\mathcal{O}(N)$ algorithm of Section 3.2. In particular, the similar methods of Briers et al. (2004) and Wills et al. (2008) use the Two-Filter and Forwards-Backwards smoothers respectively which both have a complexity of $\mathcal{O}(N^2)$ making their methods slower than ours.

Since the EM algorithm converges only to a local maximum, it is advisable to start the algorithm at multiple starting values. If different final estimates are found, the model likelihood can be used to choose between them. It may also help to choose initial parameter estimates that are close to the maximum as unreasonable values may cause the particle smoother to struggle. It is often possible to use simpler tractable analyses of the data to get good estimates of the parameter values.

Estimating state parameters

We now relax the restriction placed upon θ and allow unknown parameters to appear in the prior and particularly the state density. This gives

$$\begin{aligned} Q(\theta|\theta^{(n-1)}) = &\mathbf{E}\Big(\log(\pi(X_0|\theta)) \;\Big|\; y_{1:T}, \theta^{(n-1)}\Big) + \\ &\sum_{t=1}^{T} \mathbf{E}\Big(\log(f(X_t|X_{t-1},\theta)) \;\Big|\; y_{1:T}, \theta^{(n-1)}\Big) + \\ &\sum_{t=1}^{T} \mathbf{E}\Big(\log(g(y_t|X_t,\theta)) \;\Big|\; y_{1:T}, \theta^{(n-1)}\Big) \\ \simeq &\sum_{i=1}^{N} \Bigg(\log(\pi(x_0^{(i)}|\theta)) + \sum_{t=1}^{T} \log(f(x_t^{(i)}|x_{t-1}^{(i)},\theta)) + \\ &\qquad \sum_{t=1}^{T} \log(g(y_t|x_t^{(i)},\theta)) \Bigg) w_t^{(i)}, \end{aligned} \tag{3.18}$$

where $(x_{t-1}^{(i)}, x_t^{(i)}; w_t^{(i)})$ are weighted pairs of particles approximating $p(x_{t-1}, x_t|y_{1:T}, \theta^{(n-1)})$. We note that these are available from our algorithm as either $(\bar{x}_{t-1}^{(i)}, \tilde{x}_t^{(k_i)})$ at time $t-1$ or as $(x_{t-1}^{(j_i)}, \bar{x}_t^{(i)})$ at time t but they can also be sampled from the Filter-Smoother or the joint smoothers of Hürzeler and Künsch (1998) or Godsill et al. (2004).

We can therefore proceed much as before, using a smoother which produces pairs of particles and maximising the Q given by (3.18). However, in most cases there will be different parameters in the prior, state and observation densities so the Q given by (3.18) may be split into two or three separate components. Each component can then be maximised separately at each iteration n which reduces the dimension of the space to maximise over, improving the efficiency of the numerical routines applied. This is illustrated in Algorithm 3.5 for the case where the prior has no parameters, and the state and observation densities' parameter sets are disjoint.

Algorithm 3.5: EM algorithm for fixed parameters in the state and observation densities.

---

1. **Initialisation:** Begin with initial estimates $\theta_f^{(0)}$ for the state and $\theta_g^{(0)}$ for the observation.
2. For $n = 1, 2, \ldots$
   (a) **Smooth:** Using the current parameter estimates $\theta_f^{(n-1)}$ and $\theta_g^{(n-1)}$, run a particle smoother to generate pairs of smoothed particles $\{(x_{t-1,t}^{(i)}, w_t^{(i)})\}$ for $t = 1, \ldots, N$.
   (b) **Maximise for state:** Maximise

$$Q_f(\theta_f|\theta_f^{(n-1)}, \theta_g^{(n-1)}) := \sum_{t=1}^{T}\sum_{i=1}^{N} \log(f(x_t^{(i)}|x_{t-1}^{(i)}, \theta_f))\, w_t^{(i)}$$

   to give new state parameter estimate $\theta_f^{(n)}$.
   (c) **Maximise for observation:** Maximise

$$Q_g(\theta_g|\theta_f^{(n-1)}, \theta_g^{(n-1)}) := \sum_{t=1}^{T}\sum_{i=1}^{N} \log(g(y_t|x_t^{(i)}, \theta_g))\, w_t^{(i)}$$

   to give new observation parameter estimate $\theta_g^{(n)}$.

3.3.2 Estimating observed information

One disadvantage of the EM algorithm is that it only gives a point estimate of θ with no automatic measure of uncertainty. The standard approach is to estimate the variance of our parameter estimates by inverting the observed information matrix $\mathcal{I}(\theta|y_{1:T})$. This can be estimated from the method of Louis (1982) which has

$$\mathcal{I}(\theta|y_{1:T}) = \mathrm{E}\big(\nabla^2 \log(p(X_{0:T}, y_{1:T}|\theta)) \mid y_{1:T}, \theta\big) - \mathrm{E}\big(\nabla \log(p(X_{0:T}, y_{1:T}|\theta)) \nabla \log(p(X_{0:T}, y_{1:T}|\theta))' \mid y_{1:T}, \theta\big), \quad (3.19)$$

where we use the θ estimate given by the EM algorithm.

Since $\log(p(x_{0:T}, y_{1:T}|\theta))$ is available in closed form as

$$\log(p(x_{0:T}, y_{1:T}|\theta)) = \log(\pi(x_0|\theta)) + \sum_{t=1}^{T} \log(f(x_t|x_{t-1}, \theta)) + \sum_{t=1}^{T} \log(g(y_t|x_t, \theta)),$$

the expectations in (3.19) can be estimated as before using particle pairs from our smoothing algorithm.

Chapter 4

State space modelling of univariate extremes

In this chapter we present flexible state space models for the extremes of univariate time series with non-stationary trends. While the models are presented through two example analyses, they may be applied to a wide range of problems that involve extreme value modelling of a non-stationary series. We show how the particle filters and smoothers presented in previous chapters may be applied to fit these models.

4.1 Analysis of women's 3000m running event

4.1.1 Introduction

We first analyse an athletics dataset, shown in Figure 4.2, of the fastest annual times in the women's 3000m running event since 1972.

Robinson and Tawn (1995) first studied the fastest times from 1972 to 1992 to assess whether Wang Junxia's record in 1993 was consistent with the previous

data. They used the r-smallest order statistics likelihood (see Subsection 2.3.1 for the r-largest model) with a parametric trend to conclude that cutting 16.51s off the record, though unusual, was not exceptionally so. Smith (1997) outlined the benefits of a Bayesian analysis for calculating the probability of beating Wang Junxia's record given that a new record is set and Gaetan and Grigoletto (2004) extended this by using particle methods to model a dynamic trend.

As stated in Subsection 2.3.4, Gaetan and Grigoletto (2004) propose a GEV model whose parameters vary stochastically over years by following random walks in a state space. They use only the fastest record of each year t and negate the value so that a GEV(μ_t, σ_t, ξ_t) distribution for maxima is appropriate. They model μ_t with either the first-order random walk $\mu_t|\mu_{t-1} \sim \mathcal{N}(\mu_{t-1}, \nu^2)$ or the second-order random walk $\mu_t|\mu_{t-1}, \mu_{t-2} \sim \mathcal{N}(2\mu_{t-1} - \mu_{t-2}, \nu^2)$; if the latter is used, the state is augmented to include μ_t and μ_{t-1} so that the Markov property can be maintained. Since they wish σ and ξ to be constant over time, they assume $\log(\sigma)$ and ξ follow first-order random walks with negligible variances of $\nu^2 = 0.001^2$.

While Gaetan and Grigoletto (2004) presented an attractive model for the data, it is our belief that the particle methods they used for their inference are highly inefficient. Because they use extremely small state variances for $\log(\sigma)$ and ξ, these parameters are practically constant and therefore, as discussed in Subsection 2.1.6, their components are only explored in the initialisation of the particle filter. This is demonstrated in Figure 4.1 which repeats their analysis using the 1000 particles they suggest is sufficient. While the particle representation looks good in 1973 after a single iteration, we see that after just 6 more steps there are perhaps only 9 distinct ξ values and by 1992 this has fallen further to just 4. While their method could easily be improved by using a lot more than the 1000 particles they propose, the speed at which the particle approximation degrades suggests that a huge number may be required.

We also note that the forward-backward particle smoother of Tanizaki (2001),

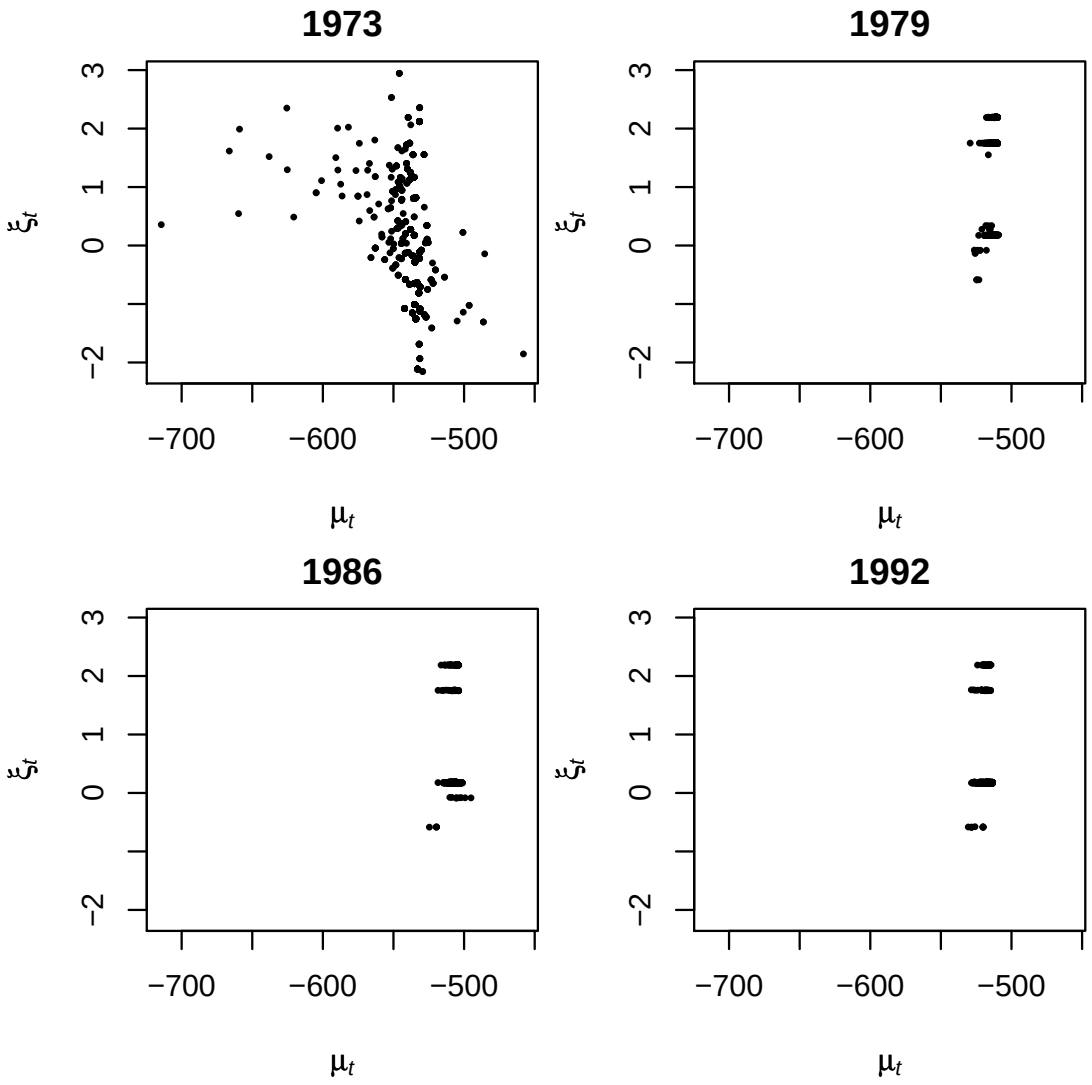


Figure 4.1: The μ_t and ξ_t components of 1000 particles from the filter at years 1973, 1979, 1983 and 1992 following the method of Gaetan and Grigoletto (2004). As time progresses the effective number of distinct ξ_t values decreases since the state variation is insufficient to create significantly different values to replace those lost to re-sampling.

that they propose to provide smooth μ_t estimates, works poorly with their model for μ. This is because their choice of second-order random walk creates a partial deterministic relationship between particles of adjacent years which, as shown in Subsection 3.2.1, causes forward-backward smoothers to degrade. Even without the second-order random walk, the constraint of an almost constant value of σ_t and ξ_t provides additional near-deterministic relationships between the particles.

4.1.2 Dynamic r-smallest order statistics model

We now propose a new model that aims to improve on the deficiencies of the Gaetan and Grigoletto (2004) approach.

Whereas Gaetan and Grigoletto (2004) used the annual minimum running times, we use the r-fastest annual times following the initial analysis of Robinson and Tawn (1995). Large amounts of data are now available on-line (for example from *Track and Field all-time Performances* at `http://www.alltime-athletics.com/`) from which the five fastest times of different athletes per year is shown in Figure 4.2. Since the data values within a year are the smallest from a population of independent running times, the r-smallest extremal order statistics model is most appropriate. However, since this model arises from the asymptotics of IID variables, we use only the best record achieved by each athlete within a year and discard any lesser times. For convenience, we refer to the data as the r-fastest times per year but we must remember that we mean the r-fastest times from different athletes per year. Following the theory of Subsection 2.3.1, the r-smallest model fit is equivalent to negating the data values and using the r-largest order statistics likelihood of (2.19).

To account for the non-stationarity visible in the series, Gaetan and Grigoletto (2004) used independent random walks for each of the three likelihood parameters in the state space. However, since they only wished to model non-stationarity in the location, they chose tiny variances for the other parameters to fix them through time which ultimately led to their method failing. Referring to other methods of modelling fixed parameters that we reviewed in Subsection 2.1.6, the method of Storvik (2002) and the rejuvenation of the sample with MCMC moves are perhaps the most promising. However, these both require sufficient statistics of σ and ξ to exist for their methods to be efficient, and for this model they do not. We therefore propose the simpler strategy of assuming they are fixed and known and use the EM algorithm of Section 3.3 to estimate them.

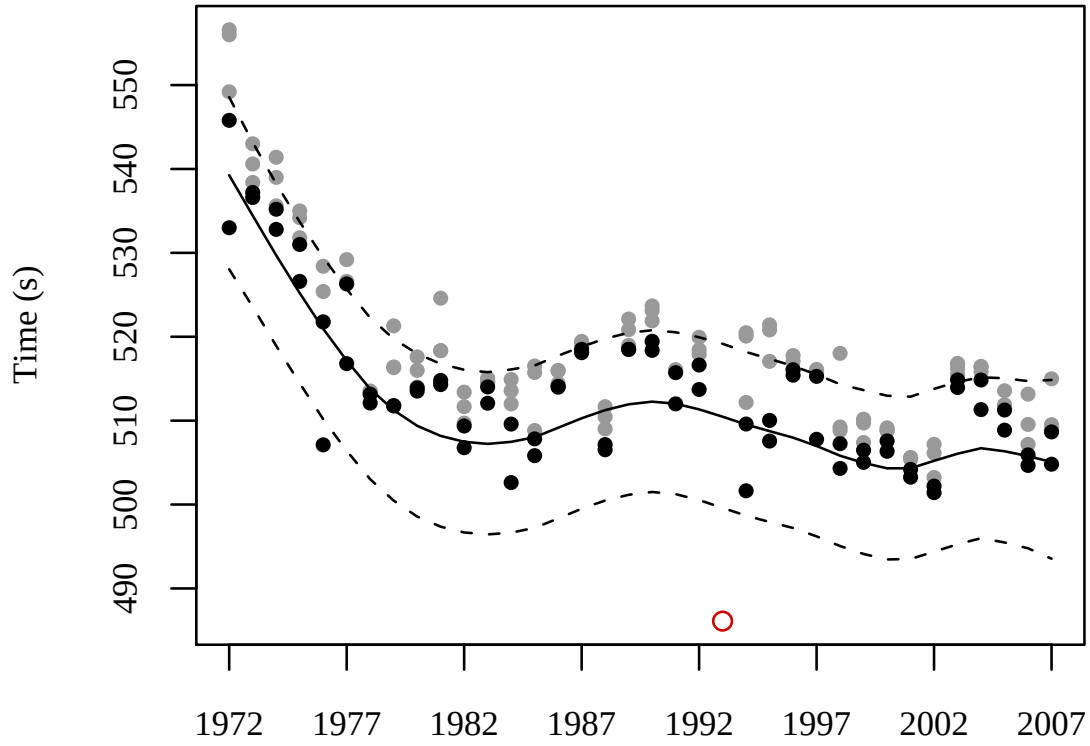


Figure 4.2: Five fastest times for the women's 3000m race between 1972 and 2007 with Wang Junxia's time in 1993. The two fastest annual times used for our fit are coloured black. Also shown is the mean and central 95% probability interval of the fitted predictive distribution for the fastest time per year.

Since the second order random walk for μ causes degeneracy problems with some smoothers, we instead adopt the smooth second order random walk given in the linear-Gaussian simulation study of Subsection 3.2.3. We therefore augment the state with $\dot{\mu}$, the velocity of μ, giving us the two-dimensional state $x_t = (\mu_t, \dot{\mu}_t)'$. Finally, for the prior we follow Gaetan and Grigoletto (2004) and use an uninformative normal distribution.

The aim of this analysis is to accurately estimate the probability of a new record in 1993 beating Wang Junxia's time of 486.11s (shown as a red circle in Figure 4.2). For this we need to approximate the marginal smoothing distribution of μ_{1993} for which we use our particle smoother of Section 3.2. Since the likelihood only depends on μ_t and the prior is Gaussian, we used Rao-Blackwellisation (see Subsection 2.1.5) to marginalise $\dot{\mu}_t$ thus improving the accuracy of the particle methods. Details of this step and the particle algorithms we used to achieve this

are given in Appendix A.4.

4.1.3 Parameter estimation

For a fixed value of r and ν^2 we can estimate the likelihood parameters σ and ξ using an EM algorithm constructed using our new smoother (see Algorithm 3.4 in Section 3.3 for details). The EM algorithm for (σ, ξ) was initialised with $(\hat{\sigma}, \hat{\xi})$ which were obtained using the following two-step procedure: We first fit the negated annual minima $(-y_t)$ to a Kalman smoother with the same prior and state as our model but with Gaussian observations to obtain trend estimates $\hat{x}_t$. As $(-Y_t) - \hat{x}_t$ should follow a GEV$(0, \sigma, \xi)$ distribution approximately, we then obtain maximum likelihood estimates $(\hat{\sigma}, \hat{\xi})$ of (σ, ξ) from a GEV fit to the $(-y_t) - \hat{x}_t$ data. For this fit and for the maximisation within the EM algorithm we work with the transformed variable $\log(\sigma)$ as this then spans the whole real line.

Simultaneously estimating ν^2 requires particles approximating the joint distribution $p(x_{t-1}, x_t | y_{1:T})$ which is possible using our smoother (as our algorithm gives approximations to $p(x_{t-1:t+1}|y_{1:T})$, see Subsection 3.2.2). It is perhaps simpler, however, to select among a few possible ν^2 by maximising the model likelihood $p(y_{1972:2007}|\nu^2)$, which we estimate using the formula of Kitagawa (1996) given by (2.10). Table 4.1 shows a selection of ν^2 values with the corresponding EM estimates of σ and ξ and the model likelihood when we take $r = 2$, using $N = 10,000$ particles.

To select the number of observations r to include per year we constructed probability-

ν^2	0.5	0.75	1	1.25	1.5	1.75	2
σ	4.36	4.25	4.22	4.15	4.12	4.04	4.01
ξ	-0.15	-0.13	-0.13	-0.11	-0.11	-0.09	-0.09
Likelihood ($\times 10^{-83}$)	0.41	2.72	3.68	3.50	2.41	1.52	1.02

Table 4.1: Model likelihood with σ and ξ estimates for different values of the smoothing parameter ν^2 and $r = 2$.

probability and quantile-quantile plots to assess the model fit. Looking at $r = 1, \ldots, 5$ we concluded that the best fit was obtained from only two observations per year. As we see from Table 4.1, this leads us to select $\nu^2 = 1$ and estimate σ and ξ to be 4.22 and -0.13 respectively. This fit is shown on Figure 4.2 as the estimated mean and central 95% probability interval of the negated annual maxima from each marginal smoothing distribution.

4.1.4 Results

To estimate the probability of a new record in 1993 beating Wang Junxia's we use the $r = 2$ fastest times from 1972 to 2007 excluding 1993, denoted $y_{1972:2007}$, to estimate the predictive distribution of the fastest time in 1993. Given this data and the parameters σ and ξ, the probability of Y_{1993}, the fastest time in 1993, beating Wang Junxia's time of 486.11s is given by

$$\begin{aligned}
&\mathbf{P}\{Y_{1993} \leq 486.11 | y_{1972:2007}, \sigma, \xi\} \\
&\qquad = \int \mathbf{P}\{-Y_{1993} > -486.11 | \mu_{1993}, \sigma, \xi\}\, p(\mu_{1993} | y_{1972:2007}, \sigma, \xi)\, \mathrm{d}\mu_{1993} \\
&\qquad = \int (1 - G(-486.11 | \mu_{1993}, \sigma, \xi))\, p(\mu_{1993} | y_{1972:2007}, \sigma, \xi)\, \mathrm{d}\mu_{1993} \\
&\qquad \simeq \sum_{i=1}^{N} (1 - G(-486.11 | \mu_{1993}^{(i)}, \sigma, \xi))\, w_{1993}^{(i)},
\end{aligned}$$

where G is the GEV cdf and $(\mu_{1993}^{(i)}, w_{1993}^{(i)})$ are weighted particles approximating $p(\mu_{1993} | y_{1972:2007}, \sigma, \xi)$. The probability that Y_{1993} beats Wang Junxia's time given it is a new world record is then

$$\begin{aligned}
&\mathbf{P}\{Y_{1993} \leq 486.11 | Y_{1993} \leq 502.62, y_{1972:2007}, \sigma, \xi\} \\
&\qquad = \frac{\mathbf{P}\{Y_{1993} \leq 486.11 | y_{1972:2007}, \sigma, \xi\}}{\mathbf{P}\{Y_{1993} \leq 502.62 | y_{1972:2007}, \sigma, \xi\}} \\
&\qquad \simeq \frac{\sum_{i=1}^{N} (1 - G(-486.11 | \mu_{1993}^{(i)}, \sigma, \xi))\, w_{1993}^{(i)}}{\sum_{i=1}^{N} (1 - G(-502.62 | \mu_{1993}^{(i)}, \sigma, \xi))\, w_{1993}^{(i)}},
\end{aligned} \tag{4.1}$$

where 502.62s was the world record prior to 1993.

Our analysis estimates the probability of a new record in 1993 beating Wang's to be 2.16×10^{-4}. This conflicts with the analysis of Gaetan and Grigoletto (2004) who showed Wang's record well within the reach of their boxplots of the conditional distribution. Apart from our doubts in the accuracy of their results, the main difference in the two analyses is that Gaetan and Grigoletto (2004) only used data on the fastest race for years up to 1992. If we repeat our analysis using only the two fastest times from 1972 to 1992 we obtain a probability estimate of 1.045×10^{-2}. While this is a lot larger than before, as no data ahead of 1993 is being used, it still disagrees with the previous conclusions of Gaetan and Grigoletto (2004).

We also admit that our analysis fails to account for the uncertainty in σ and ξ which could cause our estimate to be significantly under-estimated. However, since the attempts of Gaetan and Grigoletto (2004) to account for this lead to poor performance of the particle methods, a new approach is required. In Fearnhead et al. (2009) we consider this issue by combining likelihood estimates on a grid with priors for σ and ξ to produce a Bayesian probability estimate. Fitted using the data selected from 1972 to 2007, our estimate becomes 1.9×10^{-2} which, though naturally larger due to increased parameter uncertainty, still shows Wang's time to be fairly unlikely.

There are many ways our simple model can be enhanced to improve the probability estimate. Robinson and Tawn (1995) proposed many extensions including an additional effect to account for expected increases in performance during Olympic and World Championship years. Such an effect could be added to our model via the addition of indicator variables to the state process which then gives us more parameters to estimate with the EM algorithm. It should be noted that the women's 3000m was only a standard discipline at the major championships between 1983 and 1994 before it was replaced by the 5000m so therefore the additional term should only be applied during these years.

Our model assumes that times made in successive years are independent given the trend which is unrealistic since the fastest times are often made by the same athletes. Without the trend, we could simply remove all but the fastest time recorded over each athlete's career, but the non-stationarity in the series makes this more difficult. One solution would be to subtract the trend from our current fit and then keep only the fastest trend-adjusted times before re-fitting the modified dataset. Alternatively, we could allow multiple times per athlete and explicitly model the dependence across years using multivariate extreme value techniques.

Robinson and Tawn (1995) also proposed a model which linked the 3000m series with the similar women's 1500m event. In Section 5.1 we extend this analysis by jointly modelling dependencies between the two series in the trend and the extremes to refine our probability estimate.

4.2 Analysis of Antarctic temperature data

4.2.1 Introduction

We now study the extremal properties of a time series of temperature measurements. We have daily temperatures taken from the Faraday research station on the Antarctic peninsula from 1st January 1951 through to 31st December 1995 (in 1996 it was handed over to Ukraine and is now known as the Akademik Vernadsky Station). The dataset is complete with not one missing value. A plot of the raw data is given in Figure 4.3 with linear least-squares fits made to the annual maxima, mean and minima.

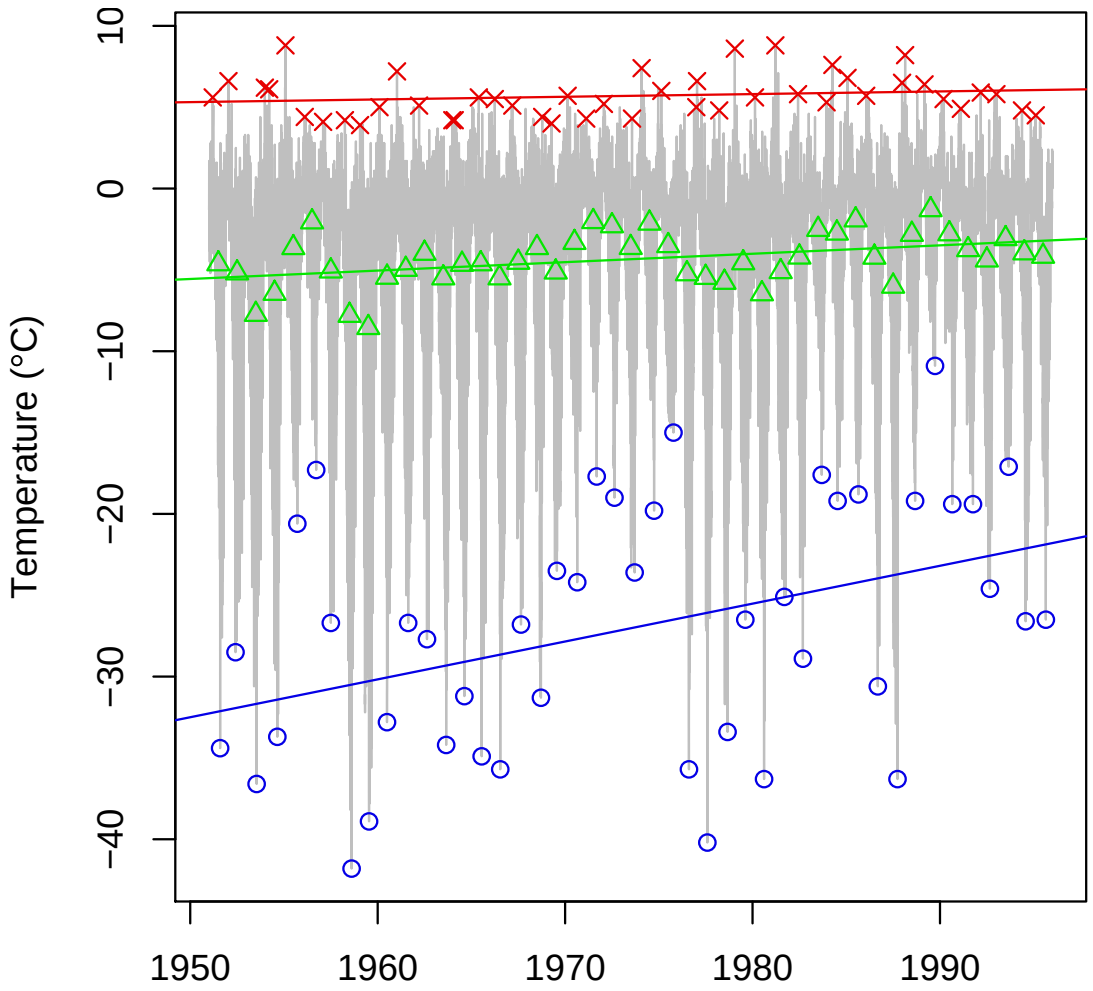


Figure 4.3: Raw Faraday temperature series showing annual maxima (×), mean (△) and minima (○) with linear least-squares fits.

The increasing trend in the mean of the process is much studied (see for example Vaughan et al. (2003)) but the change in the extremes is less well understood. From Figure 4.3 we see that a linear fit to the minima shows a clear upward trend at a much faster rate than the mean of the process. The maxima on the other hand appear to be stationary. However, it is unclear whether these effects are caused by the overall change in the bulk of the series or by further variation in the extremes.

To attempt to answer this question, we will fit extreme value models with dynamic trends to the upper and lower extreme values of the series. However, we first standardise the dataset to remove the overall trend as well as the seasonality present in the series. We can then differentiate between temperatures that are extreme only because they fall in the summer or winter and those that are extreme given the current trend.

4.2.2 Standardising the dataset

Kernel smoothing to remove seasonality

To standardise the data we use kernel smoothing to produce a smooth mean and variance which we then use to normalise the series. We use a non-parametric method to account for the variety of factors that cause the seasonality which we may miss if we attempt a parametric fit. From Figure 4.3 we can see a clear local trend that follows the seasons but also a possible year-to-year effect that we attempt to capture. It is also evident that the winters are more variable than the summers so we produce a smooth variance estimate as well as the mean.

Writing $y_{d,t}$ to denote the Faraday temperature measurement on day d of year t, we estimate the smoothed mean value on this day by the weighted sum

$$\hat{y}_{d,t} := \sum_{e,s} y_{e,s}\, w_{e,s}(d,t)$$

using weights

$$w_{e,s}(d,t) :\propto \phi\left(\frac{d-e}{b_1}\right)\phi\left(\frac{t-s}{b_2}\right), \tag{4.2}$$

normalised to sum to 1 (where $\phi(z)$ is the pdf of a standard normal variable evaluated at z). It is assumed that for days near the beginning or end of a year we wrap around to consecutive years so that $d-e$ is the smallest difference between days in a year. The amount of smoothing is determined by the bandwidths $b_1 > 0$ and $b_2 > 0$. We similarly smooth the variance using

$$\sigma^2_{d,t} := \sum_{e,s}(y_{e,s} - \hat{y}_{d,t})^2\, w_{e,s}(d,t)$$

where we use the same weights as the smoothed mean given in (4.2).

To select the bandwidths b_1 and b_2 to use for both the smoothed mean and variance we use cross-validation. For a range of different bandwidths, we randomly remove 10% of the data and calculate the smoothed mean $\hat{y}_{d,t}$ at these values. We then sum the squared difference between the removed data and its smoothed mean estimate and compare this value with that of other bandwidths choosing that which minimises the sum of squares.

Figure 4.4 shows the logged sum of squares for a range of b_1 and b_2 values. The minimum value is obtained at $b_1 = 0.88$ days and $b_2 = 0.39$ years. The resulting smoothed mean and variance are displayed in Figure 4.5. We can see that the smoothed mean still contains a lot of structure but the additional noise in the raw data is mirrored by the variable standard deviation.

Having smoothed the mean and variance of the series, we can standardise the series using

$$z_{d,t} := \frac{y_{d,t} - \hat{y}_{d,t}}{\sqrt{\sigma^2_{d,t}}}.$$

This should remove the seasonality and overall trend of the series captured by the smoothed variables to allow us to study the extremes of the residuals. The

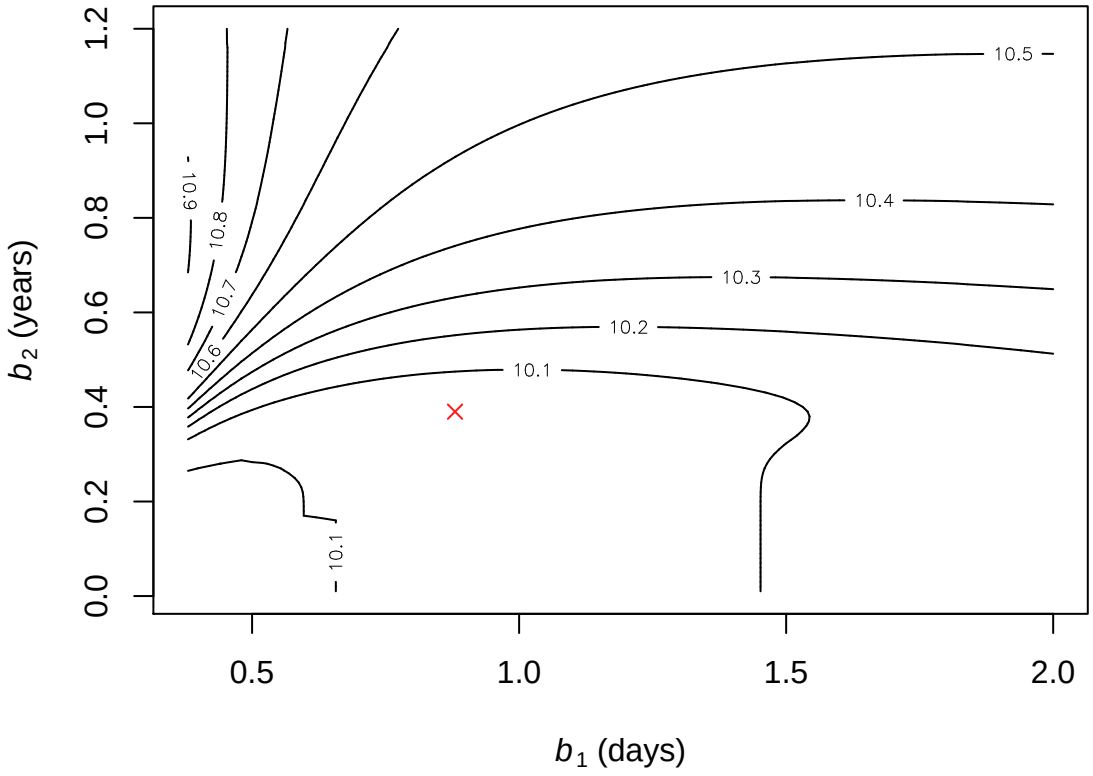


Figure 4.4: Logged sum of squares of the discrepancy between the data removed for cross-validation and its smoothed estimate for a variety of kernel bandwidths b_1 and b_2. The minimum sum of squares is shown as a red cross (×).

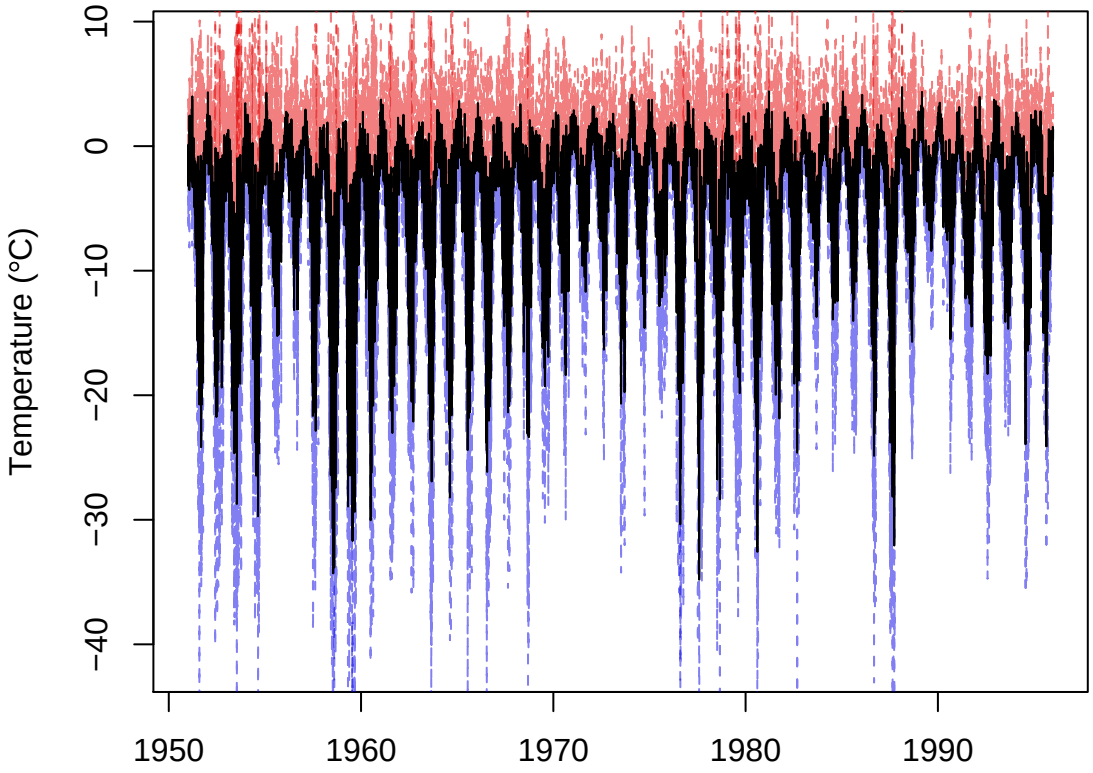


Figure 4.5: Smoothed Faraday temperature series $\hat{y}_{d,t}$ (—) with confidence interval of $\hat{y}_{d,t} - 2\sigma_{d,t}$ (– –) and $\hat{y}_{d,t} + 2\sigma_{d,t}$ (– –), where $\sigma_{d,t}$ is the smoothed standard deviation.

standardised series for 1958 is shown in Figure 4.6.

Declustering to account for temporal dependence

We now focus on modelling the upper and lower extremes of the standardised dataset. Though de-trended, we still expect an amount of dependence to remain in the series and therefore, following the theory of Subsection 2.3.3, we should first decluster the series. We can then model the cluster maxima and minima as if they are independent observations.

To decluster the standardised series, we use the runs method of Smith and Weissman (1994). This involves selecting a high threshold u and identifying clusters of threshold exceedances by consecutive runs of points separated by at least κ values below u. To analyse the lower extremes we simply decluster the upper extremes of the negated series.

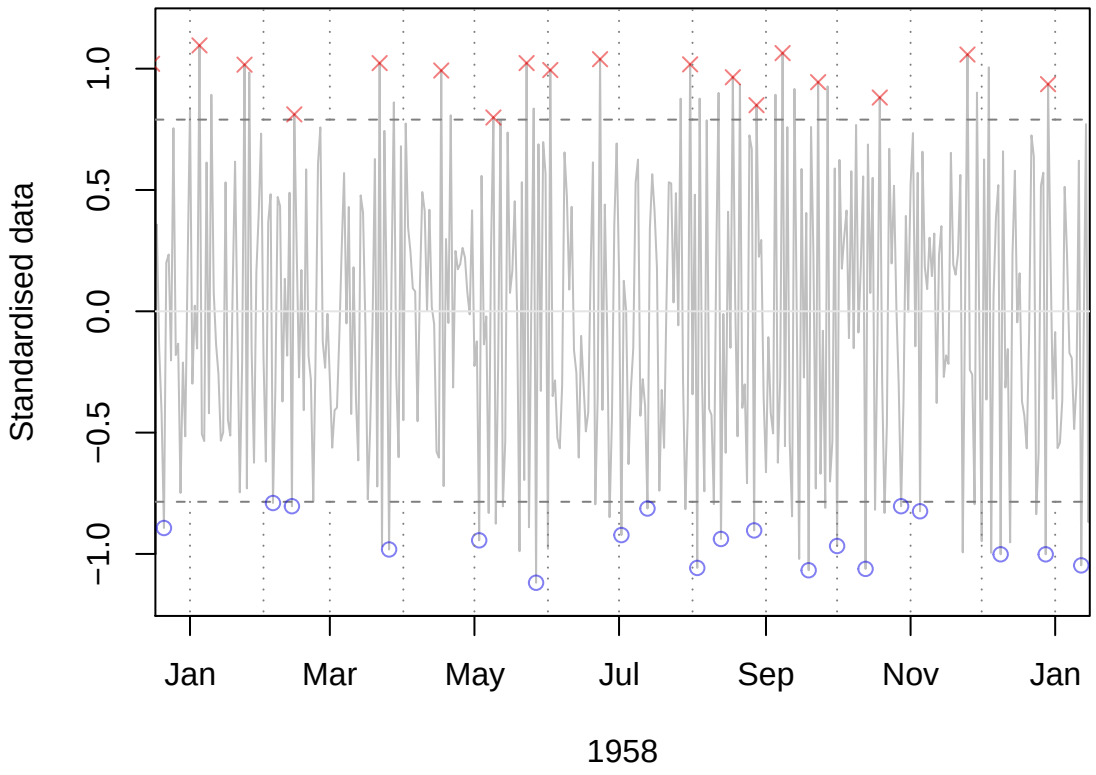


Figure 4.6: Standardised Faraday temperature series $z_{d,t}$ for $t = 1958$ with declustered threshold exceedances for the upper and lower tail. The thresholds u are shown by the dashed lines (- -) and the monthly boundaries by $(\cdots)$.

We select the threshold u to be the upper 90% quantile of the series and pick the run length κ by eye, looking to identify independent clusters. As a compromise between missing clusters by using too large a run length and having multiple points in a cluster with too small, we chose $\kappa = 7$ days for both the upper and lower extremes. The consequences of this choice can be seen in Figure 4.6 which identifies the cluster maxima for 1958.

4.2.3 Dynamic point process model

Following the theory of Subsection 2.3.3 we can model the cluster maxima (and minima) as if they were independent observations. For this we use the point process model (2.22) as a starting point, choosing to model with the GEV parameters μ, σ, ξ rather than use the peaks over threshold characterisation (see Subsection 2.3.2). Mirroring the athletics analysis of Section 4.1, we will allow μ to vary over time while fixing the remaining GEV parameters.

To construct a state space model for μ the decision of how to discretise the series into blocks has to be made. Note that this was not an issue with the women's 3000m analysis since the dataset was given as the fastest times within a year. Since the temperature series is recorded as daily measurements, we may choose to allow μ to vary between days. However, we have 45 years of observations so this would lead to a large computational burden, especially if smoothing estimates are required since particles then need to be stored for each time step. We also recall that we are modelling not the time series itself but the cluster maxima of those values which exceed the threshold u. We therefore have mostly missing data when a resolution of a single day is used.

Unless we expect μ to vary substantially between days, it is reasonable for us to assume μ is constant over a block of time. To allow the possibility of comparing the model fits of various block sizes, we propose a model whose parameters are

independent of the particular discretisation deployed. Following the point process model (2.22) of Subsection 2.3.2, we propose the following observation density $g(\mathbf{y}_t|\mu_t, \sigma, \xi)$ for cluster maxima $y_{t,1}, \ldots, y_{t,n_t}$ in block t:

$$\exp\left(-\Delta t\left[1+\xi\left(\frac{u-\mu_t}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right)\prod_{j=1}^{n_t}\left[1+\xi\left(\frac{y_{t,j}-\mu_t}{\sigma}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)}, \tag{4.3}$$

where Δt is length of time (in years) covered by block t. Note that the number n_t of threshold exceedance cluster maxima varies between blocks and may be 0. By measuring t in years we ensure that the GEV parameters correspond to a year for any discretisation we use.

For the state we again adopt the smooth second order random walk with state $x_t = (\mu_t, \dot{\mu}_t)'$ that we used in the women's 3000m analysis as well as the simulation study of Subsection 3.2.3. However, we need to extend the stated form to allow models with different block sizes to be comparable with one another. Since the model is derived from a pair of stochastic differential equations (3.16) given on page 84, this goal is achieved by using the distribution of $X_{t+\Delta t}$ given X_t where Δt is now the time between successive blocks. In Appendix B.1 this is shown to be

$$X_{t+\Delta t}|\{X_t = x_t\} \sim \mathcal{N}(F_{\Delta t}\, x_t, Q_{\Delta t}), \tag{4.4}$$

where

$$F_{\Delta t} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}, \qquad Q_{\Delta t} = \nu^2 \begin{pmatrix} \frac{(\Delta t)^3}{3} & \frac{(\Delta t)^2}{2} \\ \frac{(\Delta t)^2}{2} & \Delta t \end{pmatrix}. \tag{4.5}$$

Since we assume μ_t is constant over the block labelled by t, we take t to be the time in the centre of the block and therefore Δt in the state transition is the time difference between the centres of two successive blocks. This contrasts with Δt in the observation density (4.3) which denotes the size of the block t. These two will be equal as long as the block size remains constant over the series. However, this

distinction allows us to specify the prior at $t = 1951.0$ rather than at one time step before the first to allow the prior to remain independent of the discretisation. This is achieved by using $\Delta t/2$ in place of Δt in the state transition for the first time step since the state has half the distance to travel to get to the centre of the first block.

For the prior itself, we again choose a Gaussian distribution with large variances to represent our uncertainty in the parameters. However, since the prior choice and the form of the state transition determine the a-priori distribution of the state at future time steps $p(x_t)$, we explore the consequences of the prior correlation. Indeed, it can be shown that selecting an independent prior covariance will not give independent covariances for future time steps. Instead, the correlation of $p(x_t)$ tends towards that of the state covariance $Q_{\Delta t}$ (which itself is independent of Δt) as t increases.

This prompts us to choose a prior covariance matrix with the same correlation structure as $Q_{\Delta t}$. This, in turn, implies that the prior covariance should be $\Sigma_0 = Q_c$ for some large value of c so that prior uncertainty is still captured. In doing this we are making the assumption that the a-priori state correlation remains constant over time. This is similar to the common practice of assuming stationarity by setting the prior to be the stationary distribution of the state when it exists.

We note that this causes the prior to depend upon the state parameter ν which must be remembered when it comes to its estimation. We also note that the prior covariance of Q_c implies that $p(x_t)$ has covariance Q_{c+t} so that all a-priori state densities take this form.

We aim with this analysis to judge whether μ varies over time and so we require the smoothed estimates of μ_t for every time step. Once again we use our new particle smoother of Section 3.2 which in turn uses forwards and backwards particle filters. The similarity of this model to that proposed for the women's 3000m analysis of Section 4.1 means that our algorithms implementations are very similar. This is

summarised in Appendix A.5.

4.2.4 Results

We now apply our model to the threshold exceedance cluster maxima (and minima) we obtained from the standardised dataset. We first chose to discretise the series into months (setting $\Delta t = 1/12$) as this allows the model to capture any remaining seasonality whilst not giving too many time steps which would cause a large computational burden. Before applying our method we must select the observation parameters σ and ξ as well as the state parameter ν.

In the previous section we used an EM algorithm to estimate σ and ξ conditionally on ν, although it is also possible to estimate these jointly with Algorithm 3.5. While in principle we could use the EM algorithm here, it will take a lot longer since we now have 540 time steps. We therefore estimate the observation parameters by fitting a simpler model by maximum likelihood.

Specifically, we fit the point process likelihood (4.3) with $\mu_t = c_0 + c_1 t + c_2 \sin(2\pi t) + c_3 \cos(2\pi t)$. This gives μ a simple trend with seasonal variability which should be close enough to the smooth μ fit to give good σ and ξ estimates. Implementing this we obtain $\sigma = 0.02456$, $\xi = -0.8702$ for the upper tail and $\sigma = 0.02870$, $\xi = -0.7906$ for the lower fit. These significant negative values for ξ give a short tailed distribution for the threshold exceedances that has an upper bound close to μ.

Because we wish to compare our variable μ fit with a constant value for μ, we pick a value of the state parameter ν that gives the trend a reasonable amount of flexibility. We could, alternatively, use the model to estimate ν but since σ and ξ were estimated from a simpler model it is likely that the fitted ν value would be small thus restricting the potential variability of the trend.

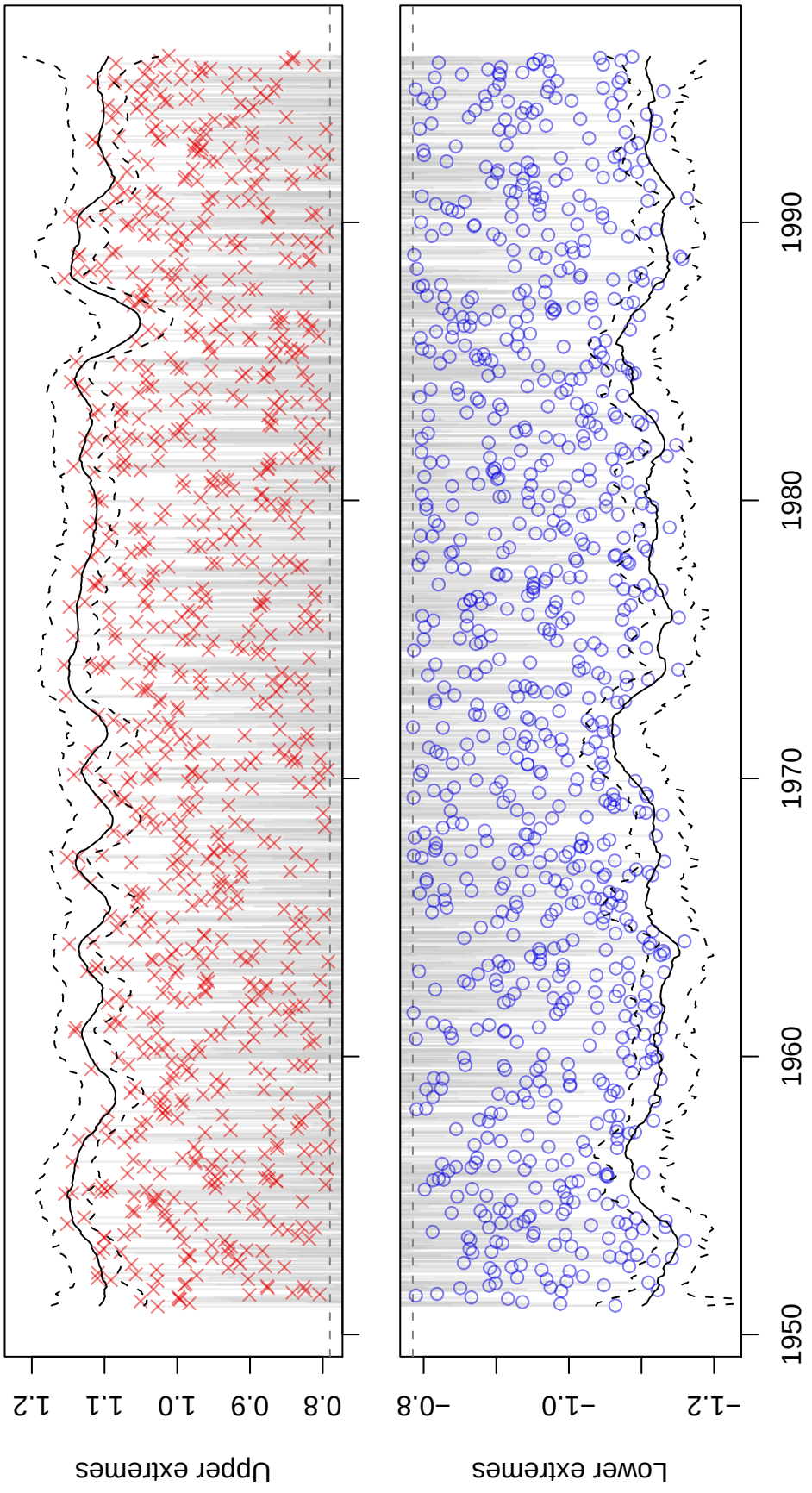


Figure 4.7: Standardised Faraday temperature series with declustered threshold exceedances and fitted trend μ_t for the upper and lower tail. The μ_t fit is represented by the mean and central 95% probability interval of the smoothing distribution. Also shown are the constant thresholds u.

Figure 4.7 shows the fitted μ_t trend from our state space model when we select $\nu = 0.1$. We can see from the variability of the μ_t estimate that this gives the trend some flexibility to follow the observations but not too much so that we overfit the data. While the expected value of μ_t varies considerably, we see that the 95% probability intervals are wide enough to allow a straight line to be put between them over large expanses of time.

For the upper tail it is only around 1987 that the fitted μ_t falls below a potential constant value of around 1.13. This suggests that the upper extreme temperatures at Faraday have little additional structure above the overall trend in the bulk of the data. Any processes bringing about change in the overall trend seem therefore to have the same effect on the warmest temperatures.

While the variability in the lower trend appears to be no greater than that of the upper trend, it is harder to place a constant value of μ through the lower probability intervals. This irregularity in the lower tail can be seen in the original dataset in Figure 4.3 which shows far greater variability in the lower tail than the upper tail. Comparing the fitted μ_t trend with the linear fit of the annual minimum temperatures, we see that regions of unusually warm winters correspond well to peaks in the fitted trend with the possible exception of around 1990. Here the summers were cool so the annual variability was low which absorbed the effect of the warm summer at normalisation. The greater variability in the lower tail suggests that the average trend in the bulk of the series is not enough to describe the distribution of the coldest temperatures.

To better account for the skew in the original dataset, we could repeat the analysis using data standardised with smoothed median and quartiles rather than the mean and variance. This should produce a better normalisation and would also lessen the chance of extremes in one tail affecting the other's final fit. To achieve this, the same weights given by (4.2) could be used to smooth over days as well as years. Given a weighted sample (y_i, w_i), the p%-quantile could be estimated by first

ordering the pairs by increasing y_i (giving $(y_{(i)}, w_{(i)})$ where $y_{(i)} \leq y_{(i+1)}$), and then selecting the jth ordered sample $y_{(j)}$ where $j = \arg\min_k \left| p - \sum_{i=1}^{k} w_{(i)} \right|$. Then, given first, second and third quartile estimates, $Q_{1,i}$, $Q_{2,i}$ and $Q_{3,i}$ respectively, observations above their median $Q_{2,i}$ could be standardised as $(y_i - Q_{2,i})/(Q_{3,i} - Q_{2,i})$ and those below as $(y_i - Q_{2,i})/(Q_{1,i} - Q_{2,i})$. The discontinuity at $Q_{2,i}$ would not matter since we model only the upper and lower extremes of the standardised series.

There are many more ways in which our temperature analysis and the dynamic point process model in general can be improved. We are in particular limited by the need to pre-process the data to remove the seasonality before the extreme values are modelled; this separation confines us to looking at the remaining non-stationarity in the extremes above that in the bulk of the data, restricting the conclusions that can be drawn on the original scale.

In the athletics analysis of the previous section we were able to model the dataset directly since it arrived as the r-smallest values in a year and we could assume that performances within and across years were independent given the trend. The temperature series, however, arises as raw, dependent data from which independent extreme values must be extracted. The usual strategy of thresholding the series before declustering the threshold exceedances is only appropriate when the seasonality is removed. It is therefore primarily this step that needs to be modified for the raw data to be taken into the model.

One way of doing this would be to decluster each block separately with a threshold that varies with each block (perhaps taking a quantile of the data within a block rather than for the series as a whole). This would allow the threshold to vary between blocks following the seasonality of the observations allowing the cluster maxima in each block to be fitted directly. Inferences could then be drawn on the original scale to, for example, answer questions about the overall trend in the extremes of the Faraday series.

A potential problem with this method is the identifiability of clusters at the edge of each block. Where two blocks meet with differing threshold values, an observation considered extreme on one side may not be so on the other. Also, cluster maxima at the edge of a block may better belong with its neighbour if the cluster itself is spread between blocks. These issues may be resolved by taking the neighbouring blocks into account when each block is declustered. Also, if a simple smooth threshold could be constructed from the data alone then this would be preferable to a threshold that jumps between blocks.

Chapter 5

State space modelling of bivariate extremes

In this final chapter we extend the univariate state space models of Chapter 4 to allow the extreme values of two series to be analysed together. This allows for joint extremal dependencies to be modelled as well as correlations between the two trends. While the models are presented for bivariate analyses it is possible to extended them to higher dimensional problems.

5.1 Pooling athletics data from two events

5.1.1 Introduction

In this section we extend the women's 3000m analysis of Section 4.1 by jointly modelling the women's 1500m running event. By capturing the dependence between the two series, we aim to improve our estimate of the probability of a new 3000m record in 1993 beating Wang Junxia's controversial time.

Robinson and Tawn (1995) first considered a joint analysis of the 3000m series with that of a similar event. The 1500m was chosen as it is run in the same style as the 3000m unlike shorter sprint events or longer distance races. Their method consisted of a linkage between the two events, assuming that the annual minima for the 3000m is roughly twice that of the 1500m with a lag to account for a lag in development between the two races. By exploiting this relationship, dependencies between the GEV parameters are introduced.

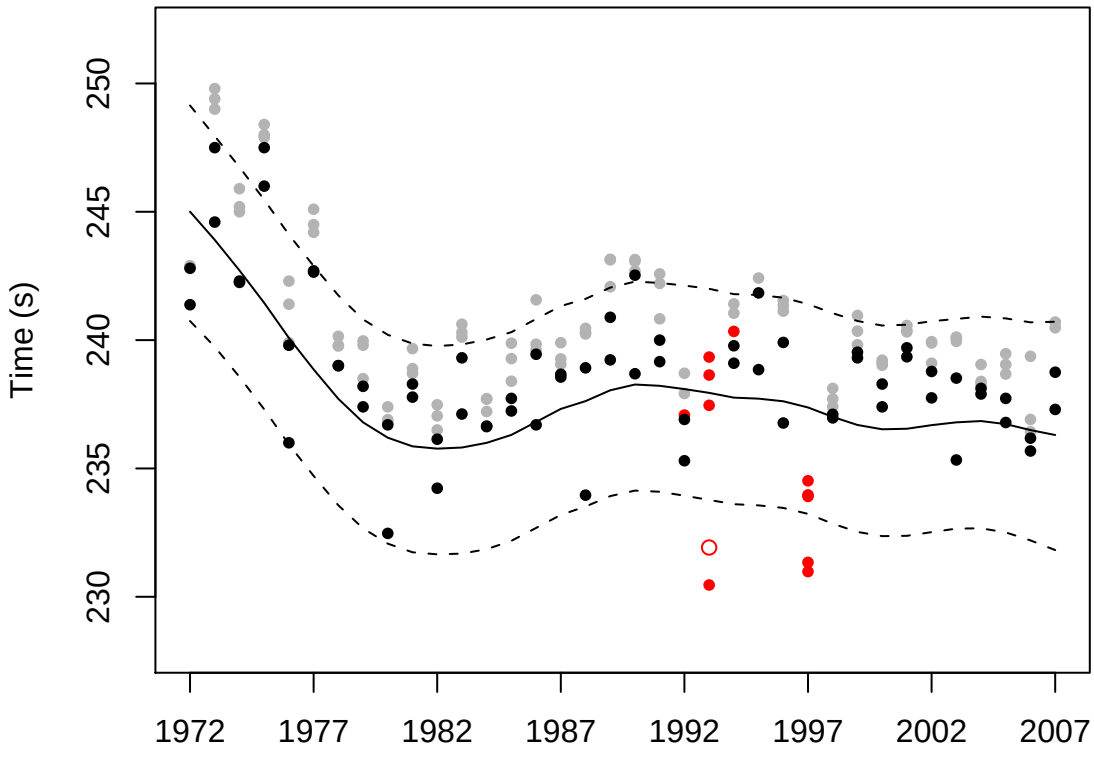
One problem with this approach is that the times run in the 1500m are assumed to be independent, given the GEV parameters, of those made in the 3000m. However, since the events are similar, many athletes will run both events; for example, Wang Junxia also ran the 1500m in 1993 achieving the second fastest time that year (see Figure 5.1). With this in mind, Barão and Tawn (1999) extend the analysis of Robinson and Tawn (1995) by using a joint model based on the bivariate logistic distribution for the fastest annual times in each series. To account for the non-stationarity, they used the parametric trend of Robinson and Tawn (1995) for each series.

While this approach provides a better treatment of the dependence within a year, little connection is made between the trends of the two series. We therefore aim to model the joint extremal properties of these two events as well as capturing any dependence in the overall trend.

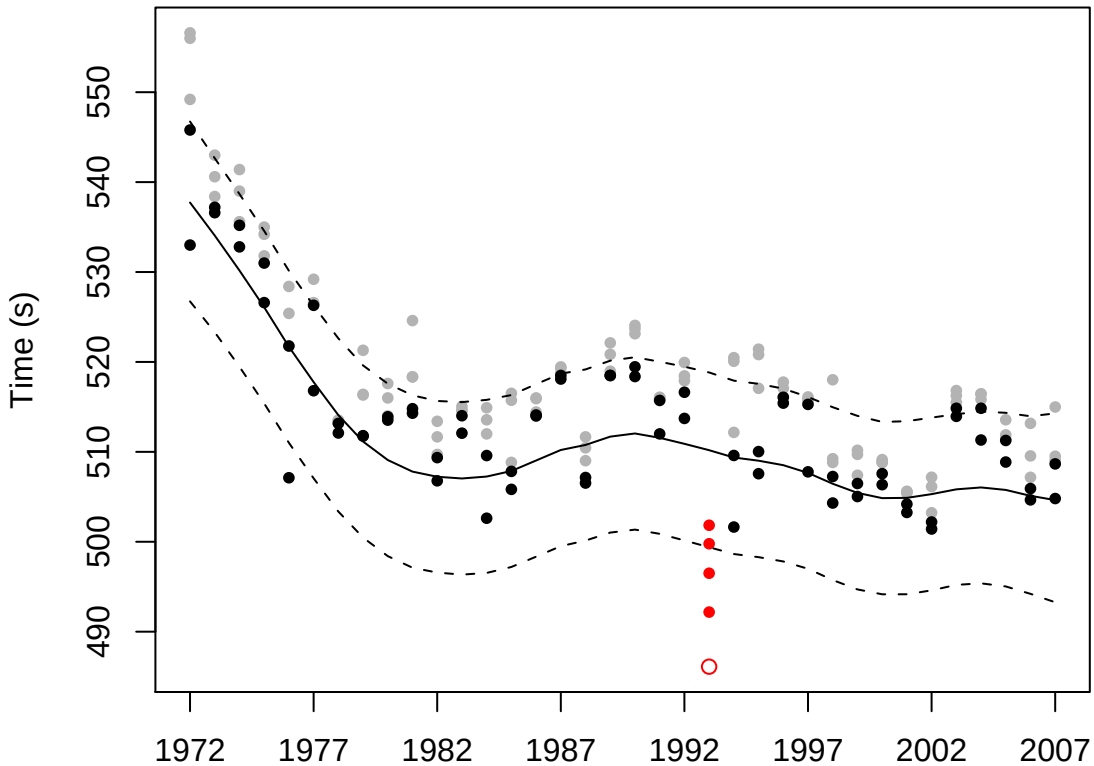
5.1.2 Dynamic logistic model with correlated random walks

We now describe our alternative joint model for the 1500m and 3000m series.

Firstly, we consider a joint distribution for the p and q-fastest annual times run in the 1500m and 3000m respectively. Referring to Section 2.4, we propose a bivariate extreme value model for the negated series. This takes each series as the marginal extremes of an underlying paired process where each athlete runs in



(a) Women's 1500m



(b) Women's 3000m

Figure 5.1: Five fastest times for the women's 1500m and 3000m races between 1972 and 2007. Those made by Chinese athletes are shown in red with Wang Junxia's times drawn as circles. Also shown is the mean and central 95% probability interval of the predictive distribution for the fastest time per year. The two fastest annual times that were used for this fit are coloured black.

both the 1500m and the 3000m races. This seems reasonable as many athletes run both races and those which do not are unlikely to have achieved an exceptionally fast time if they did. We again make sure that multiple times by the same athlete within a year are removed to preserve the IID assumption of the underlying annual process, and for convenience we refer to this dataset as the p and q-fastest times per year.

To provide a closed-form expression for the joint pdf, we are required to choose a parametric form for the spectral distribution H or equivalently for V. For this we pick the bivariate exchangeable logistic model with V given in Fréchet margins by (2.25) (where we have $d = 2$). This is chosen for its simplicity as it has a single dependence parameter $\alpha \in (0, 1]$ and the series is possibly too short to allow a more complex parametrisation. The exchangeability assumption (on common margins) is also reasonable since the two series are of similar running events as opposed to two very different measurements.

Using the point process characterisation of Subsection 2.4.1, we can derive the joint distribution of the p and q-largest variables from the logistic model. This can then be differentiated multiple times to produce the joint density function. The marginal distributions should then be transformed from Fréchet to the GEV forms we know are appropriate for each series. This gives us an overall annual likelihood $g(\mathbf{y}_{1500}, \mathbf{y}_{3000}|\mu_{1500}, \sigma_{1500}, \xi_{1500}, \mu_{3000}, \sigma_{3000}, \xi_{3000}, \alpha)$ with marginal parameters μ, σ and ξ for each series and a dependence parameter α. Details of this derivation are given in Appendix C.

This joint observation model is consistent with our previous analysis since marginally the r-largest order statistic likelihood is obtained for each series. We would like to also propose a consistent model for the state (by marginally giving the same random walk for μ_{3000}) but would also like to capture possible dependence between the two series in the trend.

We therefore propose to derive the state model from the following four-dimensional

stochastic differential equation:

$$\begin{aligned} \mathrm{d}X_{t,1} &= X_{t,3}\,\mathrm{d}t, \\ \mathrm{d}X_{t,2} &= X_{t,4}\,\mathrm{d}t, \\ \mathrm{d}X_{t,3} &= \nu_{1500}\,\mathrm{d}B_{t,1}, \\ \mathrm{d}X_{t,4} &= \nu_{3000}\rho\,\mathrm{d}B_{t,1} + \nu_{3000}\sqrt{1-\rho^2}\,\mathrm{d}B_{t,2}, \end{aligned} \tag{5.1}$$

where $X_t = (\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})'$ and $\mathrm{d}B_{t,j}$ are independent Wiener processes. Marginally $(\mu_{1500,t}, \dot{\mu}_{1500,t})'$ and $(\mu_{3000,t},\ \dot{\mu}_{3000,t})'$ both have the same form as the SDE (3.16) used in the previous 3000m analysis with the μ variable being the integrated path of the random walk $\dot{\mu}$. Jointly the random walks $\dot{\mu}_{1500}$ and $\dot{\mu}_{3000}$ have correlation ρ which will correlate the trends between the two series.

The SDE (5.1) has solution given by

$$X_{t+\Delta t}|\{X_t = x_t\} \sim \mathcal{N}(F_{\Delta t}\,x_t, Q_{\Delta t}), \tag{5.2}$$

where

$$F_{\Delta t} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad Q_{\Delta t} = \begin{pmatrix} S\frac{(\Delta t)^3}{3} & S\frac{(\Delta t)^2}{2} \\ S\frac{(\Delta t)^2}{2} & S\Delta t \end{pmatrix}, \tag{5.3}$$

with

$$S := \begin{pmatrix} \nu_{1500}^2 & \nu_{1500}\nu_{3000}\rho \\ \nu_{1500}\nu_{3000}\rho & \nu_{3000}^2 \end{pmatrix}.$$

The proof of this is outlined in Appendix B.2. Measuring time t in years, this gives us the Gaussian state transition density $f(x_{t+1}|x_t, \nu_{1500}, \nu_{3000}, \rho)$.

For the prior, we once again use a multivariate Gaussian distribution with large variances to account for our uncertainty in the trend.

With this model we aim to improve our estimate of the probability of a new 3000m record in 1993 exceeding Wang Junxia's time. As with previous analyses, this probability is used as evidence to assess the consistency of this record to the surrounding data. Referring to Figure 5.1, we see that Wang's 3000m record was not the only exceptional time made by a Chinese athlete during the 1990s. Indeed, many of these Chinese performances caused controversy at the time. We therefore make two fits to the data: one using all the data except 1993 and another after all the Chinese records have been removed.

As before, the calculation of this probability requires the smoothing distribution $g(\mu_{3000}|\mathbf{y}_{1972:2007})$, where we use $\mathbf{y}_{1972:2007}$ to generically represent all the data we are using for the fit. To estimate this distribution, we again use our new particle smoother of Section 3.2 which makes use of forwards and backwards filters. Since our observation density only depends upon the μ components of the state, we use Rao-Blackwellisation to marginalise the $\dot{\mu}_{1500}$ and $\dot{\mu}_{3000}$ components for improved accuracy. Details of the particle methods implementation is given in Appendix A.6.

5.1.3 Parameter estimation

Before we can calculate the target probability, we must estimate all the parameters in the model. We have marginal parameters σ, ξ and ν for each series as well as the correlation ρ between the states and the dependence parameter α. We also have to select p and q, the number of observations to include each year.

Beginning with p and q, we recall from Subsection 4.1.3 that for the 3000m analysis we used probability-probability and quantile-quantile plots to select the two fastest times per year. This gives us $q = 2$. To select p we did the same for the 1500m series, both excluding 1993's data and excluding all Chinese athletes, and found that using the two fastest times per year fitted best in both cases. We therefore take $p = q = 2$ for all years except 1997 where $p = 0$ when all Chinese athletes are

removed and 1993 where both sets of observations are taken as missing.

Having chosen which observations to use for each series, we now estimate the marginal parameters by fitting each series with the marginal model of Section 4.1. To do this we can estimate σ, ξ and ν jointly with the joint EM algorithm Section 3.3. Recall that we have previously estimated these parameters for the 3000m series excluding 1993 and obtained $\sigma_{3000} = 4.22$, $\xi_{3000} = -0.13$ and $\nu_{3000} = 1$. For the 1500m series, the EM algorithm gives $\sigma_{1500} = 2.1792$, $\xi_{1500} = -0.1485$ and $\nu_{1500} = 0.3831$ when we exclude 1993 and $\sigma_{1500} = 2.0033$, $\xi_{1500} = -0.2629$ and $\nu_{1500} = 0.3462$ when all Chinese athletes are removed.

To estimate the remaining parameters ρ and α we maximise the model likelihood $p(\mathbf{y}_{1972:2007}|\rho, \alpha)$. Using the approximation given by (2.10), we evaluate the likelihood on a grid of $\rho \in [0, 1)$ and $\alpha \in (0, 1]$ values. Figure 5.2 gives a contour plot of the log likelihood over a subset of these when we fit the model omitting all

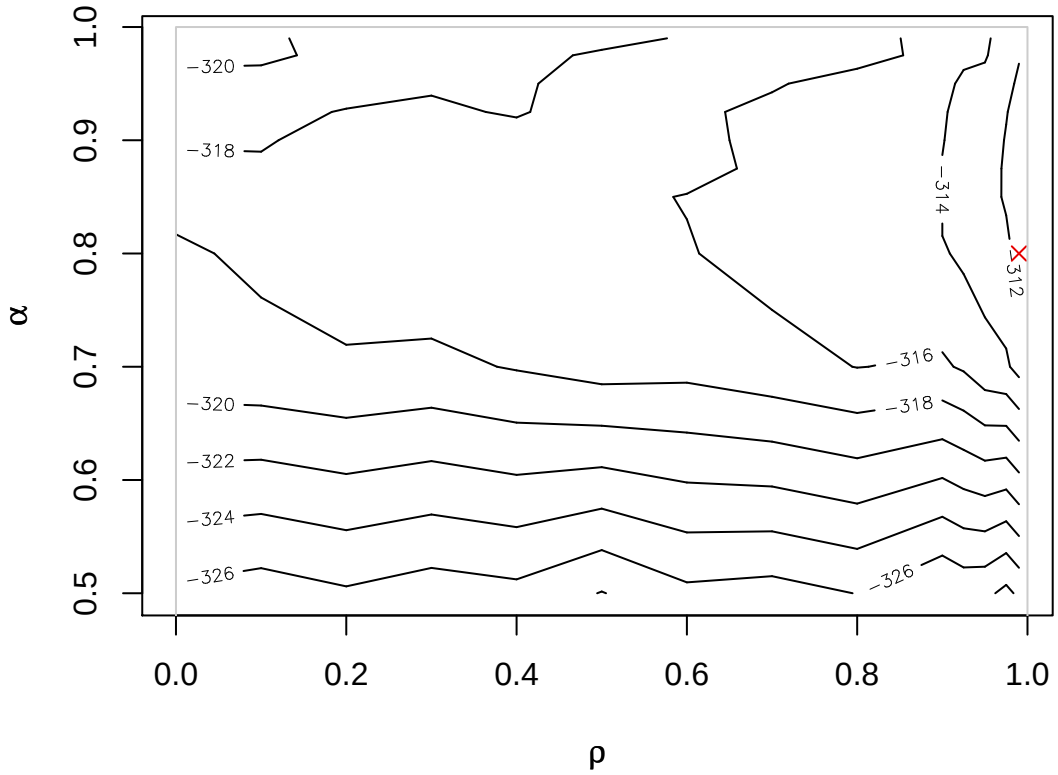


Figure 5.2: Contour plot of the model log likelihood for a range of dependence parameters ρ and α. The model was fitted with the Chinese athletes removed and the remaining parameters taken from marginal fits. The maximum value of the likelihood is displayed as a red cross (×).

Chinese athletes. We see that the maximum is obtained with $\alpha = 0.8$ and ρ very close to 1. This suggests that the model is overparametrised and the trend can be described by a single random walk. However, rather than restructure the model, we simply set $\rho = 0.99$ as this should have a similar effect. When we exclude only the 1993 values from the dataset, we obtain $\alpha = 0.875$ and have similar issues with ρ so again take the value of 0.99 for this parameter.

5.1.4 Results

Now we have selected all the parameters in the model, we run our smoothing algorithm again to revise our estimate of the probability of a new 3000m record in 1993 beating Wang Junxia's time. Having run the smoother for 1993, we estimate this probability as before using the formula (4.1) with the 3000m versions of σ and ξ.

Using all but the 1993 data we obtain a probability estimate of 1.50×10^{-4} and excluding the remaining Chinese records in 1997 from the fit we obtain 1.73×10^{-4}. This second fit is demonstrated in Figure 5.1 through the predictive distribution of the annual best time. We see that the exclusion of two extreme 1500m records in 1997 has little effect on the extreme tail for the 3000m series four years earlier as the extreme probability changes very little. This is despite the large correlation between the two trends.

Recall that the previous value of our target probability obtained from the marginal analysis of Section 4.1 was 2.16×10^{-4}. By exploiting dependencies with the 1500m series we have tightened the distribution of the fastest annual record reducing the mass in the tails. This adds extra weight to support the hypothesis that Wang Junxia's record is inconsistent with the achievements of her fellow athletes.

To further test this claim, our analysis could be repeated with a variety of alternative models for the spectral distribution function H such as those reviewed in

Kotz and Nadarajah (2000). However, with only 36 years of observations we may not have enough information to justify the use of a dependence model with very many parameters. Alternatively, other similar events such as the 5000m or the 10000m could be added into the model by mirroring the correlated state and the multivariate logistic observation model in higher dimensional cases.

If we believe that the Chinese athletes, as claimed, represent an enhanced population due to their high-altitude training and alternative diet, we could add their times to the model with a term that accounts for their potential edge over athletes of other nationalities. This could be done either as an additive term per event that indicates the expected reduction in a Chinese running time or as a multiplicative term indicating the factor by which a Chinese athlete's time is reduced. We could then produce revised probability estimates to compare the chances of Chinese and non-Chinese athletes in 1993 producing Wang's record.

5.2 Joint analysis of sea-level data

5.2.1 Introduction

In this final section we consider the variability of extremal dependencies between sea-level heights at pairs of sites along the eastern English coastline.

We have frequent sea-level data from January 1964 to April 2008 from three sites along the eastern coast of England (displayed in Figure 5.3). The records up to 1st January 1993 were taken every 60 minutes with later values taken every 15 minutes. Much data is missing as can be seen in Figure 5.4 which displays the monthly maximum sea-level surges for the months that contain no missing values. Data is missing due to a variety of causes such as gauges being replaced but also after a quality assessment has marked a value improbable.

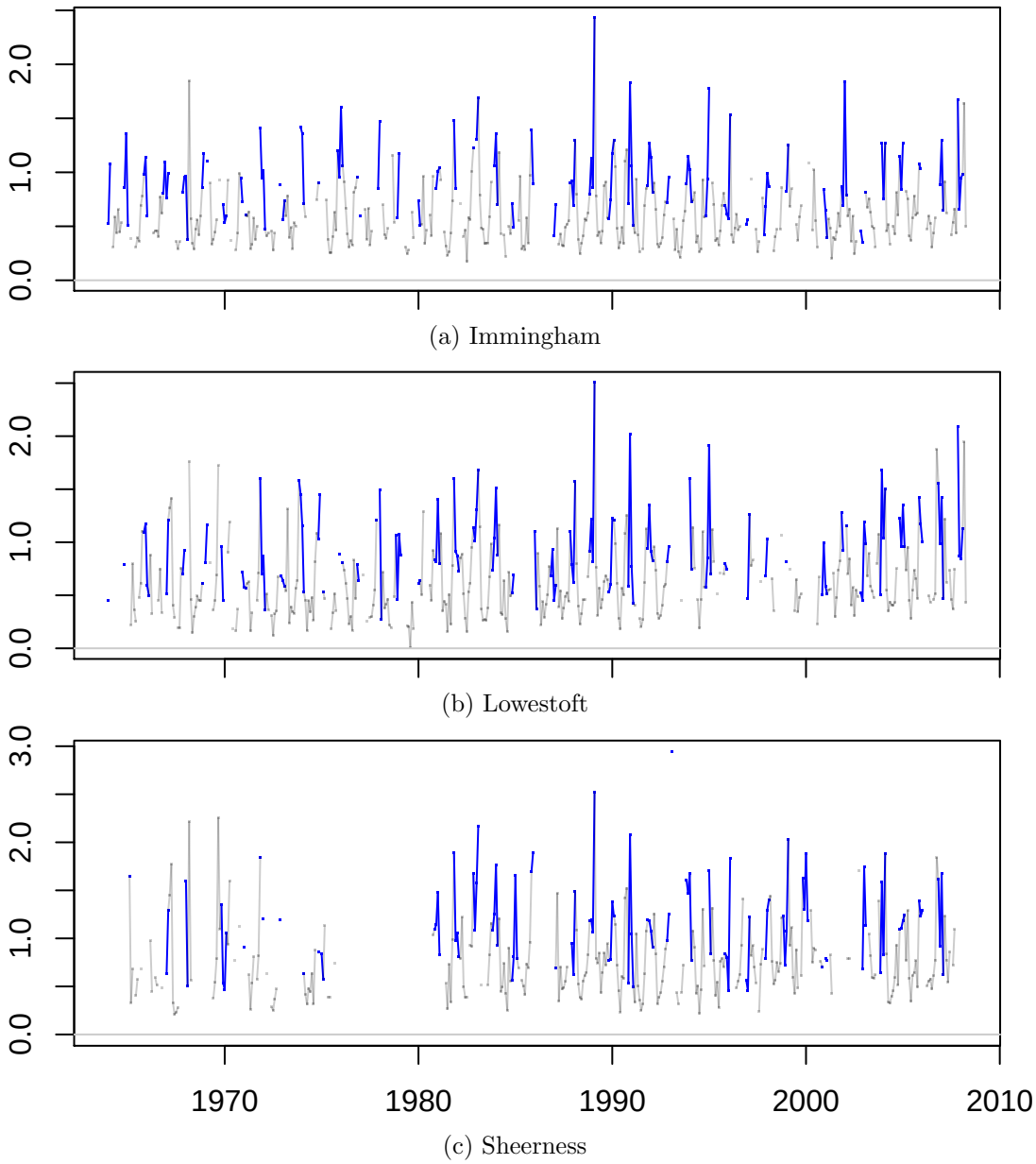
The datasets are freely available from the UK National Tide Gauge Network via the British Oceanographic Data Centre website (`http://www.bodc.ac.uk/`). As well as the raw sea-level readings, surges are provided that are calculated as the raw sea-level minus the predicted tide. These predictions are calculated from a database of tidal constants maintained by the Proudman Oceanographic Laboratory's Application Group.

Extreme sea-levels are much studied in the literature with many authors focusing on sites along the eastern English coastline. Analyses are often centred on the estimation of upper quantiles of the annual maximum sea-level distribution to aid the design of coastal defences. Tawn (1992) for example proposes an r-largest model for the sea-level surges at a single site that incorporates interactions with the tide level. Multivariate extreme value distributions are used by Tawn (1988a, 1990) and Barão and Tawn (1999) to simultaneously fit the annual maximum sea-levels at multiple sites while spatial models are proposed by Coles and Tawn (1990) and Dixon and Tawn (1992).

Figure 5.3: Map showing the locations of our sea-level data sources along the eastern coastline of England.

Many authors assume that after the tide has been subtracted from the recorded sea-level, the remaining surges form a stationary sequence. Dixon and Tawn (1999) consider the implications of this and show that a false stationarity assumption can lead to a significant underestimation of the extremal tail.

Our interest lies in the dependencies between sea-level surges at pairs of sites and how this changes with time. We see from Figure 5.4 that there is a strong relationship between the monthly maximum surges; the maximum in February



(a) Immingham

(b) Lowestoft

(c) Sheerness

Figure 5.4: Monthly maximum sea-level surges for each of the three locations. The values corresponding to the winter months (Nov–Feb) are shown in blue. Maxima are only shown for months that contain no missing values.

1989 for example is the largest value in the Immingham and Lowestoft series and the second largest for Sheerness. By eye there seems to be strong synchronicity between the location of the largest maxima as well as their size. It is harder, however, to judge whether these dependencies vary over time.

5.2.2 Dynamic logistic model with variable dependence parameter

In order to find out whether the dependence between extremal sea-levels at a pair of sites changes with time, we extend the dynamic logistic model of Subsection 5.1.2 allowing the dependence parameter α to vary. This allows us to model the non-stationarity of each series in such a way that the extremal dependence between the surges at two sites can evolve smoothly as time progresses. In Appendix C we outline the derivation of the joint distribution of the p and q-largest variables from the logistic model. Since this model is derived from the asymptotics of IID variables, we must first account for the seasonality and dependence present in each series.

For the Antarctic temperature analysis of Section 4.2 we adjusted for the seasonality by pre-processing the whole series via kernel smoothing. We could do something similar here to the raw sea-level data but as a simpler strategy we use the surge values provided which subtract the predicted tide. This does not account for seasonal variations in the variance, however, as can be seen from Figure 5.4 which shows that the monthly mean surge values are usually larger in the winter. To account for this extra effect we restrict our analysis to the winter months of November to February inclusive.

As the sea-levels are measured every 15 or 60 minutes there are clearly high levels of temporal dependence in each series. Following the theory of Subsection 2.3.3, each series should be declustered and we can then model the cluster maxima as if they are independent observations. For the temperature analysis we used the runs method for declustering since this ties in with the point process model through the use of the threshold u. However, since we intend on modelling the r-largest independent observations in each series, we have no need of a threshold for the model fitting.

We therefore use the blocks method to decluster each series. This involves sequentially selecting the largest value in the series and then removing this and all neighbouring observations within a radius of size κ. This ensures that the selected observations are separated by at least κ which should be chosen to be just large enough to identify individual clusters. We can allow for the change of resolution from one measurement per 60 minutes to one per 15 by ensuring that κ denotes a time window around each cluster maxima rather than a number of observations. Trying a range of values, we select $\kappa = 7$ days for all series as this seems to select the majority of significant peaks while ensuring that they are separate enough to be deemed independent. The consequences of this choice can be seen in Figure 5.5 for Immingham in 1989.

Having declustered each series, we can now model the p and q-largest cluster maxima within a given block of time (not to be confused with the blocks used to decluster the series). To do this we take each month as a block but only allow

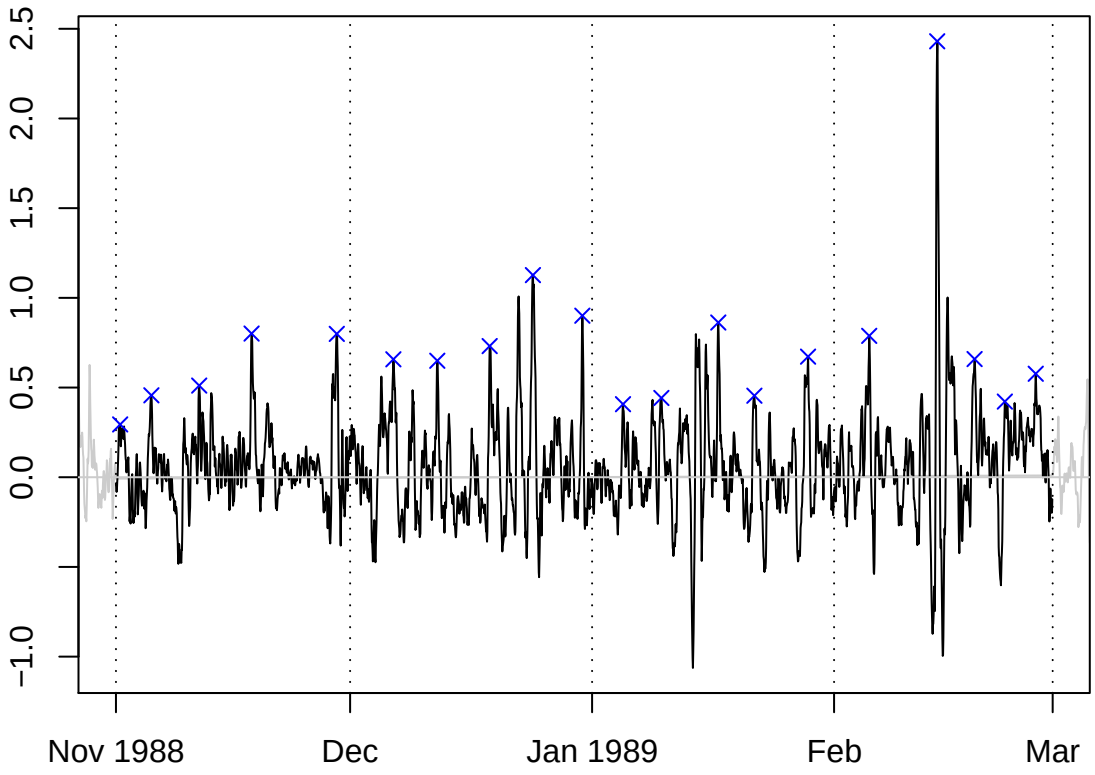


Figure 5.5: Sea-level surges at Immingham during the winter containing January 1989 with the five largest monthly cluster maximum obtained using blocks of radius $\kappa = 7$ days.

the state to vary across years. This gives us four independent blocks of winter observations per year, each of which consists of the p and q-largest values at two given sites. However, since much of the data is missing, we remove a whole month of observations at a site if there are any missing values there. As we can see in Figure 5.4 this causes many months to have no data and so we benefit from having four potential months of data pairs each year.

Our observation density for a given year is therefore the product of four p and q-largest logistic model densities, one for each of the months of November to February. When data is missing at one of the sites in a given month either p or q will be 0 and the density reduces to the r-largest density for the other site. If data is missing at both sites simultaneously that month's term is removed. Labelling each winter by the year that begins in the containing January we have 44 years of observations from 1965 to 2008.

To model any remaining non-stationarity in the location parameters μ_{A} and μ_{B} we use the correlated random walk of the previous section derived from the stochastic differential equation (5.1). This allows each μ to follow a smooth two-dimensional random walk but also allows these trends to be correlated through the use of a correlation parameter ρ_μ. Through estimating ρ_μ we get an indication of the connection between the margins at a pair of sites to contrast with α which indicates the dependence between the extreme values themselves.

To allow the extremal dependence parameter α to vary over time, we add α_t to the state X_t and propose a transition model that is independent of the model for μ. Since $\alpha \in (0, 1]$ we first transform to the real line by defining $\alpha^* := \Phi^{-1}(\alpha)$ where $\Phi(\cdot)$ is the standard Gaussian cdf. While this does not allow $\alpha = 1$ exactly, α may be arbitrarily close to 1 given a sufficiently large α^*.

To allow α to vary we could propose a simple random walk for α^* such as $\alpha^*_{t+1}|\alpha^*_t \sim \mathcal{N}(\alpha^*_t, \nu^2_\alpha)$. However, by allowing α^*_t to move away from 0 without restriction, it could become so far from 0 that α_t will be fixed at either 0 or 1 for even modest

variations in α_t^*. This would be particularly likely for sites that have asymptotically independent surges, ie when $\alpha_t \simeq 1$.

As an alternative, we propose a conditional model for α_{t+1}^* given α_t^* via a joint model for α_t^* and α_{t+1}^*. Specifically, we assume that $(\alpha_t^*, \alpha_{t+1}^*)$ follows a standard bivariate Gaussian distribution with correlation ρ_α. This implies that $\alpha_{t+1}^*|\alpha_t^* \sim \mathcal{N}(\rho_\alpha \alpha_t^*, 1-\rho_\alpha^2)$ which we take as our transition density. Assuming $\rho_\alpha \in (0,1)$, this model ensures that α_t^* can always move back towards 0.

We note that α^* now has a stationary distribution of $\mathcal{N}(0,1)$ which translates to an uninformative uniform distribution between 0 and 1 for α. We also note that this model may be generalised to $\alpha_{t+\Delta t}^*|\alpha_t^* \sim \mathcal{N}(\rho_\alpha^{\Delta t}\alpha_t^*, 1-\rho_\alpha^{2\Delta t})$ which gives a consistent model that allows time steps of arbitrary length.

We therefore have the combined state $X_t = (\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \dot{\mu}_{\mathrm{A},t}, \dot{\mu}_{\mathrm{B},t}, \alpha_t^*)'$ with a transition model that may be written as

$$X_{t+\Delta t}|\{X_t = x_t\} \sim \mathcal{N}(F_{\Delta t}\, x_t, Q_{\Delta t}), \tag{5.4}$$

where

$$F_{\Delta t} = \begin{pmatrix} 1 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \rho_\alpha^{\Delta t} \end{pmatrix}, \quad Q_{\Delta t} = \begin{pmatrix} S\frac{(\Delta t)^3}{3} & S\frac{(\Delta t)^2}{2} & 0 \\ S\frac{(\Delta t)^2}{2} & S\Delta t & 0 \\ 0 & 0 & 1-\rho_\alpha^{2\Delta t} \end{pmatrix}, \tag{5.5}$$

with

$$S := \begin{pmatrix} \nu_{\mathrm{A}}^2 & \nu_{\mathrm{A}}\nu_{\mathrm{B}}\rho_\mu \\ \nu_{\mathrm{A}}\nu_{\mathrm{B}}\rho_\mu & \nu_{\mathrm{B}}^2 \end{pmatrix}.$$

Measuring time t in years, we take $\Delta t = 1$ in the state to allow the trends and dependence parameter to vary from one winter to the next.

Note that since we have taken monthly blocks of data without introducing Δt into the observation model, the marginal parameters will correspond to a GEV distribution for the monthly maxima rather than the annual maxima. Since the annual maximum is the largest of the monthly maxima, the monthly parameters may be transformed to annual by equating the GEV distributions $G_{\text{ann}}(x) = \prod_{i=1}^{12} G_i(x)$, where G_i is the distribution of the ith monthly maxima. Since we only use the four winter months where it is likely the largest surges will occur, it may be best to equate only these four months to a single GEV distribution representing each winter maxima. We ignore these issues for now since we are only interested in the dependence between sites.

For the prior, we once again use Q_c for some large value of c to construct the prior covariance as this allows us to account for our uncertainty in the parameters while assuming a constant correlation structure. However we do this only for the trend components and their velocities as a compromise for them not having a stationary distribution. Since α_t^* has a stationary distribution of $\mathcal{N}(0,1)$ we take this for its prior which translates nicely to a non-informative uniform prior for α.

Our aim in this analysis is to fit the declustered surges from pairs of sites with a particle smoother to study how the smooth distribution of α_t varies over time. For this we use our new smoothing algorithm in its block sampling form which requires forwards and backwards particle filters. Details of our implementation of these algorithms are given in Appendix A.7.

5.2.3 Parameter estimation

To see how the amount of dependence between two sites varies with their separation distance we fit two pairs of sites: Immingham against Lowestoft and Immingham against Sheerness. Lowestoft is roughly half way between Immingham and Sheerness as can be seen in Figure 5.3. Immingham is chosen for both fits since it has

the fewest missing values.

To fit each pair of sites we must estimate the marginal parameters for each site as well as the correlation ρ_μ between each pair of random walks and the correlation ρ_α between successive α_t^*s. For this we follow a similar strategy to the joint athletics analysis in previous section (see Subsection 5.1.3).

We begin by selecting the number of observations p and q to include from each site per month, when available. For these we select 2 since much data is missing so we want to maximise the amount of information available and we do not have the observation density for p or q greater than 2. These could in principle be derived following a similar derivation to that in Appendix C but the joint distribution becomes increasingly complex making distributions of larger p and q hard to obtain. If more observations are required or if the dimension d is greater than 2, the joint density of all observations with $\sum_{i=1}^{d} y_i > u$ is easier to derive (see Subsection 2.4.1) but this is not considered here.

Having selected to use the two largest cluster maxima per month, we first estimate the parameters σ, ξ and ν for each site. For this we fit the marginal model with the joint EM algorithm of Section 3.3. The marginal model is the same as used for the athletics analysis of Section 4.1 except that the data are not negated since we are now fitting the upper extremes and the single observation density $g(y_{1:r}|\mu)$ is replaced by the product of four such densities, one for each winter month per year. The parameter estimates we obtained for each site are given in Table 5.1

We once again estimate the dependence parameters by maximising the model like-

Parameter	Immingham	Lowestoft	Sheerness
σ	0.3493	0.4382	0.4935
ξ	-0.1243	-0.1896	-0.1617
ν	0.0078	0.0078	0.0043

Table 5.1: Marginal parameter estimates obtained for each site from an EM algorithm.

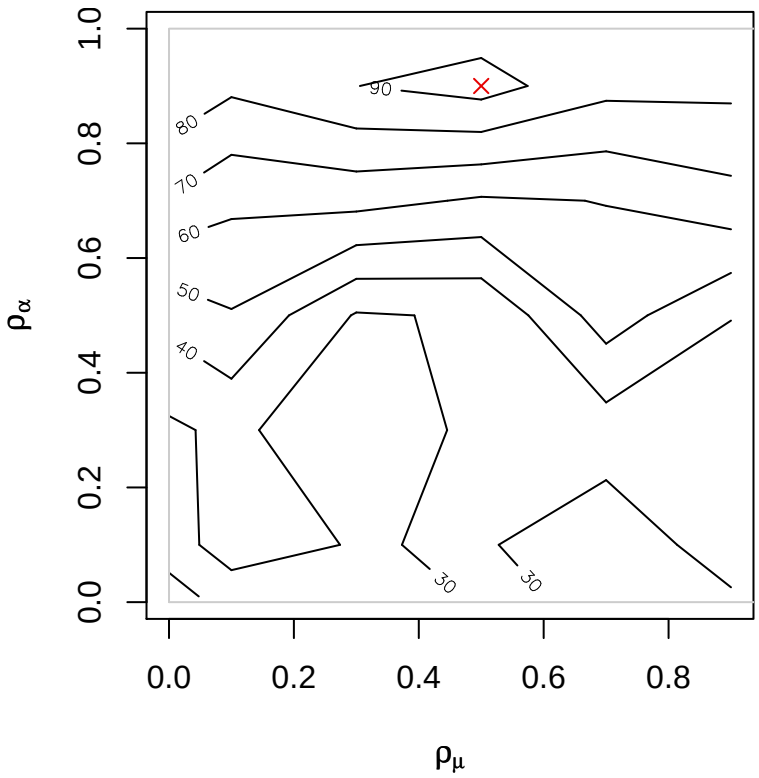


Figure 5.6: Contour plot of the model log likelihood for a range of dependence parameters ρ_μ and ρ_α with data from Immingham and Lowestoft. The maximum value of the likelihood is displayed as a red cross ($\times$).

lihood $p(\mathbf{y}_{1965:2008}|\rho_\mu, \rho_\alpha)$ using the approximation given in (2.10). The likelihood is estimated on a grid of values within $\rho_\mu, \rho_\alpha \in [0, 1)$. Figure 5.6 shows the log likelihood contour for the ρ_μ and ρ_α values fitted with data from Immingham and Lowestoft. The maximum is obtained from $\rho_\mu = 0.5$ and $\rho_\alpha = 0.9$. Similarly with Immingham and Sheerness we obtain $\rho_\mu = 0.1$ and $\rho_\alpha = 0.99$. We note that the correlation ρ_μ between the location parameter trends for each site is greater when Immingham is compared with Lowestoft which is to be expected since they are much closer to each other.

5.2.4 Results

We now estimate the marginal smoothing distributions for each pair of sites using our new parameter estimates. For this we use our block smoothing algorithm and found that block sizes of around 20 were most efficient. This is in particular

large enough to cover the periods of missing data present in the Sheerness series. The fitted distribution of α as well as of each location parameter μ is given in Figure 5.7 for the Immingham and Lowestoft fit and in Figure 5.8 for Immingham and Sheerness. Note that small values of α correspond to higher dependence in the extremes between each series.

Looking first at the fitted location parameters, it is immediately clear that they vary little from year to year which is a consequence of the small estimates for their noise parameters ν. In three out of four cases a constant value can be placed between the 95% probability intervals suggesting the model may be overparametrised but with the Lowestoft series there is some evidence to suggest that the extreme surges have been getting worse over the last ten years or so.

Focusing on the dependence parameters α we see that, while there are local variations especially with the fit containing Lowestoft, the overall dependence remains fairly stable over time. However, we can see the effect of the differing correlations between neighbouring years by the greater variability in the fit with Lowestoft. With this fit you can clearly make out troughs during years of agreement between the two sites and peaks when large extremes at one site are not accompanied by similar values in the other. While a smoother fit with larger confidence intervals could have been made to this α, the fact that this fit had a higher likelihood suggests that, while there is still a strong relationship between neighbouring values, some years' observations disagree strongly on the amount of extremal dependence with those a few years away.

Comparing with the second fit that includes Sheerness, we see that on average α is lower with the Lowestoft fit indicating the dependence between extreme surges is greater. This is expected since Immingham is closer to Lowestoft but we may have predicted this effect to be more pronounced. The comparable overall levels of α, particularly in the first half of the series, indicate a similar level of dependence between both pairs of sites which may suggest that the largest surges are often

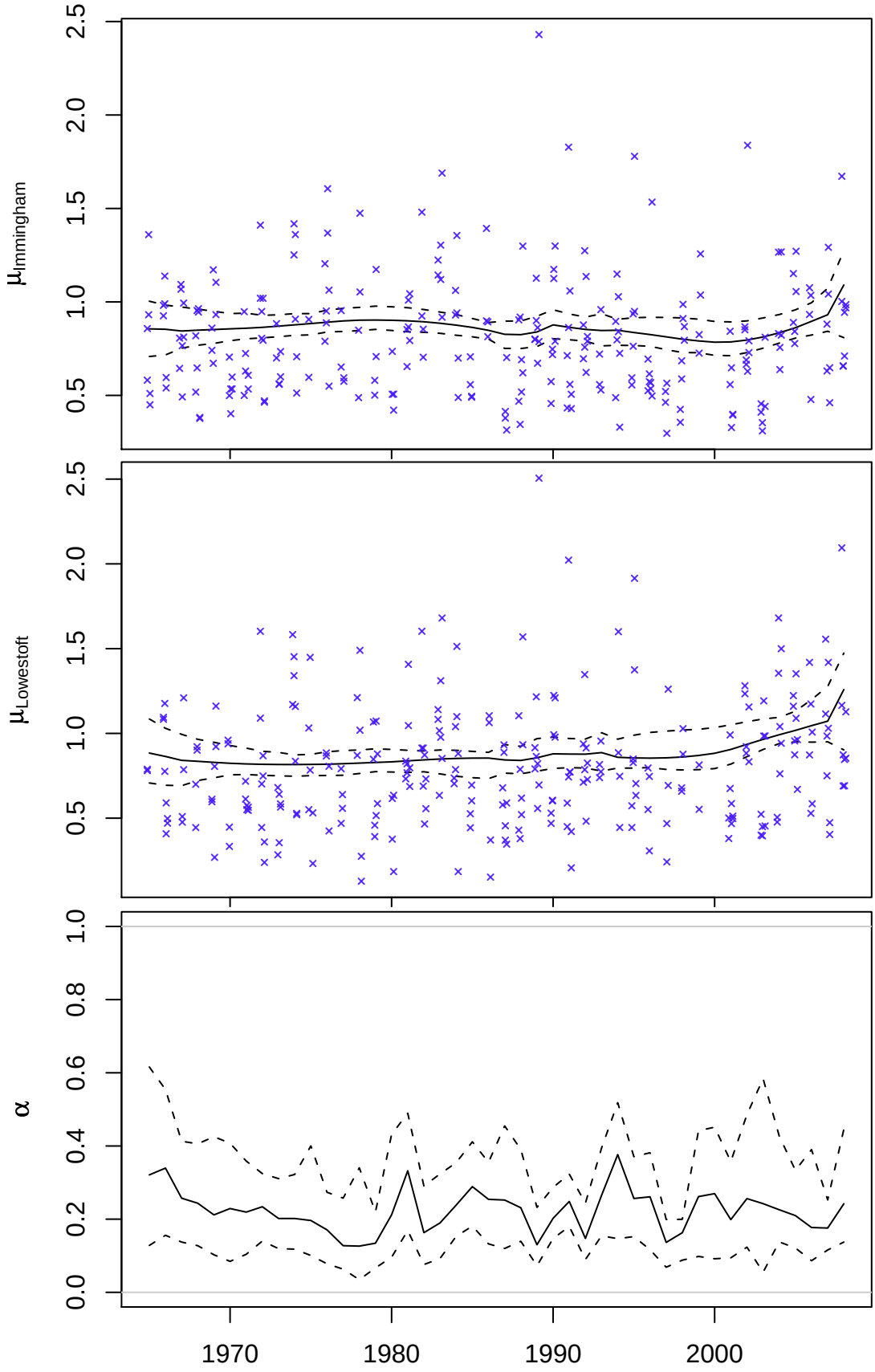


Figure 5.7: Smooth joint fit for Immingham and Lowestoft of the two largest cluster maxima during each of the winter months. The fitted smoothing distributions are shown for the location parameter μ at each site as well as for the dependence parameter α. Each fit is represented by the estimated mean as well as the central 95% probability interval.

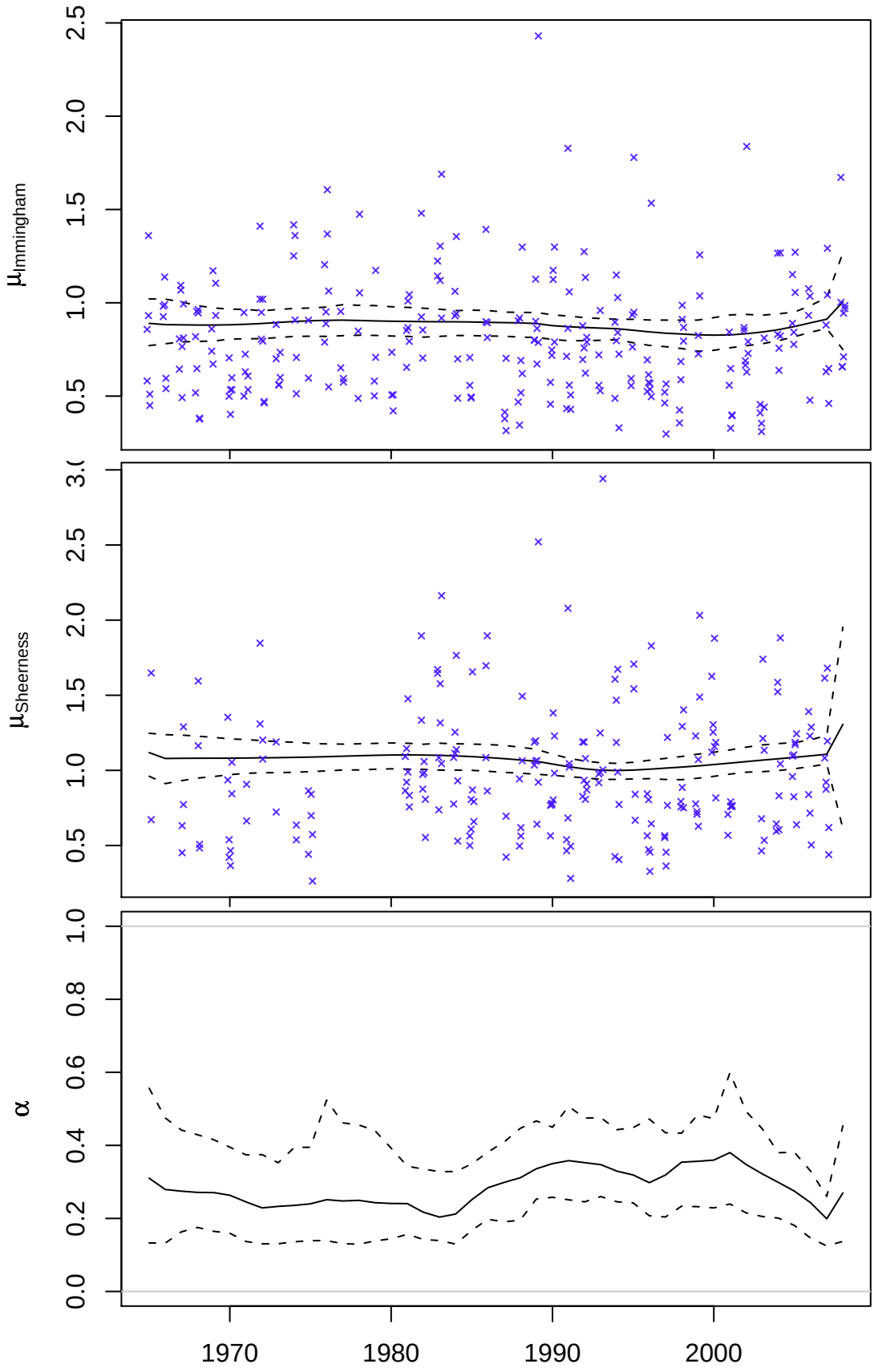


Figure 5.8: Smooth joint fit for Immingham and Sheerness as in Figure 5.7.

common to the whole eastern coastline.

While there is no overall trend in the dependence parameter for Immingham and Lowestoft, with Sheerness there is some evidence suggesting that the extremal dependence was less during the 1990s than in previous decades. However, the wide confidence intervals for α in both fits tells of the difficulty in estimating extremal dependence, which only more data can solve.

Since the fitted location parameters do not differ significantly from a constant, we could simplify our analysis by assuming each μ is fixed. This might allow us to better look at the variability of α as it could potentially be masked by the change in location. However, we expect the consequences of this change to be small since their fit is already flat.

We currently use data from November to February only to account for seasonality but if we can appropriately use every month's data we can expect efficiency gains with tighter probability intervals. This could be done with an additive sinusoidal term in the location parameter of each month's observations that adds to the annual stochastic trend, although this inevitably gives additional parameters to estimate. We may also wish to investigate whether it is the scale rather than location that should be seasonally adjusted.

Appendix A

Implementation of particle filters and smoothers

In this appendix we give the implementations of the particle filters and smoothers we use throughout this thesis.

A.1 Linear-Gaussian model

The linear-Gaussian model is defined in (2.4) of Subsection 2.1.2 for arbitrary design matrices.

To implement the various smoothing algorithms we need to choose propagation densities for a particle filter, backwards information filter and the smoother itself. Using auxiliary algorithms throughout, the linear-Gaussian model assumption allows us to calculate the optimal densities and re-sampling probabilities. Using these we have adapted algorithms giving even weights of $1/N$ whenever we resample.

Filter

Writing $\mathcal{N}(x|\mu,\Sigma)$ for the density of $\mathcal{N}(\mu,\Sigma)$ evaluated at x, it is easy to show that the optimal filter is given by

$$\begin{aligned} q(x_t|x_{t-1}^{(j)},y_t)\beta_t^{(j)} &= f(x_t|x_{t-1}^{(j)})g(y_t|x_t)w_{t-1}^{(j)} \\ &= \mathcal{N}(x_t|\mu_{t|t-1}^{(j)},\Sigma_{t|t-1})\ \mathcal{N}(y_t|GFx_{t-1}^{(j)},R+GQG')w_{t-1}^{(j)}, \end{aligned}$$

where $\Sigma_{t|t-1}=(Q^{-1}+G'R^{-1}G)^{-1}$ and $\mu_{t|t-1}^{(j)}=\Sigma_{t|t-1}(Q^{-1}Fx_{t-1}^{(j)}+G'R^{-1}y_t)$. This is used for each algorithm but we only need to keep track of our trajectories for the simple Filter-Smoother.

Backwards filter

For the backwards information filter we can use the actual prior $\gamma_t(x_t)=p(x_t)=\mathcal{N}(x_t|\mu_t,\Sigma_t)$, whose mean and covariance can be calculated sequentially using the prediction step of the Kalman filter. This gives

$$p(x_t|x_{t+1})=\mathcal{N}(x_t|\tilde{F}x_{t+1}+\tilde{Q}\Sigma_t^{-1}\mu_t,\tilde{Q}),$$

where we define $\tilde{F}:=\Sigma_tF'\Sigma_{t+1}^{-1}$ and $\tilde{Q}:=\Sigma_tF'\Sigma_{t+1}^{-1}QF'^{-1}$. We then obtain

$$\begin{aligned} \tilde{q}(x_t|y_t,\tilde{x}_{t+1}^{(k)})\tilde{\beta}_t^{(k)} &= p(x_t)g(y_t|x_t)f(\tilde{x}_{t+1}^{(k)}|x_t)\frac{\tilde{w}_{t+1}^{(k)}}{p(\tilde{x}_{t+1}^{(k)})} \\ &\propto \mathcal{N}(x_t|\mu_{t|t+1}^{(k)},\Sigma_{t|t+1})\cdot \\ &\qquad \mathcal{N}(y_t|G(\tilde{F}x_{t+1}^{(k)}+\tilde{Q}\Sigma_t^{-1}\mu_t),R+G\tilde{Q}G')\tilde{w}_{t+1}^{(k)}, \end{aligned}$$

where $\Sigma_{t|t+1}=(\Sigma_t^{-1}+G'R^{-1}G+F'Q^{-1}F)^{-1}$ and $\mu_{t|t+1}^{(k)}=\Sigma_{t|t+1}(\Sigma_t^{-1}\mu_t+G'R^{-1}y_t+F'Q^{-1}\tilde{x}_{t+1}^{(k)})$.

Smoother

Finally, for our new smoothing algorithm we have

$$\begin{aligned}\bar{q}(x_t|x_{t-1}^{(j)}, y_t, \tilde{x}_{t+1}^{(k)}) &\propto f(x_t|x_{t-1}^{(j)})g(y_t|x_t)f(\tilde{x}_{t+1}^{(k)}, x_t) \\ &\propto \mathcal{N}(x_t|\mu_{t|T}^{(j,k)}, \Sigma_{t|T}),\end{aligned}$$

where $\Sigma_{t|T} = (Q^{-1}+G'R^{-1}G+F'Q^{-1}F)^{-1}$ and $\mu_{t|T}^{(j,k)} = \Sigma_{t|T}(Q^{-1}Fx_{t-1}^{(j)}+G'R^{-1}y_t+F'Q^{-1}\tilde{x}_{t+1}^{(k)})$. The optimal re-sampling weights can be shown to be

$$\begin{aligned}\bar{\beta}_t^{(j,k)} &\propto p(\tilde{x}_{t+1}^{(k)}, y_t|x_{t-1}^{(j)})\frac{w_{t-1}^{(j)}\tilde{w}_{t+1}^{(k)}}{p(\tilde{x}_{t+1}^{(k)})} \\ &= \mathcal{N}\left(\begin{pmatrix}\tilde{x}_{t+1}^{(k)} \\ y_t\end{pmatrix}\middle|\begin{pmatrix}F^2 \\ GF\end{pmatrix}x_{t-1}^{(j)}, \begin{pmatrix}Q+FQF' & FQG' \\ GQF' & R+GQG'\end{pmatrix}\right)\frac{w_{t-1}^{(j)}\tilde{w}_{t+1}^{(k)}}{p(\tilde{x}_{t+1}^{(k)})},\end{aligned}$$

which we can see does not factorise. Therefore, for the $\mathcal{O}(N)$ version of our algorithm we use $\beta_t^{(j)}$ and $\tilde{\beta}_t^{(k)}$ from the filters as outlined in Algorithm 3.2 as this should be a good approximation of the optimal weights.

A.2 Stochastic volatility

The stochastic volatility model is given by (3.7) of Subsection 3.1.3.

The particle filter, backwards filter and block smoother for the SV model all involve target distributions of the form

$$\begin{aligned}q^{\text{opt}}(x_t) &\propto \mathcal{N}(x_t|\mu, \sigma^2)\, g(y_t|x_t) \\ &\propto \exp\left(-\frac{(x_t-\mu)^2}{2\sigma^2} - \frac{x_t}{2} - \frac{y_t^2}{2\beta^2 e^{x_t}}\right).\end{aligned}$$

Pitt and Shephard (1999a) showed how rejection sampling can be used to sample from $q^{\text{opt}}(x_t)$ exactly giving adapted auxiliary algorithms. However, to contrast

with the linear-Gaussian model above, we sample from an approximation of this instead.

We therefore consider a second order Taylor expansion of $\log q^{\text{opt}}(x_t)$ about an estimate of its mode. Setting the first derivative of this to zero and noting that the solution likely to be close to μ we get

$$\begin{aligned} x_t &= \mu + \frac{\sigma^2}{2}\left(\frac{y_t^2}{\beta^2 e^{x_t}} - 1\right) \\ &\simeq \mu + \frac{\sigma^2}{2}\left(\frac{y_t^2}{\beta^2 e^{\mu}} - 1\right) \end{aligned} \tag{A.1}$$

which we use as our estimate $\hat{x}_t$. This gives a Gaussian approximation with mean $\hat{x}_t$ and variance

$$\left(\frac{1}{\sigma^2} + \frac{y_t^2}{2\beta^2 e^{\hat{x}_t}}\right)^{-1} \tag{A.2}$$

which is close to the target $q^{\text{opt}}(x_t)$.

Filter

For the particle filter our target density satisfies

$$q^{\text{opt}}(x_t|x_{t-1}^{(j)}, y_t)\, \beta_t^{\text{opt}(j)} \propto f(x_t|x_{t-1}^{(j)})\, g(y_t|x_t)\, w_{t-1}^{(j)},$$

with $f(x_t|x_{t-1}^{(j)}) = \mathcal{N}(x_t|\phi x_{t-1}^{(j)}, \nu^2)$. Therefore, following the above, our proposal distribution $q(x_t|x_{t-1}^{(j)}, y_t)$ is Normal with mean and variance given by (A.1) and (A.2) respectively using $\mu = \phi x_{t-1}^{(j)}$ and $\sigma^2 = \nu^2$. We use the proposal mean $\hat{x}_t^{(j)}$

again to approximate the re-sampling weights by

$$\begin{aligned}\beta_t^{\mathrm{opt}(j)} &\propto \int g(y_t|x_t)\, f(x_t|x_{t-1}^{(j)})\, w_{t-1}^{(j)}\, \mathrm{d}x_t \\ &= \mathbf{E}_q\left(\frac{g(y_t|X_t)\, f(X_t|x_{t-1}^{(j)})\, w_{t-1}^{(j)}}{q(X_t|x_{t-1}^{(j)}, y_t)}\middle| x_{t-1}^{(j)}, y_t\right) \\ &\simeq \frac{g(y_t|\hat{x}_t^{(j)})\, f(\hat{x}_t^{(j)}|x_{t-1}^{(j)})\, w_{t-1}^{(j)}}{q(\hat{x}_t^{(j)}|x_{t-1}^{(j)}, y_t)}. \end{aligned} \tag{A.3}$$

Note that this is the same form as the final weight $w_t^{(i)}$ in Algorithm 2.3 without the $\beta_t^{(j)}$ term showing that the re-sampling weights may be thought of as a prediction of the final weights if we do not re-sample.

Backwards filter

For the backwards information filter we again use the actual prior $\gamma_t(x_t) = p(x_t) = \mathcal{N}(x_t|\mu_t, \sigma_t^2)$, whose mean and variance can be calculated sequentially with the Kalman filter. We then have target

$$\tilde{q}^{\mathrm{opt}}(x_t|y_t, \tilde{x}_{t+1}^{(k)})\, \tilde{\beta}_t^{(k)} \propto p(x_t|\tilde{x}_{t+1}^{(k)})\, g(y_t|x_t)\, \tilde{w}_{t+1}^{(k)},$$

where

$$p(x_t|x_{t+1}) = \mathcal{N}\left(x_t\middle|\frac{\phi\sigma_t^2}{\sigma_{t+1}^2}x_{t+1} + \frac{\nu^2}{\sigma_{t+1}^2}\mu_t, \frac{\nu^2\sigma_t^2}{\sigma_{t+1}^2}\right).$$

We therefore proceed as with the forwards filter substituting the form of $f(x_t|x_{t-1})$ for $p(x_t|x_{t+1})$ and $w_{t-1}^{(j)}$ for $\tilde{w}_{t+1}^{(k)}$.

Block smoother

For our $\mathcal{O}(N)$ smoother with inner block size n, our overall target is

$$\bar{q}^{\text{opt}}(x_{t:t+n-1}|x_{t-1}^{(j)}, y_{t:t+n-1}, \tilde{x}_{t+1}^{(k)}) \propto f(x_t|x_{t-1}^{(j)}) \left(\prod_{s=t+1}^{t+n-1} f(x_s|x_{s-1})\right) f(\tilde{x}_{t+n}^{(k)}|x_{t+n-1}) \cdot \left(\prod_{s=t}^{t+n-1} g(y_s|x_s)\right) \frac{w_{t-1}^{(j)}\, \tilde{w}_{t+n}^{(k)}}{p(\tilde{x}_{t+n}^{(k)})}.$$

As a density in $x_{t:t+n-1}$, this is

$$\begin{aligned}\bar{q}^{\text{opt}}(x_{t:t+n-1}|x_{t-1}^{(j)}, y_{t:t+n-1}, \tilde{x}_{t+1}^{(k)}) &\propto p(x_{t:t+n-1}|x_{t-1}^{(j)})\, f(\tilde{x}_{t+n}^{(k)}|x_{t+n-1}) \prod_{s=t}^{t+n-1} g(y_s|x_s) \\ &\propto p(x_{t:t+n-1}|x_{t-1}^{(j)}, \tilde{x}_{t+n}^{(k)}) \prod_{s=t}^{t+n-1} g(y_s|x_s),\end{aligned}$$

where the Brownian bridge $p(x_{t:t+n-1}|x_{t-1}, x_{t+n})$ is a multivariate Normal distribution with mean $\phi\, \Sigma(x_{t-1}, 0, \ldots, 0, x_{t+n})'/\nu^2$ and precision matrix

$$\Sigma^{-1} = \frac{1}{\nu^2} \begin{pmatrix} 1+\phi^2 & -\phi & 0 & 0 \\ -\phi & 1+\phi^2 & \ddots & 0 \\ 0 & \ddots & \ddots & -\phi \\ 0 & 0 & -\phi & 1+\phi^2 \end{pmatrix}.$$

To approximate the target we can sequentially incorporate $g(y_s|x_s)$ for $s = t, \ldots, t+n-1$ into the proposal using the Taylor approximation above. Beginning with $q(x_{t:t+n-1}) = p(x_{t:t+n-1}|x_{t-1}^{(j)}, \tilde{x}_{t+n}^{(k)})$, we can sequentially separate component s as $q(x_{t:t+n-1}) = q(x_s)\, q(x_{\setminus s}|x_s)$ and update $q(x_s)$ with the approximations (A.1) and (A.2). However, since the variance (A.2) depends on the mean $\hat{x}_s$ which varies over the particles, this strategy requires the calculation and possible storage of N covariance matrices of size $n \times n$ which could be considerable if the block size is large[1]. We therefore use the average value of the variance in (A.2) so that each

[1] N^2 matrices are required if the $\mathcal{O}(N^2)$ algorithm is used.

particle block is sampled with the same covariance matrix.

Finally, we follow Algorithm 3.3 and use the re-sampling weights $\beta_t^{(j)}$ and $\tilde{\beta}_{t+n-1}^{(k)}$ from the filters to re-sample the particles in our block smoother.

A.3 Bearings-only tracking

The bearings-only tracking model is given by (3.9) and (3.10) in Subsection 3.1.3.

To construct the proposal density for our auxiliary particle filter, we first note that there are only two degrees of freedom in the state so after sampling the position of a new particle the velocity is determined by $\dot{u}_t = u_t - u_{t-1}^{(i)}$ and $\dot{v}_t = v_t - v_{t-1}^{(i)}$. To sample the position we change to polar coordinates (r_t, α_t) and write the optimal proposal density as

$$q^{\text{opt}}(r_t, \alpha_t | x_{t-1}^{(i)}, y_t) \propto r_t \exp\left(-\frac{(r_t - \rho_t^{(i)} \cos(\alpha_t - \psi_t^{(i)}))^2}{2\nu^2}\right) \cdot \exp\left(-\frac{(y_t - \alpha_t)^2}{2\tau^2} - \frac{\rho_t^{(i)2} \sin^2(\alpha_t - \psi_t^{(i)})}{2\nu^2}\right),$$

where the prior means $(u_{t-1}^{(i)} + \dot{u}_{t-1}^{(i)}, v_{t-1}^{(i)} + \dot{v}_{t-1}^{(i)})$ are also changed to polar coordinates $(\rho_t^{(i)}, \psi_t^{(i)})$.

Fearnhead (1998) derive properties of the optimal range distribution conditional on α_t and show how to sample from it exactly using rejection sampling (details of which are omitted here). We use this method to sample the range but first need to sample the bearing α_t. The optimal bearing density is given by

$$q^{\text{opt}}(\alpha_t | x_{t-1}^{(i)}, y_t) \propto K_t^{(i)}(\alpha_t) \exp\left(-\frac{(y_t - \alpha_t)^2}{2\tau^2} - \frac{\rho_t^{(i)2} \sin^2(\alpha_t - \psi_t^{(i)})}{2\nu^2}\right),$$

where

$$K_t^{(i)}(\alpha_t) = \sqrt{2\pi\nu^2}(s_t^{(i)}\Phi(s_t^{(i)}) + \phi(s_t^{(i)})) \quad \text{with} \quad s_t^{(i)} = \frac{\rho_t^{(i)}\cos(\alpha_t - \psi_t^{(i)})}{\nu} \qquad \text{(A.4)}$$

is the normalising constant of the conditional range distribution (with Φ the cdf and ϕ the pdf of a $\mathcal{N}(0,1)$ random variable). To do this we assume $\alpha_t \simeq \psi_t^{(i)}$ to justify making the approximations $\cos(\alpha_t - \psi_t^{(i)}) \simeq 1 - (\alpha_t - \psi_t^{(i)})^2/2$, $\log\cos(\alpha_t - \psi_t^{(i)}) \simeq -(\alpha_t - \psi_t^{(i)})^2/2$ and $\Phi(s_t^{(i)}) \simeq 1 - \phi(s_t^{(i)})s_t^{(i)-1}$. Using these we get

$$q(\alpha_t|x_{t-1}^{(i)}, y_t) \propto \exp\left(-\frac{(\alpha_t - \psi_t^{(i)})^2}{2} - \frac{(y_t - \alpha_t)^2}{2\tau^2} - \frac{\rho_t^{(i)2}(\alpha_t - \psi_t^{(i)})^2}{2\nu^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{\nu^2 + (\nu^2 + \rho_t^{(i)2})\tau^2}{\nu^2\tau^2}\right)\left(\alpha_t - \frac{\nu^2 y_t + (\nu^2 + \rho_t^{(i)2})\tau^2\psi_t^{(i)}}{\nu^2 + (\nu^2 + \rho_t^{(i)2})\tau^2}\right)^2\right)$$

giving us a normal distribution for the bearing.

For the re-sampling weights $\beta_t^{(i)}$ we follow the suggestion (A.3) and use the predicted final weights about the mean of our proposal distribution for r_t and α_t.

To initialise the algorithm we use standard importance sampling for the first time step rather than sampling from the prior and propagating the particles forward as usual. Fearnhead (1998) showed that this can give a large improvement in efficiency when the prior is flat as the first observation carries a lot of information and causes only a few distinct prior particles to be propagated.

A.4 Analysis of women's 3000m running event

For the women's 3000m analysis of Section 4.1 we use the linear-Gaussian state model

$$X_{t+1}|\{X_{1:t} = x_{1:t}, Y_{1:t} = y_{1:t}\} \sim \mathcal{N}(Fx_t, Q),$$

for the two-dimensional state $X_t = (\mu_t, \dot{\mu}_t)'$ with F and Q given in (3.15). The negated observations $y_{t,1:r}$ are modelled, conditionally on μ_t, by the r-largest order statistic likelihood of (2.19).

Adapted auxiliary algorithms for this model will not be possible as the likelihood in μ is very complex. We therefore approximate the log likelihood $l(\mu_t)$ by a second-order Taylor approximation about an estimated mode $\hat{\mu}_t$ which leads to a normal approximation of the form

$$g(y_{t,1:r}|\mu_t) \simeq \mathcal{N}\left(\mu_t \;\middle|\; \hat{\mu}_t - \frac{l'(\hat{\mu}_t)}{l''(\hat{\mu}_t)}, -\frac{1}{l''(\hat{\mu}_t)}\right)\Bigg|_{A_t}, \tag{A.5}$$

where the distribution is restricted to the likelihood's support of

$$A_t := \{\mu_t | \sigma + \xi(y_{t,i} - \mu_t) > 0, \forall i\}.$$

In practice, we used the `optimize` function in R to estimate the mode at each time step.

To make the algorithms as efficient as possible we use Rao-Blackwellisation to reduce the variance of our estimates. For this we can marginalise the second component of the state $\dot{\mu}_t$ as the likelihood only depends on μ_t so the distribution of $\dot{\mu}_t|\mu_t$ can be updated by using only its mean and variance. This improves the overall approximation by allowing the second component of each particle to act as a normal distribution rather than a point mass. We therefore have particles of the form $x_t^{(i)} = (\mu_t^{(i)}, \dot{m}_t^{(i)}, \tau_t^{2(i)})'$, where $\dot{\mu}_t|\{\mu_t = \mu_t^{(i)}\} \sim \mathcal{N}(\dot{m}_t^{(i)}, \tau_t^{2(i)})$.

Filter

To create a marginalised particle filter it helps to think each particle $x_{t-1}^{(i)}$ as a kernel approximation to $p(\mu_{t-1}, \dot{\mu}_{t-1}|y_{1:t-1})$ of the form

$$\phi^{(i)}(\mu_{t-1}, \dot{\mu}_{t-1}) := \mathcal{N}(\mu_{t-1}, \dot{\mu}_{t-1}|\eta_{t-1}^{(i)}, K_{t-1}^{(i)}),$$

with

$$\eta_{t-1}^{(i)} := \begin{pmatrix} \mu_{t-1}^{(i)} \\ \dot{m}_{t-1}^{(i)} \end{pmatrix}, \qquad K_{t-1}^{(i)} := \begin{pmatrix} 0 & 0 \\ 0 & \tau_{t-1}^{2(i)} \end{pmatrix}.$$

This leads to the approximation of $p(\mu_t, \dot{\mu}_t|y_{1:t-1})$ by

$$\begin{aligned} \pi^{(i)}(\mu_t, \dot{\mu}_t) &:= \int f(\mu_t, \dot{\mu}_t|\mu_{t-1}, \dot{\mu}_{t-1})\, \phi^{(i)}(\mu_{t-1}, \dot{\mu}_{t-1})\, \mathrm{d}\mu_{t-1}\, \mathrm{d}\dot{\mu}_{t-1} \\ &= \mathcal{N}(\mu_t, \dot{\mu}_t|F\eta_{t-1}^{(i)}, Q + FK_{t-1}^{(i)}F'). \end{aligned}$$

To create the new particle $x_t^{(i)}$ we therefore use standard auxiliary particle filtering with target density

$$q^{\text{opt}}(\mu_t|x_{t-1}^{(i)}, y_t)\beta_t^{(i)} = \pi^{(i)}(\mu_t)\, g(y_t|\mu_t)\, w_{t-1}^{(i)}$$

to sample $\mu_t^{(i)}$ and then update the mean and variance of $\dot{\mu}_t|\{\mu = \mu_t^{(i)}\}$ with that of $\pi^{(i)}(\dot{\mu}_t|\mu_t^{(i)})$. For this we replace the likelihood by the approximation (A.5) to give us a constrained normal sampling density for $\mu_t^{(i)}$ and approximate the optimal re-sampling weights with

$$\beta_t^{(i)} \simeq \frac{\pi^{(i)}(\hat{\mu}_t)\, g(y_t|\hat{\mu}_t)\, w_{t-1}^{(i)}}{q(\hat{\mu}_t|x_{t-1}^{(i)}, y_t)},$$

where $\hat{\mu}_t$ is the mean of the sampling density $q(\mu_t|x_{t-1}^{(i)}, y_t)$.

Backwards filter

For the backwards filter we again start by defining $\tilde{F} := \Sigma_t F' \Sigma_{t+1}^{-1}$ and $\tilde{Q} := \Sigma_t F' \Sigma_{t+1}^{-1} Q F'^{-1}$, where Σ_t is the variance of the normal prior at time t. It can then be shown that $p(\mu_t, \dot{\mu}_t | \mu_{t+1}, \dot{\mu}_{t+1})$ is equal to

$$\mathcal{N}\left(\begin{pmatrix}\mu_t \\ \dot{\mu}_t\end{pmatrix} \middle| \tilde{F}\begin{pmatrix}\mu_{t+1} \\ \dot{\mu}_{t+1}\end{pmatrix} + \tilde{Q}\Sigma_t^{-1}\begin{pmatrix}\hat{\mu}_t \\ \hat{\dot{\mu}}_t\end{pmatrix}, \tilde{Q}\right),$$

where $(\hat{\mu}_t, \hat{\dot{\mu}}_t)'$ is the mean of the prior at time t. We then combine this with a kernel $\phi^{(i)}(\mu_{t+1}, \dot{\mu}_{t+1})$ created from $x_{t+1}^{(i)}$ to give the density

$$\tilde{\pi}^{(i)}(\mu_t, \dot{\mu}_t) := \mathcal{N}\left(\begin{pmatrix}\mu_t \\ \dot{\mu}_t\end{pmatrix} \middle| \tilde{F}\eta_{t+1}^{(i)} + \tilde{Q}\Sigma_t^{-1}\begin{pmatrix}\hat{\mu}_t \\ \hat{\dot{\mu}}_t\end{pmatrix}, \tilde{Q} + \tilde{F}K_{t+1}^{(i)}\tilde{F}'\right).$$

We now proceed in exactly the same way as with the forwards filter using $\tilde{\pi}$ instead of π to sample $x_t^{(i)}$.

Smoother

Finally, for our new smoothing algorithm, it can be shown that our target for $\mu_t^{(i)}$ in this marginalised setting is

$$\bar{q}^{\text{opt}}(\mu_t | x_{t-1}^{(j)}, y_t, x_{t+1}^{(k)})\, \bar{\beta}_t^{(j,k)} = \bar{\pi}^{(j,k)}(\mu_t)\, g(y_t|\mu_t)\, w_{t-1}^{(j)}\, \tilde{w}_{t+1}^{(k)},$$

where

$$\bar{\pi}^{(j,k)}(\mu_t, \dot{\mu}_t) :\propto \frac{\pi^{(j)}(\mu_t, \dot{\mu}_t)\, \tilde{\pi}^{(k)}(\mu_t, \dot{\mu}_t)}{p(\mu_t, \dot{\mu}_t)}.$$

This leads us to sample $\mu_t^{(i)}$ as before using $\bar{\pi}^{(j,k)}(\mu_t)$ in place of $\pi^{(j)}(\mu_t)$. We can then calculate the mean and variance of $\dot{\mu}_t | \{\mu = \mu_t^{(i)}\}$ from $\bar{\pi}^{(j,k)}(\dot{\mu}_t | \mu_t^{(i)})$. The filter and backwards filter re-sampling weights were used again for the suboptimal $\mathcal{O}(N)$ version of our algorithm.

For both the filter and the backwards filter the initial step was sampled using standard importance sampling as the target density is available in closed form and using it rather than propagating the prior greatly improves the algorithm. We also used the stratified sampling algorithm of Carpenter et al. (1999) in both the filters and our new algorithm to reduce the Monte Carlo error of re-sampling. Following the results of Section 3.1, we compare the effective sample size in (2.7) with a fixed threshold that is close to N to decide when to re-sample in the filters since this should give near optimal results.

Since we choose not to include the data from 1993, for this time step in each of the above algorithms we proceed without the likelihood term $g(y_t|\mu_t)$.

A.5 Analysis of Antarctic temperature data

The Antarctic temperature model of Section 4.2 is defined for arbitrary block sizes characterised by Δt. The state transition model for $X_t = (\mu_t, \dot{\mu}_t)'$ is given by (4.4) which depends upon matrices $F_{\Delta t}$ and $Q_{\Delta t}$ given in (4.5). In this context, Δt is the time difference (in years) between the centre of two adjacent blocks.

The point process observation density $g(\mathbf{y}_t|\mu_t)$ is given by (4.3) where here Δt denotes the size of the block in years. The observations $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,n_t})$ are the cluster maxima of the exceedances of the threshold u; it is therefore assumed that the series is first negated if the smallest values are of interest.

This model is very similar to that of the women's 3000m analysis of Section 4.1. Indeed, the athletics model is obtained within the temperature model by setting $\Delta t = 1$ and adjusting the threshold to the r-th largest value in each year. We therefore only need to adjust the particle methods' implementation for the women's 3000m analysis given in Appendix A.4.

This is simply done by replacing F and Q by $F_{\Delta t}$ and $Q_{\Delta t}$ and using the point

process observation density in place of the r-largest one. The derivations in Appendix A.4 then give us the filter, backwards filter and smoother for our Antarctic temperature model.

Filter and backwards filter initialisation

Because the number of observations in each block varies it is possible for the first few blocks at either end of the series to contain no observations. This causes the filter or backwards filter distribution to be diffuse since the prior is uninformative and this can cause the filter to collapse to a single point when the first observation is encountered.

To counter this problem we initialise the particle filter by sampling the first k time steps jointly as $p(x_{1:k}|y_{1:k})$. We select k to be just large enough for two separate time steps within $1, \ldots, k$ to contain observations, thus providing information about the velocity $\dot{\mu}_t$ as well as μ_t.

The joint distribution $p(x_{1:k}|y_{1:k})$ is given by

$$\begin{aligned} p(x_{1:k}|y_{1:k}) &\propto p(x_{1:k})\, p(y_{1:k}|x_{1:k}) \\ &= \left(p(x_1) \prod_{t=2}^{k} f(x_t|x_{t-1}) \right) \prod_{t=1}^{k} g(y_t|x_t), \end{aligned}$$

which can be sampled from directly. However, since the observation density does not depend on $\dot{\mu}_t$, we marginalised this component from the filter and should do the same here. To do this we sample instead $p(\mu_{1:k}|y_{1:k})$ and calculate the mean and covariance of $p(\dot{\mu}_{1:k}|\mu_{1:k}) = p(\dot{\mu}_{1:k}|\mu_{1:k}, y_{1:k})$ which is possible since $p(x_{1:k})$ is available in closed form.

Since k is not very large we may be able to use importance sampling to sample $p(\mu_{1:k}|y_{1:k})$ but MCMC can, alternatively, be used to make it more robust to the range of parameter values given by the EM algorithm. We create a sim-

ple Metropolis algorithm using a multivariate Gaussian distribution to iteratively sample a block $\mu^{\bigstar}_{1:k}$ and accept or reject it based on a likelihood ratio.

Specifically, we sample $\mu^{\bigstar}_{1:k} \sim \mathcal{N}(\mu^{(i-1)}_{1:k}, \Sigma)$ and accept the block by setting $\mu^{(i)}_{1:k} = \mu^{\bigstar}_{1:k}$ with probability

$$\min\left\{1, \frac{p(\mu^{\bigstar}_{1:k}|y_{1:k})}{p(\mu^{(i-1)}_{1:k}|y_{1:k})}\right\},$$

setting $\mu^{(i)}_{1:k} = \mu^{(i-1)}_{1:k}$ otherwise. For the $k \times k$ covariance matrix Σ we choose

$$\tau^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \rho \\ \vdots & \ddots & \ddots & \vdots \\ \rho & \cdots & \rho & 1 \end{pmatrix},$$

where τ and ρ are tuning constants. We sample the block as one from a correlated distribution since the μ_t components are highly correlated, especially when ν is small.

We use the same procedure sample $p(\mu_{T-k+1:T}|y_{T-k+1:T})$ to initialise the backwards filter. To reduce the correlation between sampled blocks we thin the sample by keeping only the mth value and to allow the Markov chain to converge we only start keeping values after a suitable burn-in period.

A.6 Pooling athletics data from two events

For the joint women's 1500m and 3000m analysis of Section 5.1 we use the familiar state model (5.2) with the 4-dimensional state $X_t = (\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})'$. This depends upon the matrices $F_{\Delta t}$ and $Q_{\Delta t}$ given by (5.3) where we set $\Delta t = 1$ to give time steps of a year.

The annual observations $\mathbf{y}_t = (y_{1500,1}, \ldots, y_{1500,p}, y_{3000,1}, \ldots, y_{3000,q})_t$ are the negated p and q-fastest 1500m and 3000m times in year t. The observations we include

varies between years since we choose to remove certain values from the fit. The observation density $g(\mathbf{y}_t|\mu_{1500,t}, \mu_{3000,t})$ is given in Appendix C.3 for p and q up to 2.

To construct our particle methods, we once again mirror the 3000m analysis derivation of Appendix A.4. The only real difference between this and our current model is that the state has doubled in size and now the observation density depends upon two of the four components. This allows us to marginalise the remaining two components for improved accuracy.

For all three particle algorithms we approximate the log likelihood $l(\mu_{1500,t}, \mu_{3000,t}) := \log(g(\mathbf{y}_t|\mu_{1500,t}, \mu_{3000,t}))$ by a second-order Taylor approximation about a mode estimate $(\hat{\mu}_{1500,t}, \hat{\mu}_{3000,t})$ obtained numerically. This gives the following approximation to the observation density:

$$\mathcal{N}\left(\begin{pmatrix}\mu_{1500,t}\\ \mu_{3000,t}\end{pmatrix}\middle|\begin{pmatrix}\hat{\mu}_{1500,t}\\ \hat{\mu}_{3000,t}\end{pmatrix} - \nabla^2 l\begin{pmatrix}\hat{\mu}_{1500,t}\\ \hat{\mu}_{3000,t}\end{pmatrix}^{-1}\nabla l\begin{pmatrix}\hat{\mu}_{1500,t}\\ \hat{\mu}_{3000,t}\end{pmatrix}, -\nabla^2 l\begin{pmatrix}\hat{\mu}_{1500,t}\\ \hat{\mu}_{3000,t}\end{pmatrix}^{-1}\right)\Bigg|_{A_t},$$

where the distribution is restricted to the likelihood's support of

$$A_t := \Big\{(\mu_{1500,t}, \mu_{3000,t})'\Big|\sigma_{1500} + \xi_{1500}(y_{1500,i,t} - \mu_{1500,t}) > 0, \\ \sigma_{3000} + \xi_{3000}(y_{3000,j,t} - \mu_{3000,t}) > 0, \forall i, j\Big\}.$$

To marginalise both $\dot{\mu}_{1500,t}$ and $\dot{\mu}_{3000,t}$ we need to update the distribution of $(\dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})|(\mu_{1500,t},\ \mu_{3000,t})$ algebraically. In doing so we work with particles of the form $x_t^{(i)} = (\mu_{1500,t}^{(i)}, \mu_{3000,t}^{(i)}, \dot{m}_{1500,t}^{(i)}, \dot{m}_{3000,t}^{(i)}, \tau_{1500,t}^{2(i)}, \tau_{3000,t}^{2(i)}, c_t^{(i)})'$, where

$$\begin{pmatrix}\dot{\mu}_{1500,t}\\ \dot{\mu}_{3000,t}\end{pmatrix}\Bigg|\begin{pmatrix}\mu_{1500,t}^{(i)}\\ \mu_{3000,t}^{(i)}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\dot{m}_{1500,t}^{(i)}\\ \dot{m}_{3000,t}^{(i)}\end{pmatrix}, \begin{pmatrix}\tau_{1500,t}^{2(i)} & c_t^{(i)}\\ c_t^{(i)} & \tau_{3000,t}^{2(i)}\end{pmatrix}\right).$$

These are then used to construct kernels of the form

$$\phi^{(i)}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t}) := \mathcal{N}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t} | \eta_t^{(i)}, K_t^{(i)}),$$

where

$$\eta_t^{(i)} := \begin{pmatrix} \mu_{1500,t}^{(i)} \\ \mu_{3000,t}^{(i)} \\ \dot{m}_{1500,t}^{(i)} \\ \dot{m}_{3000,t}^{(i)} \end{pmatrix}, \qquad K_t^{(i)} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \tau_{1500,t}^{2(i)} & c_t^{(i)} \\ 0 & 0 & c_t^{(i)} & \tau_{3000,t}^{2(i)} \end{pmatrix}.$$

Filter

To increment the marginalised particle filter, we begin by constructing the kernel $\phi^{(i)}(\mu_{1500,t-1}, \mu_{3000,t-1}, \dot{\mu}_{1500,t-1}, \dot{\mu}_{3000,t-1})$ from each filter particle $x_{t-1}^{(i)}$ as above. We use this filter approximation to approximate the prediction density $p(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t} | \mathbf{y}_{1:t-1})$ by

$$\pi^{(i)}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t}) = \mathcal{N}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t} | F_1 \eta_{t-1}^{(i)}, Q_1 + F_1 K_{t-1}^{(i)} F_1').$$

New particles $x_t^{(i)}$ are sampled using the proposal density

$$q(\mu_{1500,t}, \mu_{3000,t} | x_{t-1}^{(i)}, \mathbf{y}_t) \propto \pi^{(i)}(\mu_{1500,t}, \mu_{3000,t})\, \hat{g}(\mathbf{y}_t | \mu_{1500,t}, \mu_{3000,t}),$$

which is a bivariate Gaussian distribution constrained to be in A_t. The marginalised components are then taken from the mean and covariance of $\pi^{(i)}(\dot{\mu}_{1500,t}, \dot{\mu}_{3000,t} | \mu_{1500,t}^{(i)}, \mu_{3000,t}^{(i)})$.

As before we use an auxiliary algorithm with re-sampling weights of the form

$$\beta_t^{(i)} \simeq \frac{\pi^{(i)}(\hat{\mu}_{1500,t}, \hat{\mu}_{3000,t})\, g(\mathbf{y}_t | \hat{\mu}_{1500,t}, \hat{\mu}_{3000,t})\, w_{t-1}^{(i)}}{q(\hat{\mu}_{1500,t}, \hat{\mu}_{3000,t} | x_{t-1}^{(i)}, \mathbf{y}_t)},$$

where $(\hat{\mu}_{1500,t}, \hat{\mu}_{3000,t})$ is the mean of $q(\mu_{1500,t}, \mu_{3000,t}|x_{t-1}^{(i)}, \mathbf{y}_t)$. We initialise the algorithm by using standard importance sampling for the first time step to overcome the loss in efficiency caused by propagating from a diffuse prior.

Backwards filter

For the backwards filter we once again begin by defining $\tilde{F}_1 := \Sigma_t F_1' \Sigma_{t+1}^{-1}$ and $\tilde{Q}_1 := \Sigma_t F_1' \Sigma_{t+1}^{-1} Q_1 F_1'^{-1}$, where Σ_t is the prior variance at time t. We then use a kernel constructed from the backwards filter particle $\tilde{x}_{t+1}^{(i)}$ to create

$$\tilde{\pi}^{(i)} \begin{pmatrix} \mu_{1500,t} \\ \mu_{3000,t} \\ \dot{\mu}_{1500,t} \\ \dot{\mu}_{3000,t} \end{pmatrix} = \mathcal{N} \left(\begin{pmatrix} \mu_{1500,t} \\ \mu_{3000,t} \\ \dot{\mu}_{1500,t} \\ \dot{\mu}_{3000,t} \end{pmatrix} \middle| \tilde{F}_1 \eta_{t+1}^{(i)} + \tilde{Q}_1 \Sigma_t^{-1} \begin{pmatrix} \hat{\mu}_{1500,t} \\ \hat{\mu}_{3000,t} \\ \hat{\dot{\mu}}_{1500,t} \\ \hat{\dot{\mu}}_{3000,t} \end{pmatrix}, \tilde{Q}_1 + \tilde{F}_1 K_{t+1}^{(i)} \tilde{F}_1' \right),$$

where $(\hat{\mu}_{1500,t}, \hat{\mu}_{3000,t}, \hat{\dot{\mu}}_{1500,t}, \hat{\dot{\mu}}_{3000,t})'$ is the prior mean for time t. We then proceed as the forwards filter, using $\tilde{\pi}$ in place of π.

Smoother

For the smoother we combine the above filter and backwards filter to define $\bar{\pi}^{(j,k)}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})$ as proportional to

$$\frac{\pi^{(j)}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})\, \tilde{\pi}^{(k)}(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})}{p(\mu_{1500,t}, \mu_{3000,t}, \dot{\mu}_{1500,t}, \dot{\mu}_{3000,t})}.$$

We then follow a similar strategy as before, using the filters re-sampling weights to create the $\mathcal{O}(N)$ version of our algorithm.

A.7 Joint analysis of sea-level data

For the sea-level analysis of Section 5.2 we jointly model the upper extreme sea-levels at a pair of cites, hereby labelled A and B. We use the familiar state model (5.4) with the 5-dimensional state $X_t = (\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \dot{\mu}_{\mathrm{A},t}, \dot{\mu}_{\mathrm{B},t}, \alpha_t^*)'$ where $\alpha_t^* := \Phi^{-1}(\alpha_t)$. The state design matrices $F_{\Delta t}$ and $Q_{\Delta t}$ are given in (5.5) where we take $\Delta t = 1$ so that the state parameters vary between years.

We have four sets of independent observations per year arising from each of the winter months November to February. The observations within a month are the p-largest cluster maxima from site A and the q-largest from site B. Each month's observations are modelled independently by the bivariate logistic model derived in Appendix C so that the observation density for year t, $g(\mathbf{y}_t|\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \alpha_t)$, is the product of four logistic densities. The datasets have much missing data which causes either p or q to be 0 or the logistic density to be removed entirely.

The implementation of our particle filter and backwards filter is based on that of Appendix A.6 for the simpler model with α as a constant. The addition of α_t^* to the state adds an extra component to the sampling distributions but otherwise the methods remain the same. For example, the log likelihood approximation, used by each of the particle methods, now becomes

$$\mathcal{N}\left(\begin{pmatrix}\mu_{\mathrm{A},t}\\ \mu_{\mathrm{B},t}\\ \alpha_t^*\end{pmatrix}\middle|\begin{pmatrix}\hat{\mu}_{\mathrm{A},t}\\ \hat{\mu}_{\mathrm{B},t}\\ \hat{\alpha}_t^*\end{pmatrix} - \nabla^2 l\begin{pmatrix}\hat{\mu}_{\mathrm{A},t}\\ \hat{\mu}_{\mathrm{B},t}\\ \hat{\alpha}_t^*\end{pmatrix}^{-1} \nabla l\begin{pmatrix}\hat{\mu}_{\mathrm{A},t}\\ \hat{\mu}_{\mathrm{B},t}\\ \hat{\alpha}_t^*\end{pmatrix}, -\nabla^2 l\begin{pmatrix}\hat{\mu}_{\mathrm{A},t}\\ \hat{\mu}_{\mathrm{B},t}\\ \hat{\alpha}_t^*\end{pmatrix}^{-1}\right)\Bigg|_{A_t}, \quad \text{(A.6)}$$

where $(\hat{\mu}_{\mathrm{A},t}, \hat{\mu}_{\mathrm{B},t}, \hat{\alpha}_t^*)$ is a mode estimate of the log likelihood $l(\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \alpha_t^*) := \log(g(\mathbf{y}_t|\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \alpha_t^*))$ and

$$A_t := \Big\{(\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \alpha_t^*)'\Big|\sigma_{\mathrm{A}} + \xi_{\mathrm{A}}(y_{\mathrm{A},i,t} - \mu_{\mathrm{A},t}) > 0, \\ \sigma_{\mathrm{A}} + \xi_{\mathrm{B}}(y_{\mathrm{B},j,t} - \mu_{\mathrm{B},t}) > 0, \forall i, j\Big\}$$

is the likelihood's support.

As before we use Rao-Blackwellisation to marginalise the two $\dot{\mu}_t$ components to improve the overall efficiency of the algorithms. This is done by updating the distribution of $(\dot{\mu}_{\mathrm{A},t}, \dot{\mu}_{\mathrm{B},t})|(\mu_{\mathrm{A},t},\ \mu_{\mathrm{B},t}, \alpha_t^*)$ algebraically while using particles of the form $x_t^{(i)} = (\mu_{\mathrm{A},t}^{(i)}, \mu_{\mathrm{B},t}^{(i)}, \dot{m}_{\mathrm{A},t}^{(i)}, \dot{m}_{\mathrm{B},t}^{(i)}, \tau_{\mathrm{A},t}^{2(i)}, \tau_{\mathrm{B},t}^{2(i)}, c_t^{(i)}, \alpha_t^{*(i)})'$.

To gauge the accuracy of the filter and backwards filter as they run, we monitor the effective sample size using (2.7) for the re-sampling weights $\beta_t^{(j)}$ (or $\tilde{\beta}_t^{(k)}$) and for the final weights $w_t^{(i)}$ (or $\tilde{w}_t^{(i)}$). If $\mathrm{ESS}(\beta_t)$ or $\mathrm{ESS}(w_t)$ is very low there may be too few particles with significant weights to accurately approximate the filter distribution so that subsequent steps diverge substantially from the target.

In an attempt to improve the accuracy of each filter, we repeat step t if $\mathrm{ESS}(w_t)$ is lower than a predetermined threshold and repeat the previous step if $\mathrm{ESS}(\beta_t)$ is too low. With even very low thresholds we found this strategy can prevent the filters diverging to effectively a single point. Care must be taken, however, to ensure the thresholds are not so high that after repeated attempts the filter gets stuck.

Filter and backwards filter initialisation

As with the temperature analysis implementation in Appendix A.5, we use MCMC to improve the initialisation of the filters by sampling the first k time steps jointly. Since the velocity components are marginalised, we use MCMC to sample from $p(\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha_{1:k}^*|\mathbf{y}_{1:k})$ for the filter, and similar for the backwards filter.

Extending the Metropolis algorithm used for the temperature analysis, we sequentially propose a block $(\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha_{1:k}^*)^\bigstar$ and accept it or reject it with a probability equal to a likelihood ratio. Since the μ and α^* components are independent in the state density, we alternate between updating each component block.

During a μ update, we sample the proposal μ block

$$(\mu_{\mathrm{A},1}, \mu_{\mathrm{B},1}, \ldots, \mu_{\mathrm{A},k}, \mu_{\mathrm{B},k})^{\bigstar} \sim \mathcal{N}((\mu_{\mathrm{A},1}, \mu_{\mathrm{B},1}, \ldots, \mu_{\mathrm{A},k}, \mu_{\mathrm{B},k})^{(i-1)}, \Sigma_\mu),$$

and accept $(\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i)} = (\mu^{\bigstar}_{\mathrm{A},1:k}, \mu^{\bigstar}_{\mathrm{B},1:k}, \alpha^{*(i-1)}_{1:k})$ with probability

$$\min\left\{1, \frac{p(\mu^{\bigstar}_{\mathrm{A},1:k}, \mu^{\bigstar}_{\mathrm{B},1:k}, \alpha^{*(i-1)}_{1:k} | \mathbf{y}_{1:k})}{p(\mu^{(i-1)}_{\mathrm{A},1:k}, \mu^{(i-1)}_{\mathrm{B},1:k}, \alpha^{*(i-1)}_{1:k} | \mathbf{y}_{1:k})}\right\},$$

setting $(\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i)} = (\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i-1)}$ otherwise. For the $2k \times 2k$ covariance matrix Σ_μ we select

$$\begin{pmatrix} S_\mu & \rho S_\mu & \cdots & \rho S_\mu \\ \rho S_\mu & S_\mu & \ddots & \rho S_\mu \\ \vdots & \ddots & \ddots & \vdots \\ \rho S_\mu & \cdots & \rho S_\mu & S_\mu \end{pmatrix} \qquad \text{with} \qquad S_\mu = \begin{pmatrix} \tau^2_{\mathrm{A}} & \tau_{\mathrm{A}}\tau_{\mathrm{B}}\rho_\mu \\ \tau_{\mathrm{A}}\tau_{\mathrm{B}}\rho_\mu & \tau^2_{\mathrm{B}} \end{pmatrix},$$

where τ_{A}, τ_{B} and ρ are tuning constants. This gives $\mu_{\mathrm{A},t}$ and $\mu_{\mathrm{B},t}$ the same correlation ρ_μ that they have in the state but also allows us to adjust the correlation between the time steps with ρ.

During an α^* update we similarly sample $\alpha^{*\bigstar}_{1:k} \sim \mathcal{N}(\alpha^{*(i-1)}_{1:k}, \Sigma_\alpha)$ and accept $(\mu_{\mathrm{A},1:k}$, $\mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i)} = (\mu^{(i-1)}_{\mathrm{A},1:k}, \mu^{(i-1)}_{\mathrm{B},1:k}, \alpha^{*\bigstar}_{1:k})$ with probability

$$\min\left\{1, \frac{p(\mu^{(i-1)}_{\mathrm{A},1:k}, \mu^{(i-1)}_{\mathrm{B},1:k}, \alpha^{*\bigstar}_{1:k} | \mathbf{y}_{1:k})}{p(\mu^{(i-1)}_{\mathrm{A},1:k}, \mu^{(i-1)}_{\mathrm{B},1:k}, \alpha^{*(i-1)}_{1:k} | \mathbf{y}_{1:k})}\right\},$$

setting $(\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i)} = (\mu_{\mathrm{A},1:k}, \mu_{\mathrm{B},1:k}, \alpha^*_{1:k})^{(i-1)}$ otherwise. For the $k \times k$

covariance matrix Σ_α we select

$$\tau_\alpha^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \rho \\ \vdots & \ddots & \ddots & \vdots \\ \rho & \cdots & \rho & 1 \end{pmatrix}$$

which uses an additional tuning constant τ_α.

The same procedure is used to sample $p(\mu_{\mathrm{A},T-k+1:T}, \mu_{\mathrm{B},T-k+1:T}, \alpha^*_{T-k+1:T} | \, \mathbf{y}_{T-k+1:T})$ to initialise the backwards filter. We again use a burn-in period and thinning to improve the MCMC samples.

Block smoother

We use the block version of our $\mathcal{O}(N)$ smoothing algorithm given in Algorithm 3.3 since this improves the performance when the dependence in the state is high. Recall that in step t the block $x_{t:t+n-1}$ of size n is sampled given filter particles $x_{t-1}^{(j)}$ and backwards filter particles $\tilde{x}_{t+n}^{(k)}$. Our target density is given by

$$\begin{aligned} \bar{q}^{\mathrm{opt}}(x_{t:t+n-1} | x_{t-1}^{(j)}, \mathbf{y}_{t:t+n-1}, \tilde{x}_{t+1}^{(k)}) &\propto f(x_t | x_{t-1}^{(j)}) \prod_{s=t+1}^{t+n-1} f(x_s | x_{s-1}) \cdot \\ &\qquad f(\tilde{x}_{t+n}^{(k)} | x_{t+n-1}) \prod_{s=t}^{t+n-1} g(\mathbf{y}_s | \mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*) \\ &\propto p(x_{t:t+n-1} | x_{t-1}^{(j)}) \, f(\tilde{x}_{t+n}^{(k)} | x_{t+n-1}) \cdot \\ &\qquad \prod_{s=t}^{t+n-1} g(\mathbf{y}_s | \mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*) \\ &\propto p(x_{t:t+n-1} | x_{t-1}^{(j)}, \tilde{x}_{t+n}^{(k)}) \prod_{s=t}^{t+n-1} g(\mathbf{y}_s | \mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*). \end{aligned}$$

The Brownian bridge $p(x_{t:t+n-1}|x_{t-1}, x_{t+n})$ is a multivariate Normal distribution with mean $\Sigma(Q_{\Delta t}^{-1}F_{\Delta t}x_{t-1}, 0, \ldots, 0, F_{\Delta t}'Q_{\Delta t}^{-1}x_{t+n})'$ and precision matrix

$$\Sigma^{-1} = \begin{pmatrix} Q_{\Delta t}^{-1} + F_{\Delta t}'Q_{\Delta t}^{-1}F_{\Delta t} & -F_{\Delta t}'Q_{\Delta t}^{-1} & 0 & 0 \\ -Q_{\Delta t}^{-1}F_{\Delta t} & Q_{\Delta t}^{-1} + F_{\Delta t}'Q_{\Delta t}^{-1}F_{\Delta t} & \ddots & 0 \\ 0 & \ddots & \ddots & -F_{\Delta t}'Q_{\Delta t}^{-1} \\ 0 & 0 & -Q_{\Delta t}^{-1}F_{\Delta t} & Q_{\Delta t}^{-1} + F_{\Delta t}'Q_{\Delta t}^{-1}F_{\Delta t} \end{pmatrix}.$$

Since the observation density does not depend upon $\dot{\mu}_{\mathrm{A}}$ or $\dot{\mu}_{\mathrm{B}}$ we update $(\dot{\mu}_{\mathrm{A},t}, \dot{\mu}_{\mathrm{B},t})|(\mu_{\mathrm{A},t}, \mu_{\mathrm{B},t}, \alpha_t^*)$ algebraically using the Brownian bridge storing only the marginal in time components in each particle as $(\dot{m}_{\mathrm{A}}^{(i)}, \dot{m}_{\mathrm{B}}^{(i)}, \tau_{\mathrm{A}}^{2(i)}, \tau_{\mathrm{B}}^{2(i)}, c^{(i)})_{t:t+n-1}$. We first sample $(\mu_{\mathrm{A}}^{(i)}, \mu_{\mathrm{B}}^{(i)}, \alpha^{*(i)})_{t:t+n-1}$ from an approximation to

$$\bar{q}^{\mathrm{opt}}((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{t:t+n-1}|x_{t-1}^{(j)}, \mathbf{y}_{t:t+n-1}, \tilde{x}_{t+1}^{(k)}) \propto p((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{t:t+n-1}|x_{t-1}^{(j)}, \tilde{x}_{t+n}^{(k)}) \prod_{s=t}^{t+n-1} g(\mathbf{y}_s|\mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*).$$

To construct our approximation we sequentially incorporate $g(\mathbf{y}_s|\mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*)$ for $s = t, \ldots, t+n-1$ into the proposal using the approximation (A.6). Starting with $q((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{t:t+n-1}) = p((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{t:t+n-1}|x_{t-1}^{(j)}, \tilde{x}_{t+n}^{(k)})$, we sequentially separate the triple indexed at time s using $q((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{t:t+n-1}) = q((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_s) \cdot q((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_{\setminus s}|(\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_s)$ and multiply $q((\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \alpha^*)_s)$ by the approximation $\hat{g}(\mathbf{y}_s|\mu_{\mathrm{A},s}, \mu_{\mathrm{B},s}, \alpha_s^*)$.

Putting this all together gives us a $3n$ dimensional multivariate Normal with up to $2n$ restricted components. Although we can sample from this distribution, its density requires the calculation of an up to $2n$ dimensional tail probability which is hard to approximate. We therefore only restrict up to 2 of the $2n$ possible components, preferring one for μ_{A} and one for μ_{B} as then the strong dependence in the state brings the other components close to their boundaries.

As before, we use the re-sampling weights $\beta_t^{(j)}$ and $\tilde{\beta}_{t+n-1}^{(k)}$ from the filters to re-sample the corresponding particles in our block smoother.

Appendix B

Derivation of state models from stochastic differential equations

In this appendix we derive a couple of state models that we use throughout the thesis from a pair of stochastic differential equations (SDEs).

B.1 Integrated random walk

We first consider the two-dimensional stochastic differential equation (3.16) as used for the linear-Gaussian simulation study in Subsection 3.2.3 and also both analyses of Chapter 4. It can be written in vector form as

$$\mathrm{d}\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ \nu \end{pmatrix} \mathrm{d}B_t, \tag{B.1}$$

where B_t is a Wiener process. The first component $X_{t,1}$ is therefore the integrated path of the scaled Wiener process $X_{t,2}$.

This model is a linear SDE of the form

$$\mathrm{d}X_t = \big(A(t)X_t + \mathbf{b}(t)\big)\,\mathrm{d}t + \sum_{i=1}^{m}\big(C_i(t)X_t + \mathbf{d}_i(t)\big)\,\mathrm{d}B_{t,i}, \tag{B.2}$$

where $\{B_{t,i}\}_{i=1}^{m}$ are independent Wiener processes, and with $m = 1$, $\mathbf{b}(t) \equiv \mathbf{0}$, $C_1(t) \equiv 0$,

$$A(t) \equiv \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{d}_1(t) \equiv \begin{pmatrix} 0 \\ \nu \end{pmatrix}.$$

The general solution to (B.2) is given by

$$X_t = \Phi_{t,t_0}\Bigg(X_{t_0} + \int_{t_0}^{t}\Phi_{s,t_0}^{-1}\Bigg(\mathbf{b}(s) - \sum_{i=1}^{m}C_i(s)\mathbf{d}_i(s)\Bigg)\,\mathrm{d}s + \sum_{i=1}^{m}\int_{t_0}^{t}\Phi_{s,t_0}^{-1}\mathbf{d}_i(s)\,\mathrm{d}B_{s,i}\Bigg), \tag{B.3}$$

where Φ_{t,t_0} is the *fundamental matrix* satisfying $\Phi_{t_0,t_0} = I$ and the *homogeneous SDE*:

$$\mathrm{d}\Phi_{t,t_0} = A(t)\Phi_{t,t_0}\,\mathrm{d}t + \sum_{i=1}^{m}C_i(t)\Phi_{t,t_0}\,\mathrm{d}B_{t,i}. \tag{B.4}$$

A solution to (B.4) does not always exist, but if A and C_i are constants and commute (i.e. $AC_i = C_iA$ and $C_iC_j = C_jC_i$) then the fundamental matrix is given by the matrix exponential

$$\Phi_{t,t_0} = \exp\left(\left(A - \frac{1}{2}\sum_{i=1}^{m}C_i^2\right)(t-t_0) + \sum_{i=1}^{m}C_i(B_{t,i} - B_{t_0,i})\right). \tag{B.5}$$

For our two-dimensional SDE (B.1), A and C_1 are commuting constants and so, using (B.5), we have

$$\begin{aligned}\Phi_{t,t_0} &= \exp\big(A(t-t_0)\big)\\ &= I + A(t-t_0),\end{aligned}$$

since $A^2 = 0$. This gives

$$\Phi_{t,t_0} = \begin{pmatrix} 1 & t-t_0 \\ 0 & 1 \end{pmatrix} \qquad \text{with} \qquad \Phi_{t,t_0}^{-1} = \begin{pmatrix} 1 & -(t-t_0) \\ 0 & 1 \end{pmatrix}.$$

Then, by (B.3), the solution to our SDE is given by

$$\begin{aligned} X_t &= \Phi_{t,t_0}\left(X_{t_0} + \int_{t_0}^{t} \Phi_{s,t_0}^{-1}\mathbf{d}_1(s)\,\mathrm{d}B_s\right) \\ &= \begin{pmatrix} 1 & t-t_0 \\ 0 & 1 \end{pmatrix}\left(X_{t_0} + \nu I_{t,t_0}\right), \end{aligned}$$

where we define

$$I_{t,t_0} := \int_{t_0}^{t} \begin{pmatrix} -(s-t_0) \\ 1 \end{pmatrix} \mathrm{d}B_s.$$

Since the integrand of the Itō integral is non-random, I_{t,t_0} is multivariate Normal with mean 0 and variance given by

$$\begin{aligned} \int_{t_0}^{t} \begin{pmatrix} -(s-t_0) \\ 1 \end{pmatrix} \begin{pmatrix} -(s-t_0) & 1 \end{pmatrix} \mathrm{d}s &= \int_{t_0}^{t} \begin{pmatrix} (s-t_0)^2 & -(s-t_0) \\ -(s-t_0) & 1 \end{pmatrix} \mathrm{d}s \\ &= \begin{pmatrix} \frac{1}{3}(t-t_0)^3 & -\frac{1}{2}(t-t_0)^2 \\ -\frac{1}{2}(t-t_0)^2 & t-t_0 \end{pmatrix}. \end{aligned}$$

Putting this together gives the transition distribution

$$\begin{aligned} X_{t+\Delta t}|\{X_t = x_t\} &\sim \mathcal{N}\big(\Phi_{t+\Delta t,t}x_t,\, \nu^2\Phi_{t+\Delta t,t}\,\mathrm{Var}(I_{t+\Delta t,t})\Phi'_{t+\Delta t,t}\big) \\ &\sim \mathcal{N}\left(\begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} x_t,\, \nu^2 \begin{pmatrix} \frac{1}{3}(\Delta t)^3 & \frac{1}{2}(\Delta t)^2 \\ \frac{1}{2}(\Delta t)^2 & \Delta t \end{pmatrix}\right). \end{aligned}$$

For the simulation study of Subsection 3.2.3 and the athletics analysis of Section 4.1 we use unit time steps so set $\Delta t = 1$.

B.2 Correlated integrated random walks

We now consider the four-dimensional stochastic differential equation (5.1) used for both models in Chapter 5. Written in vector form, we have

$$\mathrm{d}\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \\ X_{t,4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \\ X_{t,4} \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ 0 \\ \nu_{\mathrm{A}} \\ \nu_{\mathrm{B}}\rho \end{pmatrix} \mathrm{d}B_{t,1} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \nu_{\mathrm{B}}\sqrt{1-\rho^2} \end{pmatrix} \mathrm{d}B_{t,2},$$

where $B_{t,1}$ and $B_{t,2}$ are independent Wiener process. This is again a linear SDE of the form (B.2) and so may be solved using (B.3).

Since A and C_i are commuting constants, the fundamental matrix may be found by (B.5) which gives

$$\Phi_{t,t_0} = \begin{pmatrix} 1 & 0 & t-t_0 & 0 \\ 0 & 1 & 0 & t-t_0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The solution can then be shown to equal

$$X_t = \Phi_{t,t_0}\left(X_{t_0} + I_{t,t_0} + J_{t,t_0}\right),$$

where

$$I_{t,t_0} := \int_{t_0}^{t} \begin{pmatrix} -\nu_{\mathrm{A}}(s-t_0) \\ -\nu_{\mathrm{B}}\rho(s-t_0) \\ \nu_{\mathrm{A}} \\ \nu_{\mathrm{B}}\rho \end{pmatrix} \mathrm{d}B_{s,1} \quad \text{and} \quad J_{t,t_0} := \int_{t_0}^{t} \begin{pmatrix} 0 \\ -\nu_{\mathrm{B}}\sqrt{1-\rho^2}(s-t_0) \\ 0 \\ \nu_{\mathrm{B}}\sqrt{1-\rho^2} \end{pmatrix} \mathrm{d}B_{s,2}.$$

As the vector-valued integrands $\mathbf{h}(s)$ are non-random, these Itō integrals are both

multivariate Normal with zero mean and variances given by

$$\mathrm{Var}\left(\int_{t_0}^{t} \mathbf{h}(s)\,\mathrm{d}B_s\right) = \int_{t_0}^{t} \mathbf{h}(s)\,\mathbf{h}(s)'\,\mathrm{d}s.$$

Noting that I_{t,t_0} and J_{t,t_0} are independent as their generating Wiener processes are, the final transition distribution can be shown to be

$$\begin{aligned} X_{t+\Delta t}|\{X_t = x_t\} &\sim \mathcal{N}\big(\Phi_{t+\Delta t,t}x_t,\ \Phi_{t+\Delta t,t}(\mathrm{Var}(I_{t+\Delta t,t}) + \mathrm{Var}(J_{t+\Delta t,t}))\Phi'_{t+\Delta t,t}\big) \\ &\sim \mathcal{N}\left(\begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} x_t,\ \nu^2 \begin{pmatrix} \frac{1}{3}S(\Delta t)^3 & \frac{1}{2}S(\Delta t)^2 \\ \frac{1}{2}S(\Delta t)^2 & S\Delta t \end{pmatrix}\right), \end{aligned}$$

where

$$S := \begin{pmatrix} \nu_{\mathrm{A}}^2 & \nu_{\mathrm{A}}\nu_{\mathrm{B}}\rho \\ \nu_{\mathrm{A}}\nu_{\mathrm{B}}\rho & \nu_{\mathrm{B}}^2 \end{pmatrix}.$$

Appendix C

Derivation of bivariate logistic model

In this appendix we derive the joint density of the p and q-largest components from the bivariate exchangeable logistic model for p and q up to 2. To do this we derive the cdf then the pdf in Fréchet margins before transforming to the required GEV marginal form.

To simplify the notation, throughout this appendix we use X and Y to refer to the first and second components respectively of our bivariate variable (rather than Y_1 and Y_2 used elsewhere in the thesis).

C.1 Joint distribution functions in Fréchet margins

Working initially in Fréchet margins, the bivariate logistic model is defined by (2.25) which in two dimensions is

$$V(x, y) = (x^{-\frac{1}{\alpha}} + y^{-\frac{1}{\alpha}})^{\alpha}, \tag{C.1}$$

where $\alpha \in (0, 1]$ is a dependence parameter. V is defined by (2.24) which in two dimensions is simply

$$V(x, y) = 2 \int_0^1 \max \left\{ \frac{w}{x}, \frac{1-w}{y} \right\} \mathrm{d}H(w).$$

From this we can find H although for our purposes it is easier to work directly with V.

Following the theory of Subsection 2.4.1, the limit distribution of scaled componentwise maxima with Fréchet margins is given by

$$\mathbf{P} \left\{ \frac{M_{n,X}}{n} \leq x_1, \frac{M_{n,Y}}{n} \leq y_1 \right\} \xrightarrow{n \to \infty} G(x_1, y_1) = \exp(-V(x_1, y_1)),$$

where $M_{n,X}$ is the X componentwise maximum of an IID bivariate logistic sample of size n. Writing $M_{n,X}^{(p)}$ for the p-th largest X component of the sample (and similarly for Y) we wish to find $G(x_2, x_1, y_2, y_1)$ defined by

$$\mathbf{P} \left\{ \frac{M_{n,X}^{(2)}}{n} \leq x_2, \frac{M_{n,X}^{(1)}}{n} \leq x_1, \frac{M_{n,Y}^{(2)}}{n} \leq y_2, \frac{M_{n,Y}^{(1)}}{n} \leq y_1 \right\} \xrightarrow{n \to \infty} G(x_2, x_1, y_2, y_1),$$

that is the joint limit distribution of the largest and second largest components of X and Y (in Fréchet margins).

For this we use the point process representation, mirroring the example derivation of the multivariate extreme value distribution on page 50. This involves decomposing the event $\left\{ M_{n,X}^{(2)}/n \leq x_2 \,,\, M_{n,X}^{(1)}/n \leq x_1,\, M_{n,Y}^{(2)}/n \leq y_2,\, M_{n,Y}^{(1)}/n \leq y_1 \right\}$ into the possible configurations of $M_{n,X}^{(1)}/n$ to x_2 and $M_{n,Y}^{(1)}/n$ to y_2 so that the event can be written in terms of the counting process $N(\cdot)$. The five possible configurations are shown in Figure C.1. Referring to the region labels in Figure C.1f, we can then

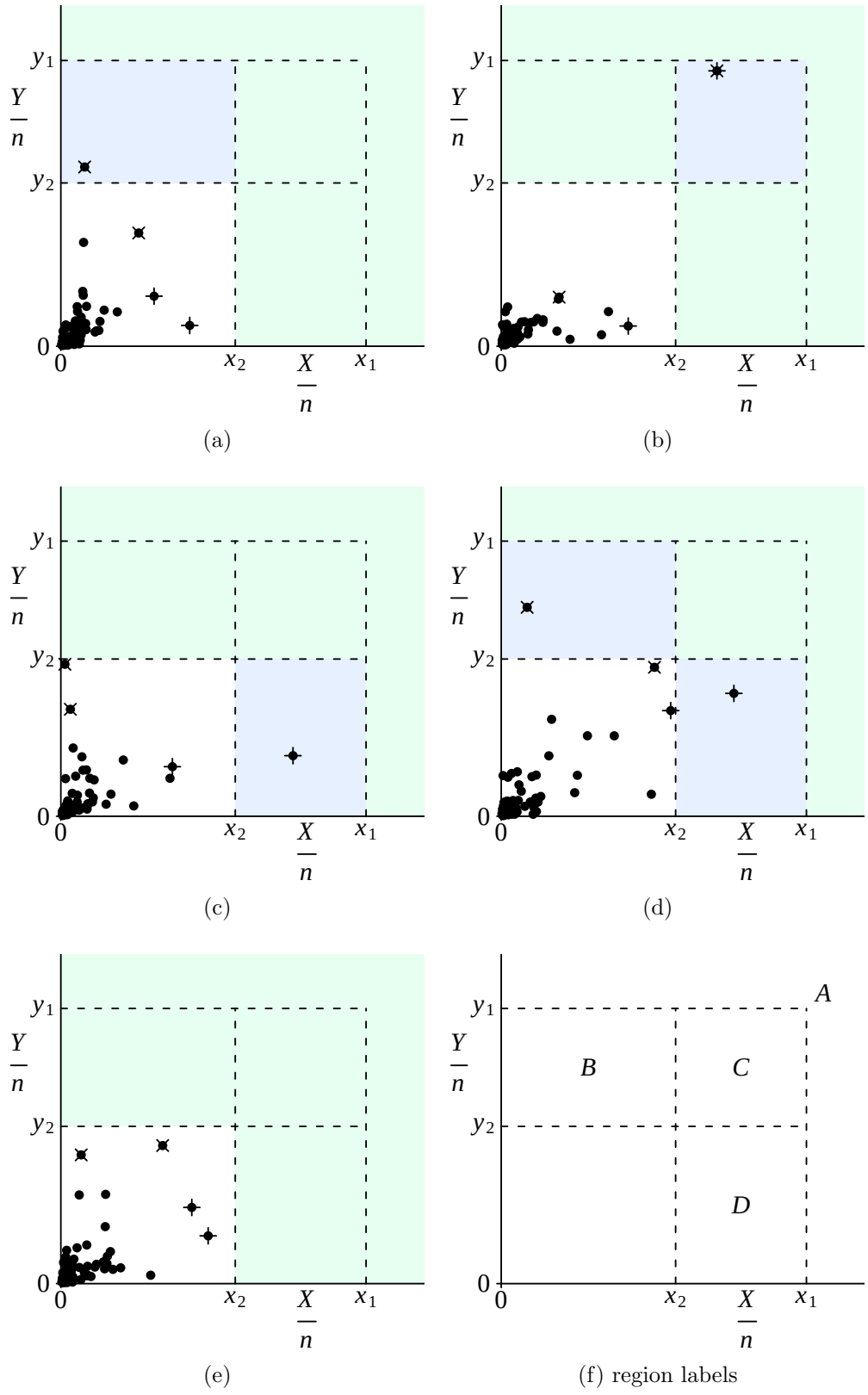


Figure C.1: Decomposition of the bivariate point process event $\left\{M_{n,X}^{(2)}/n \leq x_2,\ M_{n,X}^{(1)}/n \leq x_1,\ M_{n,Y}^{(2)}/n \leq y_2,\ M_{n,Y}^{(1)}/n \leq y_1\right\}$ in terms of the possible relative positions of the componentwise maxima $M_{n,X}^{(1)}/n$ and $M_{n,Y}^{(1)}/n$ to the constants $x_2 \leq x_1$ and $y_2 \leq y_1$. Subfigure (f) provides labels for the regions involved.

decompose the event as

$$\left\{ \frac{M_{n,X}^{(2)}}{n} \leq x_2, \frac{M_{n,X}^{(1)}}{n} \leq x_1, \frac{M_{n,Y}^{(2)}}{n} \leq y_2, \frac{M_{n,Y}^{(1)}}{n} \leq y_1 \right\} =$$
$$\big\{N(A\cup C\cup D) = 0, N(B) = 1\big\} \cup \big\{N(A\cup B\cup D) = 0, N(C) = 1\big\} \cup$$
$$\big\{N(A\cup B\cup C) = 0, N(D) = 1\big\} \cup \big\{N(A\cup C) = 0, N(B) = 1, N(D) = 1\big\} \cup$$
$$\big\{N(A\cup B\cup C\cup D) = 0\big\},$$

where we use $\cup$ to denote the union of two disjoint sets.

We therefore have

$$\begin{aligned} G(x_2, x_1, y_2, y_1) = &\mathbf{P}\big\{N(A\cup C\cup D) = 0\big\}\mathbf{P}\big\{N(B) = 1\big\} + \\ &\mathbf{P}\big\{N(A\cup B\cup D) = 0\big\}\mathbf{P}\big\{N(C) = 1\big\} + \\ &\mathbf{P}\big\{N(A\cup B\cup C) = 0\big\}\mathbf{P}\big\{N(D) = 1\big\} + \\ &\mathbf{P}\big\{N(A\cup C) = 0\big\}\mathbf{P}\big\{N(B) = 1\big\}\mathbf{P}\big\{N(D) = 1\big\} + \\ &\mathbf{P}\big\{N(A\cup B\cup C\cup D) = 0\big\}. \end{aligned}$$

Now, since $N(\mathcal{A}) \sim \text{Poisson}(\Lambda(\mathcal{A}))$ where $\Lambda(\cdot)$ is the intensity measure of the point process, this can be written as

$$\begin{aligned} G(x_2, x_1, y_2, y_1) = &\exp(-\Lambda(A\cup C\cup D))\ \Lambda(B)\exp(-\Lambda(B)) + \\ &\exp(-\Lambda(A\cup B\cup D))\ \Lambda(C)\exp(-\Lambda(C)) + \\ &\exp(-\Lambda(A\cup B\cup C))\ \Lambda(D)\exp(-\Lambda(D)) + \\ &\exp(-\Lambda(A\cup C))\ \Lambda(B)\exp(-\Lambda(B))\ \Lambda(D)\exp(-\Lambda(D)) + \\ &\exp(-\Lambda(A\cup B\cup C\cup D)) \\ = &\exp(-\Lambda(A\cup B\cup C\cup D))\ \big(\Lambda(B) + \Lambda(C) + \Lambda(D) + \Lambda(B)\Lambda(D) + 1\big). \end{aligned}$$

We now wish to write this in terms of $V(x, y)$ and hence find the particular form

for the logistic model. For this we first note that $\Lambda(A) = V(x_1, y_1)$ and similarly $\Lambda(A \cup B \cup C \cup D) = V(x_2, y_2)$; the proof of this given in the example on page 50 ending with the final result in (2.26). By writing the other regions in terms of areas with this shape, we obtain $\Lambda(B) = V(x_2, y_2) - V(x_2, y_1)$, $\Lambda(D) = V(x_2, y_2) - V(x_1, y_2)$ and $\Lambda(C) = V(x_2, y_2) - V(x_1, y_1) - \Lambda(B) - \Lambda(D)$. This gives the final result

$$G(x_2, x_1, y_2, y_1) = \exp(-V(x_2, y_2)) \cdot \\ \big(V(x_2, y_2) - V(x_1, y_1) + (V(x_2, y_2) - V(x_2, y_1))(V(x_2, y_2) - V(x_1, y_2)) + 1\big), \tag{C.2}$$

where, for our bivariate logistic model, we use $V(x, y)$ given by (C.1).

Note that the joint distribution functions for any subset of the two-largest components of X and Y can easily by found from (C.2). For example, the distribution of the componentwise maxima can be verified by setting $x_2 = x_1$ and $y_2 = y_1$ in $G(x_2, x_1, y_2, y_1)$ to obtain $G(x_1, y_1) = \exp(-V(x_1, y_1))$. Similarly

$$G(x_2, x_1, y_1) = \exp(-V(x_2, y_1)) \, \big(V(x_2, y_1) - V(x_1, y_1) + 1\big),$$

and

$$G(x_1, y_2, y_1) = \exp(-V(x_1, y_2)) \, \big(V(x_1, y_2) - V(x_1, y_1) + 1\big).$$

Larger joint distributions functions can be obtained by following a similar strategy as above although the calculation becomes increasingly complicated as the number of different decompositions of the generating event increases.

C.2 Joint density functions in Fréchet margins

We now calculate the joint density functions in Fréchet margins by differentiating the corresponding distribution functions. For the componentwise maxima we have

$$\begin{aligned} g(x_1, y_1) &= \frac{\partial^2}{\partial x_1 \partial y_1} G(x_1, y_1) \\ &= -\frac{\partial}{\partial x_1}\Big[\exp(-V(x_1, y_1))V_y(x_1, y_1)\Big] \\ &= \exp(-V(x_1, y_1))\,\big(V_x(x_1, y_1)V_y(x_1, y_1) - V_{xy}(x_1, y_1)\big), \end{aligned} \tag{C.3}$$

where we use V_x to denote $\partial V/\partial x$ and likewise for V_y and V_{xy}.

We similarly obtain

$$\begin{aligned} &g(x_2, x_1, y_1) = \exp(-V(x_2, y_1)) \cdot \\ &\big(V_x(x_2, y_1)V_{xy}(x_1, y_1) + V_x(x_1, y_1)V_{xy}(x_2, y_1) - V_x(x_1, y_1)V_x(x_2, y_1)V_y(x_2, y_1)\big), \end{aligned}$$

$$\begin{aligned} &g(x_1, y_2, y_1) = \exp(-V(x_1, y_2)) \cdot \\ &\big(V_y(x_1, y_2)V_{xy}(x_1, y_1) + V_y(x_1, y_1)V_{xy}(x_1, y_2) - V_y(x_1, y_1)V_y(x_1, y_2)V_x(x_1, y_2)\big), \end{aligned}$$

and

$$\begin{aligned} g(x_2, x_1, y_2, y_1) &= \exp(-V(x_2, y_2)) \cdot \\ &\Big(V_x(x_1, y_2)V_x(x_2, y_2)V_y(x_2, y_1)V_y(x_2, y_2) - \\ &V_x(x_1, y_1)V_y(x_1, y_1)V_{xy}(x_2, y_2) - V_x(x_1, y_2)V_y(x_1, y_2)V_{xy}(x_2, y_1) - \\ &V_x(x_2, y_1)V_y(x_2, y_1)V_{xy}(x_1, y_2) - V_x(x_2, y_2)V_y(x_2, y_2)V_{xy}(x_1, y_1) + \\ &V_{xy}(x_1, y_1)V_{xy}(x_2, y_2) + V_{xy}(x_1, y_2)V_{xy}(x_2, y_1)\Big). \end{aligned}$$

Using the bivariate logistic form of $V(x,y)$ given in (C.1) it can be shown that

$$\begin{aligned}
V_x(x,y) &= -x^{-\left(1+\frac{1}{\alpha}\right)}(x^{-\frac{1}{\alpha}}+y^{-\frac{1}{\alpha}})^{(\alpha-1)},\\
V_y(x,y) &= -y^{-\left(1+\frac{1}{\alpha}\right)}(x^{-\frac{1}{\alpha}}+y^{-\frac{1}{\alpha}})^{(\alpha-1)},\\
V_{xy}(x,y) &= \left(1-\frac{1}{\alpha}\right)(xy)^{-\left(1+\frac{1}{\alpha}\right)}(x^{-\frac{1}{\alpha}}+y^{-\frac{1}{\alpha}})^{(\alpha-2)}.
\end{aligned}$$

These can then be used in the equations above to give the required density functions.

C.3 Transforming densities to GEV margins

The current joint density functions can only be applied to data known to have Fréchet marginal distributions, in which case $M_{n,X}^{(1)}/n$ also has a Fréchet distribution. For arbitrary variables, we know from Subsection 2.3.1 that if $(M_{n,X}^{(1)}-a_n)/b_n$ has a non-degenerate distribution, it must be GEV(μ_X,σ_X,ξ_X). We can therefore obtain density functions for variables with arbitrary margins by transforming the X components with the probability integral transform of Fréchet to GEV(μ_X,σ_X,ξ_X) (and similarly for the Y components).

To do this we create the transformation $t_X(x) := F^{-1}(G(x|\mu_X,\sigma_X,\xi_X))$, where $F(x)=\exp(-1/x)$ is the Fréchet and $G(x|\mu_X,\sigma_X,\xi_X)$ the GEV(μ_X,σ_X,ξ_X) distribution functions. Then if $X_F \sim$ Fréchet, $X_G := t^{-1}(X_F) \sim$ GEV(μ_X,σ_X,ξ_X). Doing the same for Y, we can transform $g_F(x_1,y_1)$, the joint density for componentwise maxima given in (C.3), from Fréchet to GEV margins using

$$g_G(x_1,y_1) = g_F(t_X(x_1),t_Y(y_1))\,|t_X'(x_1)|\,|t_Y'(y_1)|.$$

The remaining joint densities are transformed with

$$g_G(x_2, x_1, y_1) = g_F(t_X(x_2), t_X(x_1), t_Y(y_1))\,|t'_X(x_2)|\,|t'_X(x_1)|\,|t'_Y(y_1)|,$$

$$g_G(x_1, y_2, y_1) = g_F(t_X(x_1), t_Y(y_2), t_Y(y_1))\,|t'_X(x_1)|\,|t'_Y(y_2)|\,|t'_Y(y_1)|,$$

and

$$\begin{aligned} g_G(x_2, x_1, y_2, y_1) = {} & g_F(t_X(x_2), t_X(x_1), t_Y(y_2), t_Y(y_1)) \cdot \\ & |t'_X(x_2)|\,|t'_X(x_1)|\,|t'_Y(y_2)|\,|t'_Y(y_1)|. \end{aligned}$$

Referring to the GEV distribution function (2.17), we have

$$t_X(x) = \left[1 + \xi_X \left(\frac{x - \mu_X}{\sigma_X}\right)\right]_+^{\frac{1}{\xi_X}},$$

and so

$$t'_X(x) = \frac{1}{\sigma_X}\left[1 + \xi_X \left(\frac{x - \mu_X}{\sigma_X}\right)\right]_+^{\frac{1}{\xi_X}-1}.$$

By substituting these into the equations above with the Fréchet densities and the partial derivatives of V, we can write down the exact densities for the upper extremes of the bivariate logistic model. These now have the required marginal distributions and therefore depend upon marginal parameters μ_X, σ_X, ξ_X, μ_Y, σ_Y, ξ_Y as well as the dependence parameter α.

Bibliography

Alspach, D. and Sorenson, H. (1972). Non-linear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448. 8

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal filtering*. Prentice-Hall Englewood Cliffs, New Jersey. 8, 22, 84

Andrieu, C., Doucet, A., and Tadic, V. B. (2005). On-line parameter estimation in general state-space models. In *Proceedings of the 44th Conference on Decision and Control*, pages 332–337. Citeseer. 20

Barão, M. I. and Tawn, J. A. (1999). Extremal analysis of short series with outliers: sea-levels and athletics records. *Applied Statistics*, 48(4):469–487. 121, 129

Beirlant, J. (2004). *Statistics of Extremes: Theory and Applications*. Probability and Statistics. Wiley. 47

Berzuini, C., Best, N. G., Gilks, W. R., and Larizza, C. (1997). Dynamic conditional independence models and Markov chain Monte Carlo methods. *Journal of the American Statistical Association*, 92(440):1403–1412. 10

Bolic, M., Djuric, P. M., and Hong, S. (2004). Resampling algorithms for particle filters: A computational complexity perspective. *EURASIP Journal on Applied Signal Processing*, 15(5):2267–2277. 17

Briers, M., Doucet, A., and Maskell, S. (2004). Smoothing algorithms for state-

space models. *Submitted to IEEE Transactions on Signal Processing*. 20, 25, 93, 95

Briers, M., Doucet, A., and Singh, S. S. (2005). Sequential auxiliary particle belief propagation. In *8th International Conference on Information Fusion*, volume 1, pages 705–711. 80

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings–Radar, Sonar and Navigation*, 146(1):2–7. 17, 154

Casella, G. and Robert, C. P. (2001). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94. 18

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London. 28, 32, 37, 43

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: A review and new developments. *International statistical review*, 64(1):119–136. 32

Coles, S. G. and Tawn, J. A. (1990). Statistics of coastal flood prevention. *Philosophical Transactions: Physical Sciences and Engineering*, 332(1627):457–476. 129

Coles, S. G. and Tawn, J. A. (1991). Modelling multivariate extreme events. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 53:377–392. 51

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design. *Applied Statistics*, 43(1):1–48. 47, 51

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4):463–478. 37

Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 62(1):191–208. 44

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 52(3):393–42. 37, 39, 43

de Haan, L. (1985). *Extremes in higher dimensions: the model and some statistics*. Erasmus University. 49

de Haan, L. and de Ronde, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes*, 1(1):7–45. 51

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Operations Research and Financial Engineering. Springer. 28

de Haan, L. and Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Probability Theory and Related Fields*, 40(4):317–337. 49, 51

Dekkers, A. L. M., Einmahl, J. H. J., and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17(4):1833–1855. 33

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38. 93

Dixon, M. J. and Tawn, J. A. (1992). Trends in UK extreme sea-levels: a spatial approach. *Geophysical Journal International*, 111(3):607–616. 129

Dixon, M. J. and Tawn, J. A. (1999). The effect of non-stationarity on extreme sea-level estimation. *Applied Statistics*, 48(2):135–151. 130

Doucet, A. (1998). On sequential simulation-based methods for Bayesian filtering. Technical Report CUED/F-INFENG/TR.310, Signal Processing Group, Department of Engineering, Cambridge University. 13

Doucet, A., Briers, M., and Sénécal, S. (2006). Efficient block sampling strategies for sequential Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 15(3):693–711. 82

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer. 5, 184, 186

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208. 18, 24

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 58(1):25–45. 43

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer. 28, 37, 39

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(10):143–10. 9

Fearnhead, P. (1998). *Sequential Monte Carlo methods in filter theory*. PhD thesis, University of Oxford. 19, 68, 69, 149, 150

Fearnhead, P. (2002). MCMC, sufficient statistics and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862. 19

Fearnhead, P., Wyncoll, D. P., and Tawn, J. A. (2009). A sequential smoothing algorithm with linear computational cost. *To appear in Biometrika*. iv, 105

Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 65(2):545–556. 41

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. 29

Gaetan, C. and Grigoletto, M. (2004). Smoothing sample extremes with dynamic models. *Extremes*, 7(3):221–236. 45, 99, 101, 102, 105

Gilks, W. R. and Berzuini, C. (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 63(1):127–146. 19

Godsill, S. J., Doucet, A., and West, M. (2004). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99(465):156–169. 24, 75, 96

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113. 10, 12, 19, 68

Gumbel, E. J. (1960). Distributions des valeurs extrèmes en plusieurs dimensions. *Publications de L'Institut de Statistique, Paris*, 9:171–173. 48

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, 15(2):153–167. 44

Handschin, J. E. (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6:555–563. 10

Handschin, J. E. and Mayne, D. Q. (1969). Monte Carlo techniques to estimate the

conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9:547–559. 10

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(3):497–546. 52

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261. 33

Hsing, T. (1987). On the characterization of certain point processes. *Stochastic processes and their applications*, 26(2):297–316. 40

Hsing, T., Hüsler, J., and Leadbetter, M. R. (1988). On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, 78(1):97–112. 40

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300. 64

Hürzeler, M. and Künsch, H. R. (1998). Monte Carlo approximations for general state-space models. *Journal of Computational and Graphical Statistics*, 7(2):175–193. 24, 96

Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximizing the likelihood for a general state-space model. In Doucet et al. (2001), pages 159–176. 20

Jazwinski, A. H. (1973). Stochastic processes and filtering theory. *Academic Press*. 8

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171. 29

Joe, H., Smith, R. L., and Weissman, I. (1992). Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 54(1):171–183. 51

Julier, S. J., Uhlmann, J. K., and Durrant-Whyte, H. F. (1995). A new approach for filtering nonlinear systems. In *Proceedings of the 1995 American Control Conference*, volume 3, pages 1628–1632, Seattle, Washington. 8

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45. 7, 84

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. 10, 20, 22, 25, 94, 103

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288. 13, 14

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. World Scientific. 47, 48, 128

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65(2):291–306. 41

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag New York. 29, 40, 41, 43

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187. 51, 52

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 59(2):475–499. 52

Ledford, A. W. and Tawn, J. A. (1998). Concomitant tail behaviour for extremes. *Adv. in Appl. Probab*, 30(1):197–215. 52

Liu, J. S. and Chen, R. (1995). Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576. 14, 17, 53, 55, 59

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044. 13, 18

Liu, J. S. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In Doucet et al. (2001), pages 197–224. 19

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44(2):226–233. 97

O'Brien, G. L. (1987). Extreme values for stationary and Markov sequences. *Annals of Probability*, 15(1):281–291. 39

Pauli, F. and Coles, S. (2001). Penalized likelihood inference in extreme value analyses. *Journal of Applied Statistics*, 28(5):547–560. 44

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability*, 8:745–756. 34

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131. 38

Pickands, J. (1981). Multivariate extreme value distributions. In *Proceedings of the 43rd Session of the International Statistical Institute*, pages 859–878. 48, 49

Pickands, J. (1994). Bayes quantile estimation and threshold selection for the generalized Pareto family. In Galambos, J., Lechner, J., and Simiu, E., editors,

Extreme Value Theory and Applications, pages 123–138. Kluwer Academic, Dordrecht. 39

Pitt, M. K. (2002). Smooth particle filters for likelihood evaluation and maximisation. Warwick Economic Research Papers 651, University of Warwick, Department of Economics. 20

Pitt, M. K. and Shephard, N. (1999a). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599. 14, 16, 26, 64, 66, 68, 70, 78, 145

Pitt, M. K. and Shephard, N. (1999b). Time-varying covariances: A factor stochastic volatility approach (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 547–570. Oxford University Press. 89

Poyiadjis, G., Doucet, A., and Singh, S. S. (2005). Particle methods for optimal filter derivative: Application to parameter estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 925–928. 20

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3):723–724. 32

Prescott, P. and Walden, A. T. (1983). Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *Journal of Statistical Computation and Simulation*, 16(3):241–250. 32

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*, volume 4 of *Applied Probability*. Springer, New York. 47, 49

Robinson, M. E. and Tawn, J. A. (1995). Statistics for exceptional athletics records. *Applied Statistics*, 44(4):499–511. 98, 101, 105, 106, 120, 121

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667. 89

Sibuya, M. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11:195–210. 51

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90. 32

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86:27–43. 33

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–393. 37, 43

Smith, R. L. (1997). Comment on "Statistics for exceptional athletics records", by Robinson, M. E. and Tawn, J. A. *Applied Statistics*, 46(1):123–128. 99

Smith, R. L. and Miller, J. E. (1986). A non-Gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 48(1):79–88. 45

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 56:515–515. 41, 111

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, 50(2):281–289. 19, 93, 101

Tanizaki, H. (2001). Nonlinear and non-Gaussian state space modeling using sampling techniques. *Annals of the Institute of Statistical Mathematics*, 53(1):63–81. 24, 99

Tawn, J. A. (1988a). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415. 129

Tawn, J. A. (1988b). An extreme-value theory model for dependent observations. *Journal of Hydrology*, 101(1):227–250. 33

Tawn, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253. 129

Tawn, J. A. (1992). Estimating probabilities of extreme sea-levels. *Applied Statistics*, 41(1):77–93. 129

Vaughan, D. G., Marshall, G. J., Connolley, W. M., Parkinson, C., Mulvaney, R., Hodgson, D. A., King, J. C., Pudsey, C. J., and Turner, J. (2003). Recent rapid regional climate warming on the Antarctic Peninsula. *Climatic Change*, 60(3):243–274. 108

Von Mises, R. (1954). La distribution de la plus grande de n valeurs. In *Selected Papers*, volume 2, pages 271–294. American Mathematical Society, Providence, RI. 29

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815. 33

Wills, A. G., Schön, T. B., and Ninness, B. (2008). Parameter estimation for discrete-time nonlinear systems using EM. In *Proceedings of the 17th IFAC World Congress, Seoul, South Korea*. 20, 93, 95

Index

Adapted filter, 16, 64, 143, 145, 151

Antarctic temperatures, 107–119

Artificial prior distribution, *25*, 25–27, 78, 80

Asymptotic independence, 51, 52

Asymptotic Independence of Maxima, 39, 40

Autoregressive model, 59–64, 76, 82

Auxiliary SIR filter, *14–16*, 26, 54, 60, 74, 78, 145, 151

Backwards filter, *see* Backwards information filter

Backwards information filter, *25–26*, 75, 78, 80

Bayesian bootstrap filter, *see* Sampling Importance Re-sampling filter

Bearings-only tracking model, 68–74, 149

Beta distribution, 45

Bivariate exchangeable logistic model, 123, 132, 160, 171–178

Bivariate Gaussian distribution, 135, 158, 168, *see also* Multivariate Gaussian distribution

Blocks method of declustering, 133

Clustered non-homogeneous Poisson process, 42

Coefficient of tail dependence, 52

Cross-validation, 44, 109

Declustering

blocks method, *see* Blocks method of declustering

runs method, *see* Runs method of declustering

Domain of attraction, 31, 34

Effective sample size, *14*, 16, 53–74, 85, 88, 89, 154
Ensemble Kalman Filter, 9
Expectation-Maximisation algorithm, 20, 93–97, 101, 103, 115, 126, 137
Exponential distribution, 45
Extended Kalman Filter, 8
Extremal index, 41, 42
Extremal Types Theorem, 29

Filter-Smoother, *22–24*, 75, 77, 84–88, 90, 92
Filtering
    distribution, 7
    Kalman, *see* Kalman Filter
    particle, *see* Particle filter
    problem, 7, 21
Forward-Backward Smoother, *24*, 75–77, 84
Fréchet distribution, 29, 30, 47–49, 52, 123, 171, 176–178

Gamma distribution, 45
Gaussian distribution, 29, 31, 44, 60, 73, 146, 151
Gaussian sum filter, 8
Generalised Extreme Value distribution, *29–33*, 34–43, 46, 49, 99, 103, 123, 177, 178
Generalised Pareto distribution, *37–39*
Gumbel distribution, 29, 30, 45

Hidden Markov model, 6, 45

Importance sampling, 13, 17, 150, 154, 155, 159

Joint smoothing distribution, 9, 21, 22, 25, 75, 103

Kalman Filter, *8*, 7–9, 45, 84, 144, 147

Kalman Smoother, 22, 84, 85, 103

Kernel smoothing, 108–109

Linear-Gaussian model, *7*, 60, 83–88, 90–91, 143

Log-normal distribution, 65

Logistic model, *see* Bivariate exchangeable logistic model

Marginal smoothing distribution, 21, 24, 75–77, 79, 94, 104

Marginalisation, *see* Rao-Blackwellisation

Markov Chain Monte Carlo, 9, 17, 19, 21, 32, 89, 155–156, 161–163

Max-stability, 30, 48

Metropolis algorithm, 156, 161, *see also* Markov Chain Monte Carlo

Multinomial sampling, 17, 55, 58, 59

Multivariate exchangeable logistic distribution, 48

Multivariate extremal point process model, *49–51*, 123, 172–174

Multivariate extreme value distribution, 47–50

Multivariate extreme value theory, 47–52

Multivariate Gaussian distribution, 7, 16, 22, 45, 102, 114, 124, 148, 156, 164, 168, 170, *see also* Gaussian distribution

Multivariate Normal distribution, *see* Multivariate Gaussian distribution

Negative-Weibull distribution, 29, 30

Non-homogeneous Poisson process, 34–36, 42, 49

Normal distribution, *see* Gaussian distribution

One-step prediction distribution, *see* Prediction distribution

Parameter estimation in state space models, *18–20*, 93–97, 103–104, 115, 125–127, 136–138

Pareto distribution, 45

Particle filter, *9–20*

auxiliary, *see* Auxiliary SIR filter

backwards, *see* Backwards information filter

enhancements, 16–18

implementation, 143–165

initialisation, 17

parameter estimation, *see* Parameter estimation in state space models

Particle smoother, *22–27*, 75–92

Peaks Over Threshold, 38, 43

Point process

clustered Poisson process, *see* Clustered non-homogeneous Poisson process

for multivariate extremes, *see* Multivariate extremal point process model

for univariate extremes, *see* Univariate extremal point process model

Poisson process, *see* Non-homogeneous Poisson process

Poisson distribution, 35

Poisson process, *see* Non-homogeneous Poisson process

Prediction distribution, 7, 8, 11, 20, 45

Probability integral transform, 47, 49, 177

Probability-probability plot, 33, 104, 125

Quantile-quantile plot, 33, 104, 125

r-largest order statistics model, *33*, 99, 101, 123, 151

Rao-Blackwellisation, 18, 102, 125, 151, 157, 161

Re-sampling, 11–13

multinomial, *see* Multinomial sampling

residual, *see* Residual sampling

stratified, *see* Stratified sampling

when to re-sample, 14, 53–74

Residual sampling, 17, 55, 57, 60

Runs estimator of the extremal index, 41

Runs method of declustering, 41, 111

Sampling Importance Re-sampling filter, *10–12*, 16

Sea-levels, 129–142

Sequential Importance Sampling, 13

Sequential imputation, 13, 55

Slowly varying function, 52

Smoothing

distribution, *see* Marginal smoothing distribution

joint distribution, *see* Joint smoothing distribution

particle, *see* Particle smoother

problem, 21

Spectral distribution function, 48, 51

State space model, 5–6

Stochastic differential equation, 84, 113, 124, 134, 166–170

Stochastic volatility model, 54, 64–67, 89–92, 145

Stratified sampling, 17, 57, 154

Taylor approximation, 16, 146, 151, 157

Threshold stability, 38

Two-Filter Smoother, *25–27*, 75–78, 80, 84

Unified Extremal Types Theorem, 31

Unified Extremal Types Theorem for stationary sequences, 40

Univariate extremal point process model, *34–39*, 42–43, 112, 113, 154

Univariate extreme value theory, 28–46

Unscented Kalman Filter, 8

Weibull distribution, 45, *see also* Negative-Weibull distribution

Women's 1500m running event, 120–128

Women's 3000m running event, 98–106, 120–128