Pricing Catastrophic Mortality Bonds Using State Space Models

Zhifeng Zhang

A Thesis

In the Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Science (Mathematics) at

Concordia University

Montreal, Quebec, Canada

October 2013

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Zhifeng Zhang

Entitled: Pricing Catastrophic Mortality Bonds Using State Space Models

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair
Dr. J. Garrido

_____ Examiner
Dr. W. Sun

_____ Thesis Supervisor
Dr. P. Gaillardetz

Approved by _____
Chair of the Department or Graduate Program Director

_____
Dean, Faculty of Arts and Science

Date _____

ABSTRACT

Pricing Catastrophic Mortality Bonds Using State Space Models

Zhifeng Zhang

Catastrophic mortality bonds are designed to hedge against the mortality risks. The payoff at maturity depends on the realized mortality index over the life of the bond, therefore modeling the mortality index is the main concern in our study. Since mortality shocks are detected using outlier analysis, non-Gaussian state space models with a fat-tailed error term are proposed to fit the mortality index and to handle shocks. By comparing several state space models with different fat-tailed distributions, an ARIMA process for the baseline mortality and the t-distribution for capturing mortality shocks are chosen. We obtain the price of the mortality bond using the proposed model and estimate the market price of risk. It appears that the market price of risk is lower than the ones obtained in the literature, which is consistent with the industrial empirical results from Wang (2004). This implies that our model is capable of handling mortality risks.

Acknowledgements

First and foremost, I would like to express my profound gratitude to my supervisor, Dr. Patrice Gaillardetz, for his support and patient guidance. His optimistic attitude and energetic encouragement impress me a lot. Without his help and participation during the process of writing the thesis, it may never have been completed.

Further, I would also like to extend my thanks to Dr. José Garrido, Dr. Cody Hyndman, Dr. Ewa Duma, Dr. Sun Wei, and Ms. Marie-France Leclere. They were always very kind and supportive to me during my study in Concordia University.

In addition, I thank my friends for their assistance and encouragement. Lastly, I am very grateful to my family, especially my parents for their understanding and love.

Contents

List of Figures x

List of Tables xi

Introduction 1

1 Data and Shock Detection 5

1.1 Data 5

1.2 Shock Detection 9

1.3 Remarks 19

2 State Space Models 21

2.1 The Linear State Space Model 22

2.2 Filtering and Smoothing: the Conditional Approach 23

2.2.1 Filtering 23

2.2.2 Smoothing 25

2.3 State Space Model with Normal Error Terms 26

2.3.1 Filtering and Smoothing with the Conditional Approach . . . 26

2.3.2 Kalman Filter and Smoother 29

2.4 State Space Model with Non-Normal Error Terms: Importance Sampling 32

2.4.1 Importance Sampling . . . 33
2.4.2 Importance Sampling Algorithm . . . 35
2.5 Simulation Smoother and Antithetic Variables . . . 36
2.5.1 Simulation Smoother . . . 36
2.5.2 Antithetic Variables . . . 38
2.6 Monte Carlo Likelihood Estimation . . . 38
2.6.1 Likelihood Function of Gaussian State Space Model . . . 39
2.6.2 Adjustment Factors . . . 40
2.7 Diagnostic Test . . . 41

3 Mortality Rate Models **43**

3.1 ARIMA Models . . . 44
3.2 Gaussian Mortality Models . . . 46
3.3 Non-Gaussian Mortality Models . . . 48
3.3.1 Model Calibration . . . 50
3.3.2 Estimation of Parameters . . . 51
3.3.3 Model Diagnostic . . . 53
3.3.4 Model Fitting and Forecast . . . 54
3.4 The Outlier Analysis and State Space Models . . . 57
3.5 Comparison of Mortality Rate Models . . . 59
3.5.1 Lin and Cox (2008) . . . 59
3.5.2 Milidonis et al. (2011) . . . 60
3.6 Pricing Mortality Bond . . . 62
3.6.1 Design of the Swiss Re Bond . . . 62
3.6.2 Market Price of Risk . . . 63

Conclusion **67**

References **72**

A The Simulation Smoother **73**

Appendix 73

List of Figures

1.1 Mortality rate q_t from year 1900 - 2008 7

1.2 Logit transformed mortality rate y_t, years 1900 - 2008 8

1.3 Univariate linear regression model: fitted value (left) and residuals (right) 10

1.4 QQ-plot (upper), ACF (middle) and PACF (lower) of residuals for ARIMA(3,0,1) and ARIMA(3,0,3) + AO(1918,1921,1940,1907) + IO(1946) 16

1.5 QQ-plot (upper), ACF (middle) and PACF (lower) of residuals for ARIMA(1,0,0) and ARIMA(2,0,0) + AO(1918,1921,1940,1907) + IO(1946) 17

1.6 Plot of q_t with five detected outliers 19

2.1 Filtering process 24

3.1 The observations y_t v.s. fitted value $\tilde{\boldsymbol{x}}_t$ 48

3.2 The smoothed values and normal QQ-plot of $\tilde{\eta}_t$ (lower) and $\tilde{\varepsilon}_t$ (upper) 49

3.3 Diagnostic tests on standardized one-step ahead prediction error e_t . 54

3.4 Diagnostic tests on standardized smoothed observation error $\tilde{\varepsilon}_t$ 55

3.5 The plot of data y_t, one-step ahead forecast $\hat{\theta}_t$, filtered values $\bar{\theta}_t$ and smoothed values $\tilde{\theta}_t$ 56

3.6 Thirty-year prediction of y_t: mean values of simulated paths and 1% confidence interval. 56
3.7 Thirty-year prediction of q_t: mean values of simulated paths and 1% confidence interval. 57
3.8 $\tilde{\varepsilon}_t$ marked with detected outliers: $\tilde{\varepsilon}_t^{Normal-dist}$ (upper) and $\tilde{\varepsilon}_t^{t-dist}$ (lower) 58
3.9 Empirical probability distribution $f_X(x)$ and Wang transformed probability distribution $f_X^*(x)$ with 50000 simulation paths 65

List of Tables

1.1 Weights of age intervals 6

1.2 Weights of countries 6

1.3 Detected shocks with AIC and BIC criterion 18

2.1 The dimension of variables in linear SSM 22

3.1 Gaussian SSM models selection 47

3.2 Parameter estimation for ARIMA$(1, 0, 0)$ 47

3.3 SSM models comparison 52

3.4 Estimated parameters for non-Gaussian SSM 53

3.5 Comparison of shock effects 58

3.6 Comparison of market price of risk λ among different models 66

Introduction

Catastrophic mortality bonds are designed to hedge against the mortality shock, which is a sudden increase in mortality rates over a short period. Its payoff at maturity is based on a realized mortality index, which could be the weighted average of annual mortality rates among different countries and ages. In the case of the Swiss Re Bond, for instance, the payoff would be less than the face value if the mortality rate raises above 130% of the reference mortality index. As a result, whether there would be catastrophic mortality events in the future or not affects the payoff and thus the price of catastrophic mortality bonds. Calibrating a mortality model and forecasting possible catastrophic mortality rates become the main concern of this thesis.

Mortality rates are affected by various factors. On the one hand, mortality rates were decreasing during recent decades because of improvements in medical care, hygienic conditions, the establishment of global health systems, etc. On the other hand, the possibility of catastrophic mortality rates cannot be ignored because of higher percentage of populations at older ages, increased urban population density, and the increased human mobility (see Huynh et al., 2012). During the past one hundred years, the mortality index experienced an extreme event in 1918 caused by the Spanish flu pandemic and several relatively smaller shocks. Naturally, how to model shocks that appeared in the past and that could also appear in the future

becomes our concern. Several mortality rate models have been introduced in the literature.

Lee and Carter (1992) develop a mortality model that includes mortality changes in terms of age and time. They forecast this mortality index and point out that using only the recent data decreases both the average and width of confidence intervals, since the volatile period (the earlier period) is not taken into account. This paper provides a fundamental idea to study mortality. Girosi and King (2007) provide good insights and complimentary comments on this model. To fully consider mortality shocks, a model compatible with multiple shocks is needed.

Lin and Cox (2008) model the mortality rates using a general Weiner process based on the difference of the logarithmic mortality rates, which guarantees that the mortality rates are always positive. They introduce another log-normal variable to represent the scale of possible shocks and a Bernoulli random variable for shock occurrences. This is an improvement for mortality models by introducing a stochastic process to model shocks.

Milidonis et al. (2011) propose a two-regime switching model to fit the mortality index. The scale of shocks is also independent of time but the occurrence of shocks depends on the transition probability matrix. Based on their results, the estimated mean values of two regimes are very closed and the drift level of volatile regime still has a decreasing tendency. On the other hand, the variance of volatile regime is estimated to be much higher than stable regime. This implies that the shocks are modeled through volatility parameters instead of drift parameters.

There are other different ways to model mortality index. For example, Deng et al. (2012) apply a Brownian motion to model baseline mortality rates. In terms of modeling shocks, they use a mixture distribution to measure the scale of shocks

(different parameters for positive shocks and negative shocks) and use a Poisson process to model the occurrence of shocks. There are two different assumptions in this model. First, this model differentiates positive shocks and negative shocks by using different parameters. Second, it allows more than one shock each year following a Poisson process. The paper shows that the positive shocks have larger severity but fewer occurrences.

We can use different approaches to model baseline mortality rates, the shock effects, and the occurrence of shocks. For instance, Lin and Cox (2008) consider that mortality shocks are transient and independent of time while Milidonis et al. (2011) allows the possibility of dependent mortality shocks. In this thesis, we propose a state space model (SSM) to fit the mortality index. Basically, it is a dynamic system including an observation and state equations, with the advantage, as in other time series models, that the fitted values can be compared with historical data. We found that the linear form of SSM is flexible enough to model the baseline mortality.[1] One of the advantages is that we can model the baseline mortality rates with latent variables through the state equation. Another advantage is that shocks can be modeled with an additive term in the observation equation. Unlike Lin and Cox (2008), there is no trigger (discrete regimes) to indicate the occurrence of shocks. Instead, a fat-tailed distribution to measure the shock effects is adopted. Whenever there is a shock, this distribution should be able to capture the extreme values without conditional on the trigger of occurrence. The diagnostic tests will fully explain the reasonability of our model setting.

To model mortality shocks reasonably, non-Gaussian state space models could

[1] In terms of modeling using the non-linear form, refer to Durbin and Koopman (2001), Jungbacker and Koopman (2007), and a non-linear state space model application Ward et al. (2007).

be applied. The Gaussian SSMs can be represented by matrices and be estimated using the Kalman filter and smoother (Durbin and Koopman, 2001). However, it cannot deal with most fat-tailed distributions. In terms of non-Gaussian state space models, Kitagawa (1987) proposes a numerical algorithm which approximates the integrals derived from conditional distributions. This algorithm is not practical because of its computationally inefficiency. A more practical algorithm using Monte Carlo simulation was proposed by Kitagawa (1996).

Durbin and Koopman (1997) propose another algorithm to deal with non-Gaussian SSMs, which is based on the Kalman filter and smoother and the importance sampling technique. An efficient and simple simulation smoother has been proposed by Durbin and Koopman (2002). According to their algorithm, variance matrices do not need to be calculated, which reduces computational time. Another efficient simulation smoother is introduced by De Jong and Shephard (1995). Durbin and Koopman (2002) compare these two algorithms in detail.

Finally we will price the Swiss Re mortality bond issued in 2003. Mortality bonds are securities used to hedge the catastrophic mortality events. The discounted cash flow mainly depends on the realized mortality rates, which will be calculated from our state space model. Given that the actuarial present value of the bond will not match its actual face value, Lin and Cox (2008) explain this discrepancy by introducing the mortality market price of risk, which can be calculated using the Wang Transform (Wang, 2002). Finally, we will compare our market price of risk with other models.

Chapter 1

Data and Shock Detection

In this chapter, we construct the mortality index that will be used in our analysis of the Swiss Re mortality bond. As discussed previously, we want to identify the shocks (catastrophic mortality rates) to remove them in order to construct baseline mortality rates (defined as moratlity rates without shock effects; see Huynh et al., 2012). Statistically, shocks can be regarded as outliers which are extreme values that the standard normal error cannot handle. The outlier detection technique proposed by Chen and Liu (1993) and Cryer and Chan (2008) can be used to identify the presence of shocks where the baseline mortality rates are modeled using ARIMA models. The main goals are to confirm the existence of shocks and show that ARIMA models can fit baseline mortality rates.

1.1 Data

We construct the mortality index used in Swiss Re bond discussed by Krutov (2010). The index is a weighted average of mortality rates among genders, five countries (including USA, England, France, Italy and Switzerland) and twelve age intervals.

The weighted average mortality index at a given year can be calculated as follow:

$$q_t = \sum_{j=1}^{5} C_j \sum_{i=1}^{12} A_i (G^m q_{ijt}^m + G^f q_{ijt}^f), \qquad t = 1900, 1901, \cdots, 2008,$$

where A_i is the weight for age interval i (shown in Table 1.1), C_j is the weight for country j (shown in Table 1.2), $G_m = 65\%$ and $G_f = 35\%$ represent the weights of male and female respectively, and q_{ijt}^m and q_{ijt}^f represent the mortality rates of male and female for age interval i, country j, and year t. Therefore, we use mortality rate tables for these five countries between age 20 to 79 from year 1900 to 2008. The mortality tables for USA between 1900-1932 are taken from the Human Life Table Database[1]. The mortality tables for the other four countries and USA, between 1933-2008, are taken from the Human Mortality Database[2]. Figure 1.1 shows the annual mortality rates from 1900-2008.

Table 1.1: Weights of age intervals

Age Interval	20-24	25-29	30-34	35-39	40-44	45-49
A_i	1%	5%	12.5%	20%	20%	16%
Age Interval	50-54	55-59	60-64	65-69	70-74	75-79
A_i	12%	7%	3%	2%	1%	0.5%

Table 1.2: Weights of countries

Country	U.S.	U.K.	France	Switzerland	Italy
C_j	70%	15%	7.5%	5%	2.5%

Several observations can be drawn from Figure 1.1. First, there is a general decreasing tendency for mortality rates from 1900 - 2008 which could be explained

[1] http://www.lifetable.de/.

[2] http://www.mortality.org/.

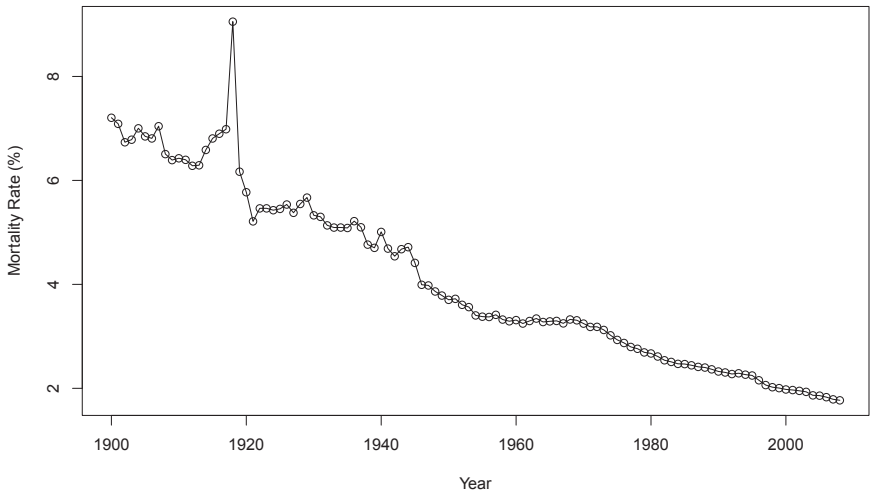

Figure 1.1: Mortality rate q_t from year 1900 - 2008

by medical improvement, health monitoring systems, improved living environments, etc. Second, the mortality rates before 1945 had larger volatilities than mortality rates after 1945. Two possible reasons for this gap are the improvement of data collection techniques and reduction of the number of wars in the reference countries. In addition, the slope of the decreasing tendency before 1945 is steeper than after 1945; in other words, the mortality index decreases at a slower rate after 1945. Mortality rates may have a structural change between these two periods (see Li et al. 2011 for structural change detection). Finally, there was a big shock in 1918, which was caused by the Spanish flu epidemic (Morens et al., 2010). The shock was transient and it disappeared right after that year. Actually, there were more shocks during the past one hundred years which were also transient, such as positive shocks caused by World War II in 1940 (Li and Chan, 2005).

We consider to apply the Logit transform on q_t so that the variation gap can be

reduced and negative q_t values would be avoided. Let

$$y_t = \ln(\frac{100q_t}{1-q_t}), \tag{1.1}$$

where q_t is the mortality index from Figure 1.1. The transformed mortality index is shown in Figure 1.2. The differences between q_t and y_t are mainly due to the convexity of the Logit transformation function.

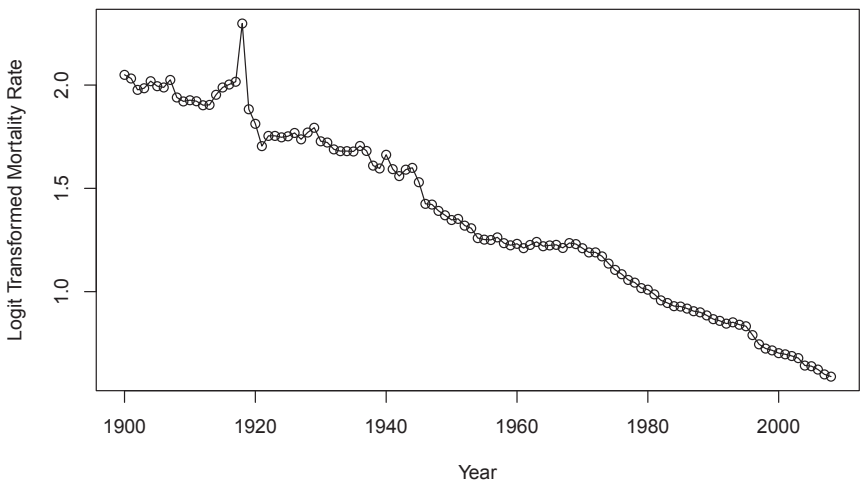

Figure 1.2: Logit transformed mortality rate y_t, years 1900 - 2008

The transformed data y_t includes some modifications compared to q_t, which are better for modeling. First, the difference in slopes for y_t is not as obvious as for q_t, so the general decreasing tendency becomes more linear. It would be easier to apply a simple regression model. In addition, while the shock in 1918 is still obvious, it is much smaller in scale. Hence the transformed mortality index possesses a more stable simulation result which would help fit a standard ARIMA model.

Unless specified, we will use these transformed mortality rates to construct the mortality models and conduct the analysis.

1.2 Shock Detection

In this section, we investigate whether or not it is reasonable to apply an ARIMA model to fit the baseline mortality index. To achieve this, we first need to remove the shocks. Statistically, mortality shocks can be treated as outliers which are regarded as extreme values that standard time series model cannot handle; in other words, they are extreme values that violate the normal error assumption. The outlier detection technique from Cryer and Chan (2008) and Chen and Liu (1993) will be applied to detect shocks and specify the ARIMA model for baseline mortality rates.

Before fitting an ARIMA model to the baseline mortality rates, it would be easier to handle if this baseline process has a constant mean. Based on Figure 1.2, a simple linear regression model is applied to y_t to remove the decreasing tendency. The fitted linear regression model is $y_t = -0.01388t + 2.1285,\ t = 1, 2, \cdots, n$. The t-test for the slope parameter is not rejected and the R-square statistic is 0.9772, which is also significant[3]. The fitted values are shown in Figure 1.3[4], which also shows the graph of residuals for this regression model. The residuals, as expected, fluctuate around zero but include shocks.

As shown in Figure 1.3, the residuals include a huge shock in year 1918 as well as some small shocks. Cryer and Chan (2008) introduce an approach to detect outliers

[3]Using the R package: stats

[4]We compared using the original data q_t and the transformed data y_t. When q_t is used, a second-order polynomial model fits better, indicating that the decreasing tendency behaves as a curve more than a straight line (with R-square statistic $0.95 > 0.93$). The shape of this second-order polynomial graph also verifies our previous observation, that the slope of the decreasing tendency for mortality rate is steeper before 1942. While we realize that the second-order polynomial regression fits y_t well, it would be more complicated than a linear model. On the other hand, if y_t is used, the fitted model would be simpler and has an even larger R-square statistic of 0.9772.

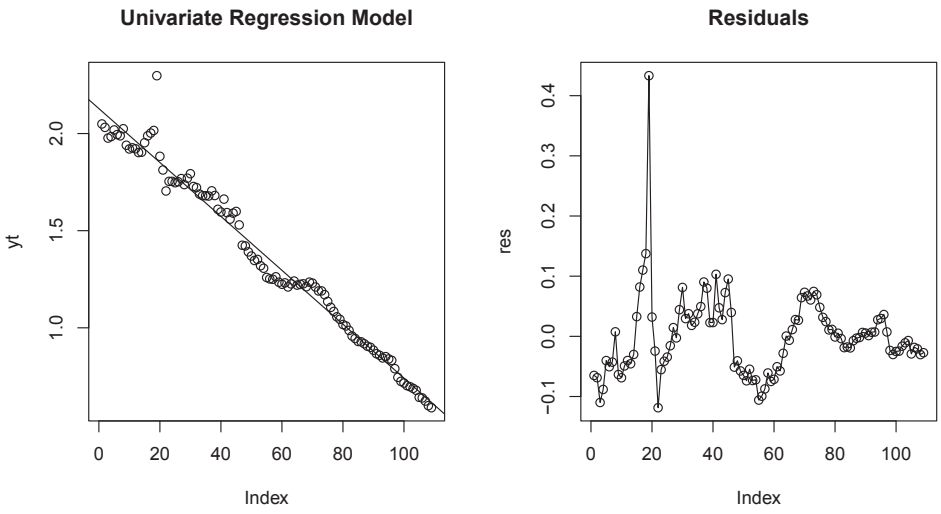

Figure 1.3: Univariate linear regression model: fitted value (left) and residuals (right)

and Chen and Liu (1993) provide an algorithm to get an outlier-free ARIMA model after removing their impacts. Li and Chan (2005, 2007) apply these techniques to study mortality rates in European and North American countries, respectively. Denote the residuals as r_t, which are our observations y_t after removing the decreasing tendency. Following Cryer and Chan (2008), two types of outliers are introduced: additive outliers (AO) and innovative outliers (IO).

Additive outliers behave as a one-time shock on r_t, that is

$$r_t = \begin{cases} r_t^0 + \delta_t^{AO} & \text{if AO occurs,} \\ r_t^0 & \text{otherwise,} \end{cases}$$

where r_t^0 is the baseline process for the residuals and δ_t^{AO} is the scale of the AO at time t.

Innovative outliers, on the other hand, are one-time shocks on ϵ_t, which is the

error term of r_t. Let δ_t^{IO} denote the scale of this IO at time t, hence

$$\epsilon_t = \begin{cases} \epsilon_t^0 + \delta_t^{IO} & \text{if IO occurs,} \\ \epsilon_t^0 & \text{otherwise,} \end{cases}$$

where ϵ_t^0 is the error term of r_t^0. The error term ϵ_t^0 should be independent and have an identical distribution ($i.i.d$).

While AO behaves as the transient effect, IO has lasting effects on r_t^0. For a stationary process, this effect would decay as time proceeds. Shocks in the mortality rates are probably additive outliers since these shocks are expected to be transient, such as a pandemic flu. Our main goal is to confirm the existence of shocks as well as the reasonability of using an ARIMA model to fit the baseline process. The difference between AO and IO is not as important as the fact that they are both outliers.

Now we briefly introduce the hypothesis tests for AO and IO (see Cryer and Chan, 2008). Consider the $AR(\infty)$ representation for the baseline process r_t^0,

$$\epsilon_t^0 = r_t^0 - \pi_1 r_{t-1}^0 - \pi_2 r_{t-2}^0 - \cdots .$$

Since we have observations r_t instead of r_t^0, the above model is replaced by

$$\epsilon_t = r_t - \pi_1 r_{t-1} - \pi_2 r_{t-2} - \cdots .$$

As mentioned above, while ϵ_t^0 and r_t^0 do not include shocks, ϵ_t and r_t may include shocks. Let λ_t^{IO} and λ_t^{AO} be the statistics for IO and AO at time t, respectively. λ_t^{IO} is the standardized innovative outlier effect statistic and it is defined as

$$\lambda_t^{IO} = \frac{\epsilon_t}{\sigma}, \tag{1.2}$$

where σ is the standard deviation of δ_t^{IO}. Similarly, λ_t^{AO} is the standardized additive outlier effect statistic defined as

$$\lambda_t^{AO} = \frac{-\rho^2 \sum_{k=t}^{n} \pi_{k-t} r_k}{\rho\sigma}, \tag{1.3}$$

where $\rho^2 = (\sum_{k=t}^{n} \pi_{k-t}^2)^{-1}$, $\pi_0 = -1$, and $\rho\sigma$ is the standard deviation of δ_t^{AO}. In both cases, σ can be robustly estimated by the mean absolute residual times $\sqrt{\frac{\pi}{2}}$ (Cryer and Chan, 2008). AO or IO would be detected as outliers whenever these two statistics are greater than the upper percentile with a certain significant level α. If λ_T^{IO} is significant at time T, the scale of this IO is $\delta_T^{IO} = \sigma\lambda_T^{IO}$; if λ_T^{AO} is significant, the scale of this AO is $\delta_T^{AO} = \rho\sigma\lambda_t^{AO}$. If λ_t^{IO} and λ_t^{AO} are both significant, choose the larger one and set r_t to be the corresponding type of outlier. For example, if both IOs and AOs are detected at time T and $\lambda_T^{AO} \geq \lambda_T^{IO}$, then r_T is set to be AO, and vice versa.

Basically, we want to find the best ARIMA(p, d, q) model where p is the order of AR process, q is the order of MA process, and d is the order of differentiation. In practice, a good model will not have large values for p, d, and q. While more parameters lead to a higher log-likelihood value, the AIC and BIC criteria are used to avoid parameter redundancy. They are given by

$$AIC = -2\ln L + 2k \qquad \text{and} \qquad BIC = -2\ln L + k\ln n, \tag{1.4}$$

where L is the maximized likelihood, k is the number of parameters, and n is the sample size.

Li and Chan (2005) introduce the following algorithm to find the outliers and the corresponding underlying model.

1. Initially, set $r_t^0 = r_t$. Choose the best model to determine certain p, d, and q without recognizing any outlier from a range of ARIMA models. The choice can be according to either the AIC or BIC criteria.

2. Apply the outlier detection test to calculate λ_t^{IO} and λ_t^{AO} using (1.2) and (1.3)

for all t.[5] AOs or IOs would be detected as outliers whenever these two statistics are greater than the upper percentile with a certain significant level α. If there is at least one detected outlier, determine the largest $|\lambda_t|$ and set it to be the corresponding type of outlier, then proceed to the next step. Otherwise go to Step 4.

3. If the new outlier at time T is an AO, then $r_T^0 = r_T - \delta_T^{AO}$ at time T. Use the updated r_t^0 to estimate parameters and calculate the AIC value. Similarly, if the new outlier is an IO, then $\epsilon_T^0 = \epsilon_T - \delta_T^{IO}$ at time T. With the updated r_t^0 and ϵ_t^0, fit the model with the same p, d, and q. Go to Step 2.

4. Similarly to Step 1, choose the best model to determine the updated p, d, and q with the updated r_t^0, according to the AIC or BIC criteria. The chosen criteria should be consistent with Step 1. If the updated parameters p, d and q remain the same, it means that the baseline ARIMA model has converged. Otherwise, go back to Step 2 to start a new iteration with the updated parameters.

As an illustration, we perform the outlier detection algorithm for the mortality index in terms of AIC criteria. The outlier detection process using the BIC criterion could be performed similarly.

1. Define r_t as residuals in Figure 1.3. Set $r_t^0 = r_t$ and $\alpha = 5\%$.

2. Initially, we selected models from a range of ARIMA models with maximal order[6] $p_{max} = 5$, $d_{max} = 2$ and $q_{max} = 5$. With no identified outliers and initial y_t (the residuals in Figure 1.3), the best model was ARIMA(3,0,1) with $AIC = -336.95$.

[5] Using the R package: TSA.

[6] The maximum bound can be adjusted if the optimization result reaches this bound.

3. Then we detected the outliers with this model, showing that there are possible AOs and IOs both at time 1918 and 1919, with $\lambda_{1918}^{AO} = 12.02$, $\lambda_{1919}^{AO} = -4.12$, $\lambda_{1918}^{IO} = 10.1$, and $\lambda_{1919}^{IO} = -7.21$. We chose the largest $|\lambda|$, therefore $\lambda_T^{AO} = 12.02$ at time 1918 was selected. That is to say, r_{1918} is identified as an AO. We calculated $\delta_{1918}^{AO} = 0.3219$, then updated the data r_{1918} to remove the outlier effect: $r_{1918}^0 = r_{1918} - \delta_{1918}^{AO}$. With this updated data and the ARIMA(3,0,1) model, the resulting AIC is -460.92. As we can see, the AIC value is improved significantly after identifying the AO at $t = 1918$.

4. We performed again the outlier detection in Step 3. This time, r_{1921} and r_{1940} were detected as possible AOs. We chose the largest $|\lambda_T|$ which was $\lambda_{1921} = |-4.359|$ at $t = 1921$. With $\delta_{1921}^{AO} = -0.0738$, we then updated the data r_{1921} to remove the outlier effect: $r_{1921}^0 = r_{1921} - \delta_{1921}^{AO}$. Without the outliers in year 1918 and 1921, the ARIMA(3,0,1) model was improved with an $AIC = -478.52$.

5. We repeated Step 3. This time, r_{1940} was detected as a possible AO. No IOs were detected. Since there is only one possible outlier, we chose it to be the AO. We calculated $\delta_{1940}^{AO} = 0.055$, then updated the data r_{1940} to remove the outlier effect: $r_{1940}^0 = r_{1940} - \delta_{1940}^{AO}$. Without the outliers in years 1918, 1921 and 1940, the ARIMA(3,0,1) model was improved, with an $AIC = -491.48$.

6. We again repeated Step 3. r_{1907} was detected as a possible AO and r_{1908} was detected as a possible IO. The largest $|\lambda_T|$ was $|\lambda_{1907}^{AO}| = 3.8929$. With $\delta_{1907}^{AO} = 0.05465$, we updated the data r_{1907} to remove the outlier effect: $r_{1907}^0 = r_{1907} - \delta_{1907}^{AO}$. Without the outliers in years 1918, 1921, 1940 and 1907, the ARIMA(3,0,1) model improved to an $AIC = -505.98$.

7. No further outliers were detected in repeating Step 3.

8. Then we re-selected models from a range of ARIMA models by minimizing the AIC. The fitted model changed to be ARIMA(1,0,4) with an improved $AIC = -514.22$.

9. We again repeated Step 3. No AOs were detected. r_{1946} was detected as the only possible IO with $|\lambda^{IO}_{1946}| = -3.582$ and $\delta^{IO}_{1946} = -0.07327$. Updating r^0_t, we fitted the ARIMA(1,0,4) model taking the IO into account. The model was improved with the $AIC = -523.83$.

10. Step 3 did not detect any further AO or IO.

11. With the outlier-free r^0_t (four AOs and one IO), we re-selected from a range of ARIMA models using the AIC criterion. The best model was ARIMA(3,0,3) with $AIC = -525.59$ which improved that of the ARIMA(1,0,4).

12. Step 3 did not detect any further AO or IO.

13. We re-selected the model once again and the best model converged at ARIMA(3,0,3). This terminated the outlier detection procedure. Therefore, the final model is $ARIMA(3,0,3) + AO(1918, 1921, 1940, 1907) + IO(1946)$.

We verify this final model by comparing the ACF, PACF graphs, and QQ-plots for the residuals of the ARIMA(3,0,3) model, after removal of the outliers and the residuals for the ARIMA(3,0,1) model before the outlier detection. As shown in Figure 1.4, the residuals do not follow a normal distribution and there are autocorrelations and partial correlations before the outlier detection. Therefore, after the outlier detection, the baseline model follows an ARIMA model with $i.i.d$ normal errors. That is to say, the model $ARIMA(3,0,3) + AO(1918, 1921, 1940, 1907) +$

$IO(1946)$ not only successfully detects outliers, but also indicates that an ARIMA model can be fitted to the baseline mortality rates.

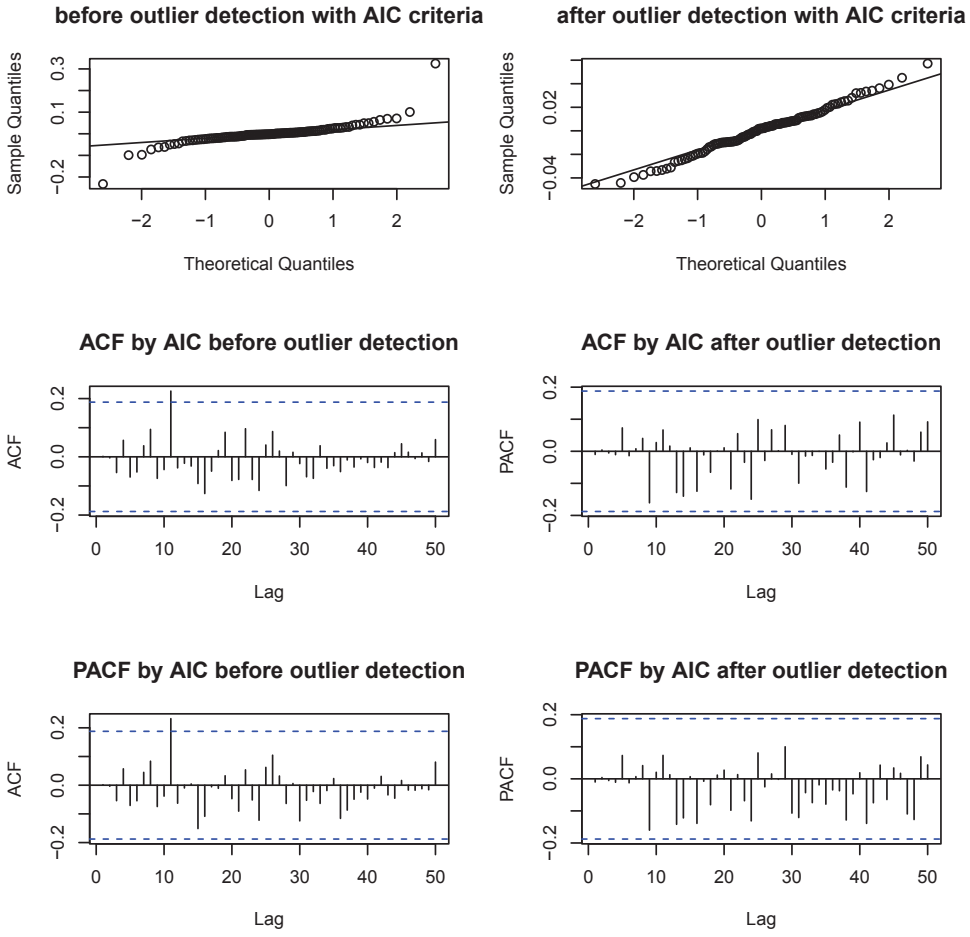

Figure 1.4: QQ-plot (upper), ACF (middle) and PACF (lower) of residuals for ARIMA(3,0,1) and ARIMA(3,0,3) + AO(1918,1921,1940,1907) + IO(1946)

Similarly, another model was determined using the BIC criterion instead of the AIC, following the above steps. The best model according to the BIC criterion is $ARIMA(2,0,0)+AO(1918,1921,1940,1907)+IO(1946)$ with a $BIC = -506.49$. As we can see, the BIC imposes a heavier penalty than AIC on parameter redundancy. Similarly, Figure 1.5 compares the residuals of this ARIMA(2,0,0) model and the

ARIMA(1,0,0) model before outlier detection. Again, the model chosen after outlier detection constitutes an improved baseline ARIMA model.

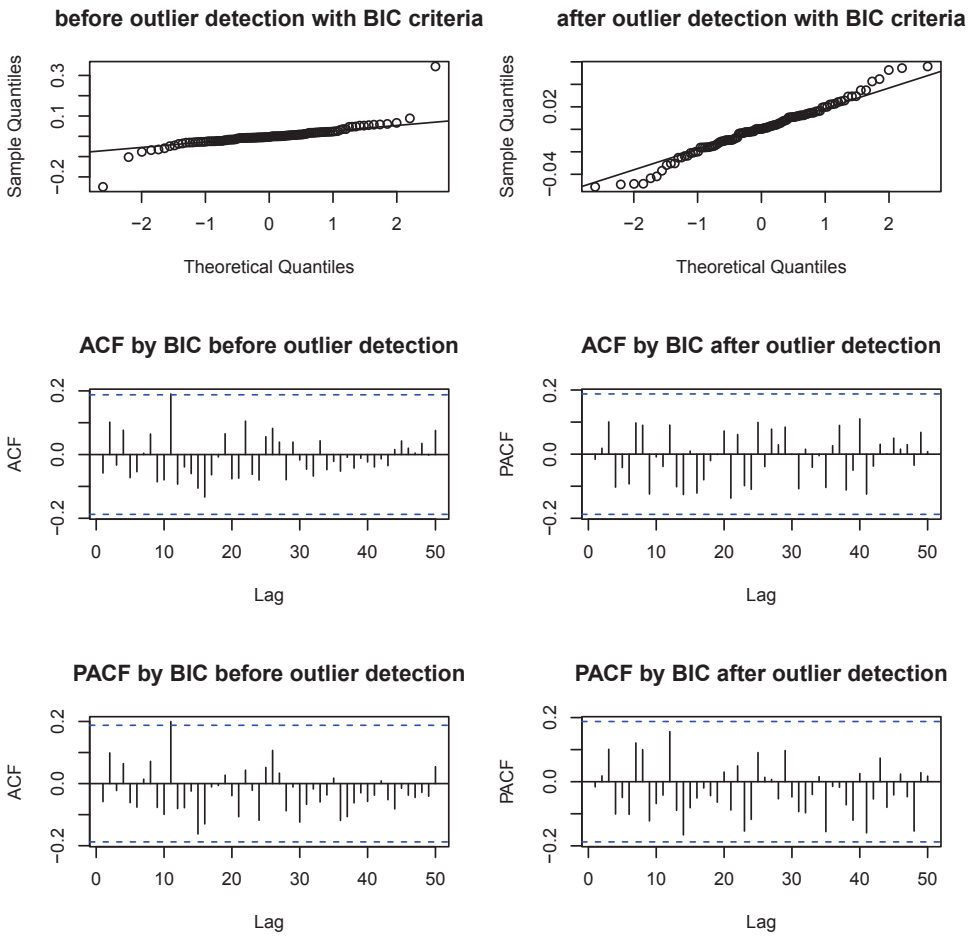

Figure 1.5: QQ-plot (upper), ACF (middle) and PACF (lower) of residuals for ARIMA(1,0,0) and ARIMA(2,0,0) + AO(1918,1921,1940,1907) + IO(1946)

Moreover, the Shapiro-Wilk normality test indicates that the p-values for the above two models after outlier detection, chosen with the AIC and BIC criteria are 0.8371 and 0.2377. While our two models pass the normality test, the model residuals by AIC are closer to a normal distribution.

Here we provide another reason to use transformed y_t values instead of q_t. This

data transformation can reduce the variation gap. If q_t is used instead, we would not get normal residuals here even if the same outlier detection procedure was applied, because larger volatility would lead to fatter tails than normal.

The above diagnostic tests indicate that both of these two outlier detection models provide a good fit. Although they indicate different baseline ARIMA models, ARIMA(3,0,3) and ARIMA(2,0,0), due to the different model selection criteria, they both detect the same five outliers, i.e. AO(1918,1921,1940,1907) and IO(1946). We summarize the detected shock information in Table 1.3. The shock effects are similar between the two criteria. We also find that all five shocks have similar shock effects ranging from 0.05 to 0.08 (in terms of absolute values), except for the extremal positive shock in 1918, which is about five times of other shock effects and was caused by a devastating pandemic (see Morens et al., 2010). Figure 1.6 visually points out these five outliers on the plot of q_t. Looking back at the history, some reasons could explain these shocks. For instance, 1907 was a peak year in European immigration to the U.S. and the positive shock in 1940 may be caused by World War II, while medical improvement and better protection of health led to mortality decreases in 1946 (Li and Chan, 2005).

Table 1.3: Detected shocks with AIC and BIC criterion

Year T	1907	1918	1921	1940	1946
AO/IO	AO	AO	AO	AO	IO
Positive/Negative Shocks	+	+	−	+	−
Shock Effects $\delta_T^{AO/IO}$ with AIC	0.05465	0.3219	−0.0738	0.055	−0.07327
Shock Effects $\delta_T^{AO/IO}$ with BIC	0.05872	0.356	−0.07922	0.0684	−0.08466

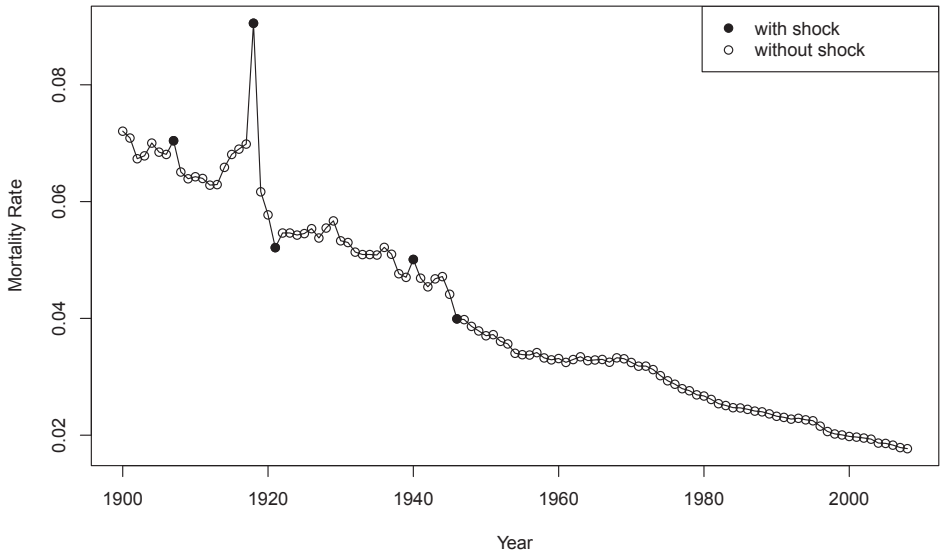

Figure 1.6: Plot of q_t with five detected outliers

1.3 Remarks

We compared the results between the AIC and BIC criteria. On the one hand, these two models detect the same outlier occurrences, though the scale and the detection order of outliers are not the same. On the other hand, the baseline ARIMA models obtained by these two criteria are different. This implies that the baseline model searched by the outlier detection procedure is not unique; accordingly, the outlier effect and detection order may not be unique either. It depends on the model selection criteria that we use. However, good models by certain criteria should detect similar outliers and finally pass all the diagnostic tests.

As mentioned earlier in this section, the outlier detection process described here is a simplified version. Chen and Liu (1993) introduce four types of outliers and also consider the joint effect for multiple outliers using multiple regression model. In

Chen and Liu's model, the baseline process may be more consistent among different model selection criteria, because differences of outlier effects and the detection order would be reduced if the joint effect is considered. Li and Chan (2005) study the mortality rates of the United Kingdom and Scandinavian countries. A similar study in North America is introduced in Li and Chan (2007). While the huge shock in 1918 is detected in both of these two papers, other detected shocks are somehow different mainly because the data on which their results are based is different.

In conclusion, we show the existence of shocks and the feasibility in using an ARIMA process to model the baseline mortality rate, where two models are provided to remove both the decreasing tendency and outlier effects. This important conclusion is used in our mortality index analysis in Chapter 3.

Chapter 2

State Space Models

This chapter introduces the state space model (SSM) which is a branch of time series analysis. The state space model is a system of equations that includes an observation equation as well as a state equation and they each contain a stochastic error term. The SSM can be constructed using linear or non-linear equations. On the one hand, the non-linear state space model is developed due to the improvement of simulation techniques. On the other hand, the linear state space model is still being used more often since they are easier to program and interpret. We will focus on linear SSMs in our study. Filtering and smoothing are required because of the existence of latent variables and two error terms. An intuitive derivation of filtering and smoothing techniques using conditional distributions is introduced by Kitagawa (1987). With Gaussian errors, the efficient Kalman filter and smoother (see Kalman, 1960) can be used. Simulation techniques, such as importance sampling, are required in the case of non-Gaussian SSMs. Durbin and Koopman (1997) propose an algorithm to deal with non-Gaussian SSMs. It is based on the Kalman filter and smoother as well as the importance sampling technique. The simulation smoother is required to calculate the likelihood function and expectations (see Section 2.5 for definitions). Durbin and

Koopman (2002) propose a simple and efficient algorithm for simulation smoother. Finally, diagnostic tests will be discussed.

2.1 The Linear State Space Model

This section introduces the linear state space model, which is given by

$$y_t = Z_t \boldsymbol{x}_t + \varepsilon_t, \tag{2.1}$$

$$\boldsymbol{x}_{t+1} = T_t \boldsymbol{x}_t + R_t \boldsymbol{\eta}_t, \tag{2.2}$$

where (2.1) is called the *observation equation*, (2.2) the *state equation*, y_t are the observations, $\boldsymbol{x}_t$ is the latent variable vector, and ε_t and $\boldsymbol{\eta}_t$ are error terms. Table 2.1 summarizes the dimensionality of variables in (2.1) and (2.2) given that y_t is 1×1. The state equation (2.2) connects two consecutive latent variables $\boldsymbol{x}_t$ and the observation equation (2.1) links the latent variables $\boldsymbol{x}_t$ with the observations y_t. The dimension of $\boldsymbol{x}_t$ accounts for the complexity of the state space model because it represents the inner connection of the equations. While $Z_t \boldsymbol{x}_t$ and $T_t \boldsymbol{x}_t$ represent the deterministic parts of the system, the error terms represent the stochastic part.

Table 2.1: The dimension of variables in linear SSM

Variable	y_t	$\boldsymbol{x}_t$	Z_t	T_t	R_t	ε_t	$\boldsymbol{\eta}_t$
Dimensions	1×1	$m \times 1$	$1 \times m$	$m \times m$	$m \times r$	1×1	$r \times 1$

2.2 Filtering and Smoothing: the Conditional Approach

The observation and state equations contain deterministic and stochastic parts. At each time t, the conditional probability density function (pdf) could be calculated given the distribution of ε_t and $\boldsymbol{\eta}_t$. However, $\boldsymbol{x}_t$ is not observable, so quantities like $\mathrm{E}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ and $\mathrm{E}(y_t|\boldsymbol{x}_t)$ cannot be calculated. In fact, we are only able to calculate expectations that are conditional on Y_t, such as $\mathrm{E}(y_{t+1}|Y_t)$, $\mathrm{E}(\boldsymbol{x}_{t+1}|Y_t)$, and $\mathrm{E}(\boldsymbol{x}_t|Y_n)$, where $Y_t = \{y_1, y_2, \cdots, y_t\}$. To obtain these probabilities, filtering and smoothing are necessary.

This section briefly shows the work by Kitagawa (1987), which provides a straightforward way to understand SSM filtering and smoothing. To simplify the problem, we suppose that

$$\begin{aligned} y_t &= Z\, x_t + \varepsilon_t, \\ x_{t+1} &= T\, x_t + R\, \eta_t, \end{aligned} \tag{2.3}$$

where matrices Z, T, and R are time invariant, y_t and x_t are both one-dimensional.

2.2.1 Filtering

Filtering is a forward process that obtains the probability density function $p(x_t|Y_t)$ for $t = 0, 1, \cdots, n$. This process starts from $p(x_0)$, which is the initial distribution of x_0, to $p(x_n|Y_n)$, which is the last step of filtering. Figure 2.1 shows the filtering process.

Suppose that we have done the filtering process up to time $t-1$, saying that $p(x_{t-1}|Y_{t-1})$ is known. As shown in Figure 2.1, the conditional pdf of $x_t|Y_{t-1}$ needs

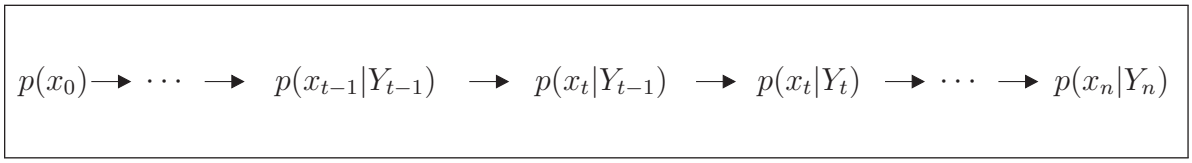

Figure 2.1: Filtering process

to be defined in order to find $p(x_t|Y_t)$. That is

$$p(x_t|Y_{t-1}) = \int p(x_t, x_{t-1}|Y_{t-1})dx_{t-1}.$$

Using Bayes theorem,

$$p(x_t, x_{t-1}|Y_{t-1}) = p(x_t|x_{t-1}, Y_{t-1})p(x_{t-1}|Y_{t-1}),$$

where $p(x_{t-1}|Y_{t-1})$ is known. Since x_t could be derived given x_{t-1} using (2.2), we have $p(x_t|x_{t-1}, Y_{t-1}) = p(x_t|x_{t-1})$. Therefore,

$$p(x_t|Y_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})dx_{t-1}. \tag{2.4}$$

The probability density function of the one-step prediction is

$$p(y_t|Y_{t-1}) = \int p(x_t, y_t|Y_{t-1})dx_t = \int p(y_t|x_t, Y_{t-1})p(x_t|Y_{t-1})dx_t.$$

Note that $p(y_t|x_t, Y_{t-1}) = p(y_t|x_t)$ by (2.1), then

$$p(y_t|Y_{t-1}) = \int p(y_t|x_t)p(x_t|Y_{t-1})dx_t, \tag{2.5}$$

where $p(x_t|Y_{t-1})$ is given by (2.4) and $p(y_t|x_t)$ is obtained using the observation equation. Note that the conditional pdf of $y_t|Y_{t-1}$ given in (2.5) is used to derive the likelihood function. The conditional pdf of $x_t|Y_t$ is given using (2.5) by

$$p(x_t|Y_t) = p(x_t|y_t, Y_{t-1}) = \frac{p(x_t, y_t|Y_{t-1})}{p(y_t|Y_{t-1})} = \frac{p(y_t|x_t)p(x_t|Y_{t-1})}{\int p(y_t|x_t)p(x_t|Y_{t-1})dx_t}. \tag{2.6}$$

The distribution of $x_n|Y_n$ could be obtained using (2.4), (2.5), and (2.6) iteratively.

2.2.2 Smoothing

The smoothing is a backward process used to find the distribution of $x_t|Y_n$. It starts with $p(x_n|Y_n)$ which is given by filtering, and finishes with $p(x_1|Y_n)$. Suppose that the probability density function $p(x_{t+1}|Y_n)$ is given, the pdf $p(x_t|Y_n)$ by smoothing is obtained as follows,

$$\begin{aligned} p(x_t, x_{t+1}|Y_n) &= p(x_{t+1}|Y_n)p(x_t|x_{t+1}, Y_n) = p(x_{t+1}|Y_n)p(x_t|x_{t+1}, Y_t) \\ &= \frac{p(x_{t+1}|Y_n)p(x_t, x_{t+1}|Y_t)}{p(x_{t+1}|Y_t)} = \frac{p(x_{t+1}|Y_n)p(x_{t+1}|x_t)p(x_t|Y_t)}{p(x_{t+1}|Y_t)}, \end{aligned} \tag{2.7}$$

where $p(x_{t+1}|x_t)$ is given by the state equation, $p(x_t|Y_t)$ and $p(x_{t+1}|Y_t)$ are obtained from (2.6) and (2.4), respectively. Using (2.7) leads to

$$p(x_t|Y_n) = \int p(x_t, x_{t+1}|Y_n)dx_{t+1} = \int \frac{p(x_{t+1}|Y_n)p(x_{t+1}|x_t)p(x_t|Y_t)}{p(x_{t+1}|Y_t)}dx_{t+1}. \tag{2.8}$$

With the above filtering and smoothing results, some quantities can be calculated. For example, using the filtering probability function $p(x_t|Y_t)$, the smoothing probability function $p(x_t|Y_n)$, and the one-step prediction probability function $p(y_{t+1}|Y_t)$, the expected filtering and smoothing values are given by,

$$\begin{aligned} E(y_{t+1}|Y_t) &= \int y_{t+1}p(y_{t+1}|Y_t)dy_{t+1}, \\ E(x_t|Y_t) &= \int x_t p(x_t|Y_t)dx_t, \\ E(x_t|Y_n) &= \int x_t p(x_t|Y_n)dx_t. \end{aligned} \tag{2.9}$$

For another example, using (2.5), the likelihood function is

$$L(\psi|Y_n) = p(Y_n|\psi) = \prod_{t=1}^{n} p(y_t|Y_{t-1}, \psi), \tag{2.10}$$

where ψ is a set of parameters of the state space model.

In general, the filtering probability function $p(x_t|Y_t)$ and smoothing probability function $p(x_t|Y_n)$ are hard to calculate since they include solving multiple integrals.

Multidimensional variables will also add more complexity to this problem. The conditional probability approach could be used when distributions of error terms are conjugate. For non-conjugate distributions, it is unrealistic to proceed by definition. An efficient algorithm is necessary to apply to state space models, such as the important sampling or Gibbs sampling.

2.3 State Space Model with Normal Error Terms

In this section, two different approaches are introduced to process the filtering and smoothing iterations. The first approach calculates filtering and smoothing distribution functions using (2.4), (2.5), (2.6), and (2.8). The second approach capitalizes on the fact that normal distributions are conjugate. This approach is known as Kalman filter and smoother, which is much more computationally efficient.

To simplify the notations, we use the following,

$$\begin{aligned} &\hat{\boldsymbol{x}}_t = E(\boldsymbol{x}_t|Y_{t-1}), && \bar{\boldsymbol{x}}_t = E(\boldsymbol{x}_t|Y_t), && \tilde{\boldsymbol{x}}_t = E(\boldsymbol{x}_t|Y_n), \\ &\hat{V}_t = Var(\boldsymbol{x}_t|Y_{t-1}), && \bar{V}_t = Var(\boldsymbol{x}_t|Y_t), && \tilde{V}_t = Var(\boldsymbol{x}_t|Y_n). \end{aligned} \tag{2.11}$$

2.3.1 Filtering and Smoothing with the Conditional Approach

The conditional pdf is obtained from one-dimension x_t and y_t. That is

$$y_t|x_t \sim N(x_t, \sigma_\varepsilon^2), \qquad x_{t+1}|x_t \sim N(x_t, \sigma_\eta^2). \tag{2.12}$$

Now given

$$x_{t-1}|Y_{t-1} \sim N(\bar{x}_{t-1}, \bar{V}_{t-1}), \tag{2.13}$$

we derive the probability density function $p(x_t|Y_{t-1})$ using (2.4), (2.12), and (2.13).

$$\begin{aligned}
p(x_t|Y_{t-1}) &= \int_{-\infty}^{\infty} p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})dx_{t-1} \\
&\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{(x_t-x_{t-1})^2}{\sigma_\eta^2}+\frac{(x_{t-1}-\bar{x}_{t-1})^2}{\bar{V}_{t-1}}\right]}dx_{t-1} \\
&\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_\eta^2}+\frac{1}{\bar{V}_{t-1}}\right)\left(x_{t-1}-\frac{\bar{V}_{t-1}x_t+\sigma_\eta^2\bar{x}_{t-1}}{\sigma_\eta^2+\bar{V}_{t-1}}\right)^2+\frac{x_t^2}{\sigma_\eta^2}+\frac{\bar{x}_{t-1}^2}{\bar{V}_{t-1}}-\left(\frac{1}{\sigma_\eta^2}+\frac{1}{\bar{V}_{t-1}}\right)\left(\frac{\bar{V}_{t-1}x_t+\sigma_\eta^2\bar{x}_{t-1}}{\sigma_\eta^2+\bar{V}_{t-1}}\right)^2\right]}dx_{t-1} \\
&\propto e^{-\frac{1}{2}\left[\frac{x_t^2}{\sigma_\eta^2}+\frac{\bar{x}_{t-1}^2}{\bar{V}_{t-1}}-\left(\frac{1}{\sigma_\eta^2}+\frac{1}{\bar{V}_{t-1}}\right)\left(\frac{\bar{V}_{t-1}x_t+\sigma_\eta^2\bar{x}_{t-1}}{\sigma_\eta^2+\bar{V}_{t-1}}\right)^2\right]} \\
&\propto e^{-\frac{1}{2}\frac{(x_t-\bar{x}_{t-1})^2}{\sigma_\eta^2+\bar{V}_{t-1}}}.
\end{aligned}$$

Hence

$$x_t|Y_{t-1} \sim N(\hat{x}_t, \hat{V}_t), \tag{2.14}$$

where

$$\hat{x}_t = \bar{x}_{t-1}, \qquad \hat{V}_t = \bar{V}_{t-1} + \sigma_\eta^2. \tag{2.15}$$

Note that $x_t|Y_{t-1}$ has the same mean as $x_{t-1}|Y_{t-1}$ but a larger variance. Using (2.6) and (2.14), we have

$$\begin{aligned}
p(x_t|Y_t) &= \frac{p(y_t|x_t)p(x_t|Y_{t-1})}{\int p(y_t|x_t)p(x_t|Y_{t-1})dx_t} \\
&\propto p(y_t|x_t)p(x_t|Y_{t-1}) \\
&\propto e^{-\frac{1}{2}\left[\frac{(y_t-x_t)^2}{\sigma_\varepsilon^2}+\frac{(x_t-\hat{x}_t)^2}{\hat{V}_t}\right]} \\
&\propto e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_\varepsilon^2}+\frac{1}{\hat{V}_t}\right)\left(x_t-\frac{\hat{V}_t y_t+\sigma_\varepsilon^2\hat{x}_t}{\hat{V}_t+\sigma_\varepsilon^2}\right)^2\right]}.
\end{aligned}$$

Therefore,

$$x_t|Y_t \sim N(\bar{x}_t, \bar{V}_t),$$

where

$$\bar{x}_t = \frac{\hat{V}_t y_t + \sigma_\varepsilon^2 \hat{x}_t}{\hat{V}_t + \sigma_\varepsilon^2}, \qquad \bar{V}_t = \frac{\hat{V}_t \sigma_\varepsilon^2}{\hat{V}_t + \sigma_\varepsilon^2}.$$

Replacing $\hat{x}_t$ and $\hat{V}_t$ by (2.15), we will get the relationship between two consecutive filtering values,

$$\bar{x}_t = \frac{(\bar{V}_{t-1} + \sigma_\eta^2) y_t + \sigma_\varepsilon^2 \bar{x}_{t-1}}{\sigma_\eta^2 + \bar{V}_{t-1} + \sigma_\varepsilon^2}, \qquad \bar{V}_t = \frac{(\sigma_\eta^2 + \bar{V}_{t-1})\sigma_\varepsilon^2}{\sigma_\eta^2 + \bar{V}_{t-1} + \sigma_\varepsilon^2}. \tag{2.16}$$

As we can see, the filtering value at time t is a weighted average of the filtering value of previous period and the observation at time t.

Next, we deal with the smoothing, which starts with $p(x_{t+1}|Y_n)$, where

$$x_{t+1}|Y_n \sim N(\tilde{x}_{t+1}, \tilde{V}_{t+1}).$$

The smoothing iteration given by $p(x_t|Y_n)$ is obtained by using (2.8) and (2.12).

$$\begin{aligned}
p(x_t|Y_n) &= \int \frac{p(x_{t+1}|Y_n)p(x_{t+1}|x_t)p(x_t|Y_t)}{p(x_{t+1}|Y_t)} dx_{t+1} \\
&\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{(x_{t+1}-\tilde{x}_{t+1})^2}{\tilde{V}_{t+1}} + \frac{(x_{t+1}-x_t)^2}{\sigma_\eta^2} + \frac{(x_t-\bar{x}_t)^2}{\bar{V}_t} - \frac{(x_{t+1}-\hat{x}_{t+1})^2}{\hat{V}_{t+1}}\right]} dx_{t+1} \\
&\propto e^{-\frac{1}{2}\left[\frac{(x_t-\bar{x}_t)^2}{\bar{V}_t} + \frac{x_t^2}{\sigma_\eta^2}\right]} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\left(\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}\right)\left(x_{t+1} - \frac{\frac{\tilde{x}_{t+1}}{\tilde{V}_{t+1}} + \frac{x_t}{\sigma_\eta^2} - \frac{\hat{x}_{t+1}}{\hat{V}_{t+1}}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}}\right)^2 - \frac{\left(\frac{\tilde{x}_{t+1}}{\tilde{V}_{t+1}} + \frac{x_t}{\sigma_\eta^2} - \frac{\hat{x}_{t+1}}{\hat{V}_{t+1}}\right)^2}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}}\right]} dx_{t+1} \\
&\propto e^{-\frac{1}{2}\left[\frac{(x_t-\bar{x}_t)^2}{\bar{V}_t} + \frac{x_t^2}{\sigma_\eta^2} - \frac{\left(\frac{\tilde{x}_{t+1}}{\tilde{V}_{t+1}} + \frac{x_t}{\sigma_\eta^2} - \frac{\hat{x}_{t+1}}{\hat{V}_{t+1}}\right)^2}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}}\right]} \\
&\propto e^{-\frac{1}{2}\left[\left(\frac{1}{\bar{V}_t} + \frac{1}{\sigma_\eta^2} - \frac{\frac{1}{(\sigma_\eta^2)^2}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}}\right) x_t^2 - 2\left(\frac{\bar{x}_t}{\bar{V}_t} + \frac{\frac{\tilde{x}_{t+1}}{\tilde{V}_{t+1}\sigma_\eta^2} - \frac{\hat{x}_{t+1}}{\hat{V}_{t+1}\sigma_\eta^2}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}}\right) x_t\right]}.
\end{aligned} \tag{2.17}$$

Using $\hat{x}_{t+1} = \bar{x}_t$ and $\hat{V}_{t+1} = \bar{V}_t + \sigma_\eta^2$, we can simplify the two terms before x_t^2 and x_t in (2.17), that is

$$\begin{aligned}
\frac{1}{\bar{V}_t} + \frac{1}{\sigma_\eta^2} - \frac{\frac{1}{(\sigma_\eta^2)^2}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}} &= \frac{\left(\frac{1}{\bar{V}_t} + \frac{1}{\sigma_\eta^2}\right)\left(\frac{1}{\tilde{V}_{t+1}} - \frac{1}{\hat{V}_{t+1}}\right) + \frac{1}{\bar{V}_t\sigma_\eta^2}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}} \\
&= \frac{\frac{1}{\bar{V}_t\sigma_\eta^2} \cdot \frac{\hat{V}_{t+1}}{\tilde{V}_{t+1}}}{\frac{1}{\tilde{V}_{t+1}} + \frac{1}{\sigma_\eta^2} - \frac{1}{\hat{V}_{t+1}}} \\
&= \frac{(\hat{V}_{t+1})^2}{(\sigma_\eta^2\hat{V}_{t+1} + \bar{V}_t\tilde{V}_{t+1})\bar{V}_t},
\end{aligned} \tag{2.18}$$

and

$$\begin{aligned}\frac{\bar{x}_t}{\bar{V}_t}+\frac{\frac{\tilde{x}_{t+1}}{\tilde{V}_{t+1}\sigma_\eta^2}-\frac{\hat{x}_{t+1}}{\hat{V}_{t+1}\sigma_\eta^2}}{\frac{1}{\tilde{V}_{t+1}}+\frac{1}{\sigma_\eta^2}-\frac{1}{\hat{V}_{t+1}}} &= \frac{\bar{x}_t}{\bar{V}_t}+\frac{\hat{V}_{t+1}\tilde{x}_{t+1}-\tilde{V}_{t+1}\hat{x}_{t+1}}{\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1}} \\ &= \frac{(\sigma_\eta^2\bar{x}_t+\bar{V}_t\tilde{x}_{t+1})\hat{V}_{t+1}}{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}.\end{aligned} \tag{2.19}$$

Now substituting (2.18) and (2.19) into (2.17) leads to

$$\begin{aligned}p(x_t|Y_n) &\propto e^{-\frac{1}{2}\left[\left(\frac{(\hat{V}_{t+1})^2}{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}\right)x_t^2-2\left(\frac{(\sigma_\eta^2\bar{x}_t+\bar{V}_t\tilde{x}_{t+1})\hat{V}_{t+1}}{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}\right)x_t\right]} \\ &\propto e^{-\frac{1}{2}\left(\frac{(\hat{V}_{t+1})^2}{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}\right)\left(x_t-\frac{\sigma_\eta^2\bar{x}_t+\bar{V}_t\tilde{x}_{t+1}}{\hat{V}_{t+1}}\right)^2} \\ &\propto e^{-\frac{1}{2}\left(\frac{(\hat{V}_{t+1})^2}{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}\right)\left(x_t-\frac{\sigma_\eta^2\bar{x}_t+\bar{V}_t\tilde{x}_{t+1}}{\sigma_\eta^2+\bar{V}_t}\right)^2}.\end{aligned}$$

Hence,

$$x_t|Y_n \sim N(\tilde{x}_t, \tilde{V}_t),$$

where

$$\tilde{x}_t=\frac{\sigma_\eta^2\bar{x}_t+\bar{V}_t\tilde{x}_{t+1}}{\sigma_\eta^2+\bar{V}_t}, \qquad \tilde{V}_t=\frac{(\sigma_\eta^2\hat{V}_{t+1}+\bar{V}_t\tilde{V}_{t+1})\bar{V}_t}{(\hat{V}_{t+1})^2}. \tag{2.20}$$

The smoothing iteration (2.20) is also a weighted average for $\tilde{x}_t$, averaging between $\bar{x}_t$ and $\tilde{x}_{t+1}$.

2.3.2 Kalman Filter and Smoother

We now generalize the above filtering and smoothing results for a multidimensional state space model capitalizing on the Gaussian distribution. Consider the following state space model where $\boldsymbol{x}_t$ is multidimensional,

$$\begin{aligned}y_t &= Z_t\,\boldsymbol{x}_t+\varepsilon_t, & \varepsilon_t &\sim N(0,\sigma_\varepsilon^2), \\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t+R_t\,\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(0,Q_t),\end{aligned} \tag{2.21}$$

where Z_t, T_t, R_t, Q_t are therefore matrices, y_t is a one-dimensional vector, the covariance matrix of $\boldsymbol{\eta}_t$ is Q_t, and the initial condition for $\boldsymbol{x}_t$ is that $\boldsymbol{x}_1 \sim N(\hat{\boldsymbol{x}}_1, \hat{V}_1)$.

Durbin and Koopman (2001) develop the Kalman filter and smoother by calculating the expectation and variance of the normal distributions instead of probability density functions as in Section 2.3.1.

We first introduce the filtering step, which aims to obtain the distribution of $x_{t+1}|Y_t$ from the distribution of $x_t|Y_{t-1}$. Because of Gaussian error terms, it can be achieved by iterating from $\hat{\boldsymbol{x}}_t$ and $\hat{V}_t$ to $\hat{\boldsymbol{x}}_{t+1}$ and $\hat{V}_{t+1}$, respectively. According to (2.21),

$$\hat{\boldsymbol{x}}_{t+1} = \mathrm{E}(T_t\boldsymbol{x}_t + R_t\boldsymbol{\eta}_t|Y_t) = T_t\mathrm{E}(\boldsymbol{x}_t|Y_t) = T_t\bar{\boldsymbol{x}}_t, \tag{2.22}$$

and

$$\hat{V}_{t+1} = \mathrm{Var}(T_t\boldsymbol{x}_t + R_t\boldsymbol{\eta}_t|Y_t) = T_t\bar{V}_tT_t' + R_tQ_tR_t'. \tag{2.23}$$

Define the one-step prediction error as $v_t = y_t - Z_t\hat{\boldsymbol{x}}_t$. By the Regression Lemma (Durbin and Koopman, 2001), the filtering values $\bar{\boldsymbol{x}}_t$ and $\bar{V}_t$ are given by

$$\bar{\boldsymbol{x}}_t = \mathrm{E}(\boldsymbol{x}_t|Y_t) = \mathrm{E}(\boldsymbol{x}_t|Y_{t-1}, v_t) = \hat{\boldsymbol{x}}_t + \hat{V}_tZ_t'F_t^{-1}v_t, \tag{2.24}$$

and

$$\bar{V}_t = \mathrm{Var}(\boldsymbol{x}_t|Y_t) = \mathrm{Var}(\boldsymbol{x}_t|Y_{t-1}, v_t) = \hat{V}_t + \hat{V}_tZ_t'F_t^{-1}Z_t\hat{V}_t, \tag{2.25}$$

where $F_t = \mathrm{Var}(v_t) = Z_t\hat{V}_tZ_t' + \sigma_\varepsilon^2$. Substitute (2.24) into (2.22),

$$\hat{\boldsymbol{x}}_{t+1} = T_t(\hat{\boldsymbol{x}}_t + \hat{V}_tZ_t'F_t^{-1}v_t) = T_t\hat{\boldsymbol{x}}_t + K_tv_t, \tag{2.26}$$

where $K_t = T_t\hat{V}_tZ_t'F_t^{-1}$. Using (2.25), (2.23) becomes

$$\hat{V}_{t+1} = T_t(\hat{V}_t + \hat{V}_tZ_t'F_t^{-1}Z_t\hat{V}_t)T_t' + R_tQ_tR_t' = T_t\hat{V}_tL_t' + R_tQ_tR_t', \tag{2.27}$$

where $L_t = T_t - K_tZ_t$.

Therefore, as we can see in (2.26) and (2.27), the Kalman filter forwardly calculates $\hat{\boldsymbol{x}}_t$ and $\hat{V}_t$, $t = 1, 2, \cdots, n$.

In the smoothing iterations, we aim to obtain the smoothing values $\tilde{\boldsymbol{x}}_t$, $\tilde{V}_t$, $\tilde{\boldsymbol{\eta}}_t$, and $\tilde{\varepsilon}_t$, where $\tilde{\boldsymbol{\eta}}_t = E(\boldsymbol{\eta}_t|Y_n)$ and $\tilde{\varepsilon}_t = E(\varepsilon_t|Y_n)$. By the definition of smoothing, the mean and variance of $\boldsymbol{x}_t|Y_n$ (i.e. $\tilde{\boldsymbol{x}}_t$ and $\tilde{V}_t$) are calculated iteratively given $\tilde{\boldsymbol{x}}_{t+1}$ and $\tilde{V}_{t+1}$ from the distribution of $\boldsymbol{x}_{t+1}|Y_n$. Since

$$\tilde{\boldsymbol{x}}_t = \mathrm{E}(\boldsymbol{x}_t|Y_n) = \mathrm{E}(\boldsymbol{x}_t|Y_{t-1}, v_t, \cdots, v_n),$$

using the Regression Lemma (Durbin and Koopman, 2001), we have

$$\mathrm{E}(\boldsymbol{x}_t|Y_{t-1}, v_t, \cdots, v_n) = \hat{\boldsymbol{x}}_t + \sum_{j=t}^{n} \mathrm{Cov}(\boldsymbol{x}_t, v_j) F_j^{-1} v_j = \hat{\boldsymbol{x}}_t + \sum_{j=t}^{n} \mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j)'] Z_j' F_j^{-1} v_j,$$

where

$$\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j)'] = \mathrm{E}\{\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j)'|Y_n]\} = \hat{V}_t L_t' \cdots L_{j-1}'.$$

To summarize, we have

$$\tilde{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_t + \hat{V}_t r_{t-1}, \tag{2.28}$$

where r_t is obtained by backwards recursions, that is

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \tag{2.29}$$

for $t = n, n-1, \cdots, 0$, and initiates with $r_n = 0$.

Following the same idea of calculating $\tilde{\boldsymbol{x}}_t$, we have

$$\tilde{\varepsilon}_t = H_t(F_t^{-1} v_t - K_t' r_t) \qquad \text{and} \qquad \tilde{\boldsymbol{\eta}}_t = Q_t R_t' r_t.$$

Define $N_t = \mathrm{Var}(r_t)$. Using (2.29), we have

$$\begin{aligned} N_{t-1} &= \mathrm{Var}(r_{t-1}) = \mathrm{Var}(Z_t' F_t^{-1} v_t + L_t' r_t) \\ &= Z_t' F_t^{-1} \mathrm{Var}(v_t) (F_t^{-1})' Z_t + L_t' \mathrm{Var}(r_t) L_t = Z_t' F_t^{-1} Z_t + L_t' N_t L_t. \end{aligned}$$

N_t is iterated backwardly, initiated with $N_n = 0$. Similarly, using (2.28) and the Regression Lemma (Durbin and Koopman, 2001),

$$\tilde{V}_t = \hat{V}_t - \hat{V}_t \mathrm{Var}(r_{t-1}) \hat{V}_t = \hat{V}_t - \hat{V}_t N_{t-1} \hat{V}_t.$$

As we can see, $\tilde{V}_t = \text{Var}(\boldsymbol{x}_t|Y_n) < \hat{V}_t = \text{Var}(\boldsymbol{x}_t|Y_{t-1})$ since $\boldsymbol{x}_t|Y_n$ contains more information and thus less variation than $\boldsymbol{x}_t|Y_{t-1}$.

The Kalman filter and smoother is efficient because only the inverse of F_t is needed, which usually has lower dimensions than V_t.

2.4 State Space Model with Non-Normal Error Terms: Importance Sampling

This section generalizes the SSM to non-Gaussian error terms. The algorithm proposed in Section 2.3 works for normal error terms. The generalized SSM is similar to (2.21) but the distributions of error terms are not constrained to the Gaussian distribution. Hence the system of equations is

$$\begin{aligned} y_t &= Z_t\,\boldsymbol{x}_t + \varepsilon_t, \\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t + R_t\,\boldsymbol{\eta}_t, \end{aligned} \tag{2.30}$$

where ε_t and $\boldsymbol{\eta}_t$ are not necessarily Gaussian distributions. For illustration, we assume that ε_t and $\boldsymbol{\eta}_t$ have non-Gaussian pdfs $p(\varepsilon_t)$ and $p(\boldsymbol{\eta}_t)$, respectively. In terms of conjugate distributions for error terms, such as the beta distribution (see Zhen and Basawa, 2009), an algorithm similar to the one presented in Section 2.3.1 can be derived. However, if the distributions are not conjugate, Section 2.3.1 could be applied but it is not computationally efficient. Since there is no analytical form for the likelihood function with non-conjugate error terms, simulation techniques are required. Two techniques have been widely used: Importance Sampling and Gibbs Sampling. The importance sampling is a simulation technique using another distribution as an approximation for the distribution of interest. In this thesis, the importance sampling is used since it is easy to derive and takes full advantage of the

Kalman filter and smoother.

2.4.1 Importance Sampling

Suppose we want to calculate the expected value of $f(\boldsymbol{x})$ given Y_n. That is

$$\mathrm{E}_p[f(\boldsymbol{x})|Y_n] = \int f(\boldsymbol{x})p(\boldsymbol{x}|Y_n)d\boldsymbol{x}, \tag{2.31}$$

where $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ and $\mathrm{E}_p(\cdot)$ represents the expectation under the pdf $p(\boldsymbol{x}|Y_n)$. Simulation techniques could be used to approximate (2.31) when there is no closed form solutions. Sometimes the sampling of random variables $\boldsymbol{x}|Y_n$ could be difficult as well. We will tackle this problem by approximating the distribution of $\boldsymbol{x}_t|Y_n$ using an importance density distribution. In terms of the state space model, the purpose of importance sampling is to approximate the non-Gaussian distribution by a Gaussian distribution and then use the Kalman filter and smoother. Durbin and Koopman (2001) illustrate how the importance sampling technique is implemented in SSM.

Let $g(\varepsilon_t)$ and $g(\boldsymbol{\eta}_t)$ be *importance probability density functions*, which approximate $p(\varepsilon_t)$ and $p(\boldsymbol{\eta}_t)$, respectively. This simulation approach is more efficient when $g(\cdot)$ is closer to $p(\cdot)$. If $g(\varepsilon_t)$ and $g(\boldsymbol{\eta}_t)$ are both Gaussian distributions, $g(\boldsymbol{x}|Y_n)$ would also be Gaussian which is used to approximate (2.31), where $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$. The Equation (2.31) under the importance density $g(\boldsymbol{x}|Y_n)$ becomes

$$\begin{aligned}\mathrm{E}_p[f(\boldsymbol{x})|Y_n] &= \int f(\boldsymbol{x})p(\boldsymbol{x}|Y_n)d\boldsymbol{x} \\ &= \int f(\boldsymbol{x})\frac{p(\boldsymbol{x}|Y_n)}{g(\boldsymbol{x}|Y_n)}g(\boldsymbol{x}|Y_n)d\boldsymbol{x} = \mathrm{E}_g\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x}|Y_n)}{g(\boldsymbol{x}|Y_n)}\right],\end{aligned}$$

where $\mathrm{E}_g(\cdot)$ represents the expectation under probability distribution $g(\boldsymbol{x}|Y_n)$. Using Bayes Theorem, we have

$$\mathrm{E}_p[f(\boldsymbol{x})|Y_n] = \frac{g(Y_n)}{p(Y_n)}\int f(\boldsymbol{x})\frac{p(\boldsymbol{x},Y_n)}{g(\boldsymbol{x},Y_n)}g(\boldsymbol{x}|Y_n)d\boldsymbol{x} = \frac{g(Y_n)}{p(Y_n)}E_g\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x},Y_n)}{g(\boldsymbol{x},Y_n)}\right].$$

Similarly, we have $\frac{g(Y_n)}{p(Y_n)} E_g\left[\frac{p(\boldsymbol{x},Y_n)}{g(\boldsymbol{x},Y_n)}\right] = 1$. Therefore,

$$\mathrm{E}_p[f(\boldsymbol{x})|Y_n] = \frac{E_g[f(\boldsymbol{x})w(\boldsymbol{x},Y_n)]}{E_g[w(\boldsymbol{x},Y_n)]},$$

where

$$w(\boldsymbol{x},Y_n) = \frac{p(\boldsymbol{x},Y_n)}{g(\boldsymbol{x},Y_n)} = \frac{p(\boldsymbol{x}_1)\prod_{t=1}^n p(\boldsymbol{\eta}_t)p(y_t|\boldsymbol{x}_t)}{g(\boldsymbol{x}_1)\prod_{t=1}^n g(\boldsymbol{\eta}_t)g(y_t|\boldsymbol{x}_t)}.$$

Suppose that N samples $\check{\boldsymbol{x}}^{(1)}, \check{\boldsymbol{x}}^{(2)}, \cdots, \check{\boldsymbol{x}}^{(N)}$ are drawn from the Gaussian density $g(\boldsymbol{x}|Y_n)$. Hence, the Monte Carlo estimate of the expected value (2.31) is given by

$$\hat{\mathrm{E}}_p[f(\boldsymbol{x})|Y_n] = \frac{\sum_{i=1}^N f(\check{\boldsymbol{x}}^{(i)})w(\check{\boldsymbol{x}}^{(i)},Y_n)}{\sum_{i=1}^N w(\check{\boldsymbol{x}}^{(i)},Y_n)}. \tag{2.32}$$

The importance sampling is now obtained in a specific SSM that will be used for the mortality index. Suppose that

$$\begin{aligned} y_t &= Z_t\,\boldsymbol{x}_t + \varepsilon_t, & &\varepsilon_t \text{ has pdf } p(\varepsilon_t), \\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t + R_t\,\boldsymbol{\eta}_t, & &\boldsymbol{\eta}_t \sim N(0,Q_t). \end{aligned} \tag{2.33}$$

Then $w(\check{\boldsymbol{x}}^{(i)},Y_n)$ in (2.32) can be simplified where $p(\boldsymbol{\eta}) = g(\boldsymbol{\eta})$ and $p(\boldsymbol{x}_1) = g(\boldsymbol{x}_1)$, that is

$$w(\check{\boldsymbol{x}}^{(i)},Y_n) = \frac{p(\check{\boldsymbol{x}}^{(i)},Y_n)}{g(\check{\boldsymbol{x}}^{(i)},Y_n)} = \frac{p(\check{\boldsymbol{x}}_1^{(i)})\prod_{t=1}^n p(\boldsymbol{\eta}_t)p(y_t|\check{\boldsymbol{x}}_t^{(i)})}{g(\check{\boldsymbol{x}}_1^{(i)})\prod_{t=1}^n g(\boldsymbol{\eta}_t)g(y_t|\check{\boldsymbol{x}}_t^{(i)})} = \prod_{t=1}^n \frac{p(y_t|\check{\boldsymbol{x}}_t^{(i)})}{g(y_t|\check{\boldsymbol{x}}_t^{(i)})}, \tag{2.34}$$

where $\check{\boldsymbol{x}}^{(i)} = \{\check{\boldsymbol{x}}_1^{(i)}, \check{\boldsymbol{x}}_2^{(i)}, \cdots, \check{\boldsymbol{x}}_n^{(i)}\}$. Define $\check{\theta}_t^{(i)} = Z_t\check{\boldsymbol{x}}_t^{(i)}$ and $\check{\varepsilon}_t^{(i)} = y_t - \check{\theta}_t^{(i)}$. Equation (2.34) can be obtained using

$$w(\check{\boldsymbol{\theta}}^{(i)},Y_n) = \prod_{t=1}^n \frac{p(y_t|\check{\theta}_t^{(i)})}{g(y_t|\check{\theta}_t^{(i)})}, \tag{2.35}$$

where $\check{\boldsymbol{\theta}}^{(i)} = \{\check{\theta}_1^{(i)}, \check{\theta}_2^{(i)}, \cdots, \check{\theta}_n^{(i)}\}$. Under model (2.33) where y_t is a linear combination of $\theta_t = Z_t\boldsymbol{x}_t$ and ε_t, (2.34) can also be written as

$$w(\check{\boldsymbol{\varepsilon}}^{(i)},Y_n) = \prod_{t=1}^n \frac{p(y_t - \check{\theta}_t^{(i)}|\check{\theta}_t^{(i)})}{g(y_t - \check{\theta}_t^{(i)}|\check{\theta}_t^{(i)})} = \prod_{t=1}^n \frac{p(\check{\varepsilon}_t^{(i)})}{g(\check{\varepsilon}_t^{(i)})}, \tag{2.36}$$

where $\check{\boldsymbol{\varepsilon}}^{(i)} = (\check{\varepsilon}_1^{(i)}, \check{\varepsilon}_2^{(i)}, \cdots, \check{\varepsilon}_n^{(i)})$.

In fact, there is no difference in using smoothing values among $\check{\boldsymbol{x}}$, $\check{\varepsilon}$ and $\check{\theta}$ for $w(\cdot)$ under model (2.33). This conclusion derives from Lemma A.2 shown in Appendix A. According to Lemma A.2, we have $p(y_t|\check{\boldsymbol{x}}_t) = p(y_t - \check{\theta}_t) = p(\check{\varepsilon}_t)$. This means (2.34)-(2.36) are equivalent. We prefer to use $\check{\boldsymbol{\theta}}$ or $\check{\boldsymbol{\varepsilon}}$ instead of $\check{\boldsymbol{x}}$ to implement the simulation because $\check{\theta}_t$ or $\check{\varepsilon}_t$ are one-dimensional at each time t.

Based on (2.34), (2.35), and (2.36), the importance sampling relies on:

- Determine the importance sampling distribution $g(\varepsilon_t)$;
- Sample $\check{\boldsymbol{x}}$ from $g(\boldsymbol{x}|Y_n)$.

2.4.2 Importance Sampling Algorithm

This section proposes the Gaussian importance density function $g(\varepsilon_t)$ when (2.33) is approximated by a Gaussian SSM, which is given by

$$\begin{aligned} y_t &= Z_t\,\boldsymbol{x}_t + \varepsilon_t, & \varepsilon_t &\sim N(0, H_t), \\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t + R_t\,\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(0, Q_t), \end{aligned} \tag{2.37}$$

where H_t means the variance of ε_t at time t. H_t can be estimated in the following way (Durbin and Koopman, 1997),

$$\tilde{H}_t^{-1} = -\frac{1}{\varepsilon_t}\frac{d\log p(\varepsilon_t)}{d\varepsilon_t}\bigg|_{\varepsilon_t = y_t - \tilde{\theta}_t}, \tag{2.38}$$

where $\tilde{\theta}_t = Z_t\tilde{\boldsymbol{x}}_t$, which represents the smoothing value calculated using the Kalman filter and smoother. Since $\tilde{H}_t$ is the estimated variance, the pdf $p(\varepsilon_t)$ must satisfy the following condition

$$-\frac{1}{\varepsilon_t}\frac{d\log p(\varepsilon_t)}{d\varepsilon_t} > 0, \tag{2.39}$$

for all ε_t. Durbin and Koopman (1997) propose a way to find the converged $\tilde{H}_t$ and $\tilde{\theta}_t$ under the following fixed-point iteration:

1. Start with an initial value $\tilde{H}_t$;

2. Get the smoothed values $\tilde{\theta}_t$ using Kalman filter and smoother (Section 2.3.2);

3. Use $\tilde{\theta}_t$ to calculate an updated $\tilde{H}_t$ using (2.38);

4. Repeat the second and third steps until convergence.

This number of iterations before convergence is usually less than 15 to 20. Since the Kalman filter and smoother is efficient, this algorithm usually runs in less than one second.

2.5 Simulation Smoother and Antithetic Variables

2.5.1 Simulation Smoother

Now we discuss how to simulate samples $\check{\boldsymbol{x}}$ from $g(\boldsymbol{x}|Y_n)$. Since we already have the approximated normal state space model, given in (2.33) where H_t is approximated by $\tilde{H}_t$. That is

$$\begin{aligned} y_t &= Z_t\,\boldsymbol{x}_t + \varepsilon_t, & \varepsilon_t &\sim N(0, \tilde{H}_t), \\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t + R_t\,\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(0, Q_t). \end{aligned} \tag{2.40}$$

Durbin and Koopman (2002) introduce an efficient simulation smoother algorithm that we will be able to generate $\check{\theta}_t = Z_t\check{\boldsymbol{x}}_t$. In general, there are two ways to implement this algorithm. On the one hand, $\check{\boldsymbol{x}}_t$ can be simulated directly. On the other hand, $\check{\boldsymbol{\eta}}_t$ and $\check{\varepsilon}_t$ can be generated and then $\check{\boldsymbol{x}}_t$ can be calculated using the Kalman filter and smoother. While the former is more straightforward, the latter may be more computationally efficient since the dimension of $\boldsymbol{\eta}_t$ is usually less than that of $\boldsymbol{x}_t$.

To simulate the smoothing $\check{\boldsymbol{x}}$, we need to generate the smoothing $\check{\varepsilon}_t$ from $g(\varepsilon_t|Y_n)$ and $\check{\boldsymbol{\eta}}_t$ from $g(\boldsymbol{\eta}_t|Y_n)$. Once we simulate $\check{\boldsymbol{\eta}}_t$ from $g(\boldsymbol{\eta}|Y_n)$, $\check{\theta}_{t+1}$ can be calculated using

$\check{\theta}_{t+1} = Z_t(T_t\check{\boldsymbol{x}}_t + R_t\check{\boldsymbol{\eta}}_t)$. Suppose $g(\boldsymbol{\eta}_t|Y_n) \sim N(\tilde{\boldsymbol{\eta}}_t, \dot{Q}_t)$, where $\tilde{\boldsymbol{\eta}}_t = \mathrm{E}(\boldsymbol{\eta}_t|Y_n)$ and $\dot{Q}_t$ is the variance matrix of $\boldsymbol{\eta}_t|Y_n$.

We start with simulating samples $\boldsymbol{\eta}_t^+$ from $N(0, Q_t)$ and ε_t^+ from $N(0, \tilde{H}_t)$. Using (2.40) we get the corresponding y_t^+. Given the general result that in a multivariate normal distribution the conditional variance-covariance matrix of a vector given that a second vector is fixed does not depend on the second vector (see, for example, Anderson, 1984, Theorem 2.5.1), we have $\mathrm{Var}(\boldsymbol{\eta}_t^+|Y_n^+) = \mathrm{Var}(\boldsymbol{\eta}_t|Y_n) = \dot{Q}_t$, where $Y_n^+ = \{y_1^+, y_2^+, \cdots, y_n^+\}$. Denote $\tilde{\boldsymbol{\eta}}_t^+ = \mathrm{E}(\boldsymbol{\eta}_t^+|Y_n^+)$; then $\boldsymbol{\eta}_t^+|Y_n^+ \sim N(\tilde{\boldsymbol{\eta}}_t^+, \dot{Q}_t)$. Therefore, $(\boldsymbol{\eta}_t^+ - \tilde{\boldsymbol{\eta}}_t^+)|y^+ \sim N(0, \dot{Q}_t)$, which is the simulation sample that we want. To sum up, the simulation smoothing samples for $g(\boldsymbol{\eta}_t|y)$ are $\check{\boldsymbol{\eta}}_t = \tilde{\boldsymbol{\eta}}_t + (\boldsymbol{\eta}_t^+ - \tilde{\boldsymbol{\eta}}_t^+) \sim N(\tilde{\boldsymbol{\eta}}_t, \dot{Q}_t)$. The simulation smoothing samples for $g(\varepsilon_t|y)$ can be generated similarly by $\check{\varepsilon}_t = \tilde{\varepsilon}_t + (\varepsilon_t^+ - \tilde{\varepsilon}_t^+) \sim N(\tilde{\varepsilon}_t, \dot{H}_t)$. Algorithm 2.1 summarizes the above ideas.

Algorithm 2.1 Simulation Smoother

The following algorithm is used to generate simulation smoothing samples $\check{\boldsymbol{x}}$ from $g(\boldsymbol{x}|Y_n)$.

1. Calculate $\check{\boldsymbol{x}} = E(\boldsymbol{x}|Y_n)$ using the Kalman filter and smoother in Section 2.3.2.

2. Simulate samples $\boldsymbol{\eta}_t^+$ from $N(0, Q_t)$ and ε_t^+ from $N(0, \tilde{H}_t)$, for $t = 1, 2, \cdots, n$. Using (2.40), $\boldsymbol{x}_t^+$, θ_t^+, and y_t^+ can be obtained iteratively, for $t = 1, 2, \cdots, n$, based on generated samples for the initial value $\boldsymbol{x}_1^+ \sim N(\hat{\boldsymbol{x}}_1, \hat{V}_1)$.

3. Under the classical state space model in (2.40), apply the Kalman filter and smoother in Section 2.3.2 to the simulated values y^+ and get the smoothing values $\tilde{\boldsymbol{x}}_t^+ = E(\boldsymbol{x}_t^+|Y_n^+)$ and $\tilde{\theta}_t^+ = Z_t\tilde{\boldsymbol{x}}_t^+$, for $t = 1, 2, \cdots, n$.

4. Let $\check{\theta}_t = \tilde{\theta}_t + (\theta_t^+ - \tilde{\theta}_t^+)$, where $\check{\theta}_t \sim N(\tilde{\theta}_t, \mathrm{Var}(\theta_t|Y_n))$, for $t = 1, 2, \cdots, n$.

Repeating Algorithm 2.1 could generate N samples from the distribution of $\theta|Y_n$, i.e. $\check{\theta}^{(1)}, \check{\theta}^{(2)}, \cdots, \check{\theta}^{(N)}$, where each simulation path $\check{\theta}^{(i)} = \{\check{\theta}_1^{(i)}, \check{\theta}_2^{(i)}, \cdots, \check{\theta}_n^{(i)}\}$.

2.5.2 Antithetic Variables

Antithetic variables can improve the simulation efficiency by increasing the sample size for each sampling path. We will use three antithetic variables introduced by Durbin and Koopman (1997).

The first antithetic variable is defined as

$$\check{\theta}^{A_1} = 2\tilde{\theta} - \check{\theta}, \tag{2.41}$$

which is easy to calculate and implement.

The second antithetic variable method can be calculated using Algorithm 2.2. Once we generate N samples of $\check{\theta}$, $4N$ samples could be generated including $\check{\theta}$, $\check{\theta}^{A_1}$, $\check{\theta}^{A_2}$, and $\check{\theta}^{A_3}$.

2.6 Monte Carlo Likelihood Estimation

After the construction of the state space model, we need to estimate the parameters using maximum likelihood estimation. Since we use the importance sampling technique with a Gaussian state space model as an approximation of a non-Gaussian state space model, the likelihood function of (2.33) would be based on the approximated model (2.40). The general idea is to obtain the likelihood function of the Gaussian model, and then modify this likelihood function with an adjustment factor to take the approximation into consideration.

Algorithm 2.2 Antithetic Variables Balanced for Scale

1. For each simulation sample based on Algorithm 2.1, let u be a vector containing $r \times n$ variables, all following $N(0,1)$, in order to generate $\boldsymbol{\eta}^+$. Find $c = u'u$. Therefore, c is a univariate random variable which follows a Chi-squared distribution with rn degrees of freedom.

2. Find $q = Pr(\chi^2_{rn} < c)$, and let $\bar{c}^{(i)} = F^{-1}_{\chi^2_{rn}}(1-q)$.

3. A second antithetic variable can be defined as

$$\check{\theta}^{A_2} = \tilde{\theta} + \sqrt{\frac{\bar{c}}{c}}(\check{\theta} - \tilde{\theta}). \tag{2.42}$$

4. Based on (2.41) and (2.42), a third antithetic variable can be constructed:

$$\check{\theta}^{A_3} = \tilde{\theta} + \sqrt{\frac{\bar{c}}{c}}(\check{\theta}^{A_1} - \tilde{\theta}). \tag{2.43}$$

2.6.1 Likelihood Function of Gaussian State Space Model

Denote by $L_g(\psi)$ be the likelihood function of the Gaussian model given in (2.40), where $g(\cdot)$ represents a Gaussian distribution and ψ is the parameter set. Assuming ψ is predetermined, $L_g(\psi)$ can be written briefly as L_g. The likelihood function of model (2.40) can be derived as following:

$$L_g = g(y_1, y_2, \cdots, y_n) = \prod_{t=1}^{n} g(y_t|Y_{t-1}).$$

So we need to calculate the distribution of $g(y_t|Y_{t-1})$. Similar to the derivation in Section 2.3, $g(y_t|Y_{t-1})$ follows a normal distribution because of the conjugate property. Moreover,

$$E_g(y_t|Y_{t-1}) = E_g(Z_t\boldsymbol{x}_t + \varepsilon_t|Y_{t-1}) = Z_t E_g(\boldsymbol{x}_t|Y_{t-1}) = Z_t\hat{\boldsymbol{x}}_t$$

and

$$\begin{aligned} Var_g(y_t|Y_{t-1}) &= Var_g(Z_t\boldsymbol{x}_t + \varepsilon_t|Y_{t-1}) \\ &= Z_t Var_g(\boldsymbol{x}_t|Y_{t-1})Z_t' + Var_g(\varepsilon_t|Y_{t-1}) \\ &= Z_t\hat{V}_t Z_t' + \tilde{H}_t = F_t. \end{aligned}$$

Therefore,

$$y_t|Y_{t-1} \sim N(Z_t\hat{\boldsymbol{x}}_t, F_t). \tag{2.44}$$

The conditional pdf is given by

$$g(y_t|Y_{t-1}) = \frac{1}{\sqrt{2\pi|F_t|}} e^{-\frac{1}{2}(y_t - Z_t\hat{\boldsymbol{x}}_t)'F_t^{-1}(y_t - Z_t\hat{\boldsymbol{x}}_t)} = \frac{1}{\sqrt{2\pi|F_t|}} e^{-\frac{1}{2}v_t'F_t^{-1}v_t}.$$

The log of likelihood function is

$$\log L_g = \sum_{t=1}^{n} \log g(y_t|Y_{t-1}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}(log|F_t| + v_t'F_t^{-1}v_t). \tag{2.45}$$

2.6.2 Adjustment Factors

Denote L_p as the likelihood function of the non-Gaussian SSM in (2.33). By definition,

$$L_p = \int p(\boldsymbol{x}, Y_n)d\boldsymbol{x} = \int p(\theta, Y_n)d\theta.$$

Using the importance sampling technique leads to

$$L_p = \int \frac{p(\theta, Y_n)}{g(\theta, Y_n)} g(\theta, Y_n)d\theta = g(Y_n)\int \frac{p(\theta, Y_n)}{g(\theta, Y_n)} g(\theta|Y_n)d\theta = L_g \int \frac{p(\theta, Y_n)}{g(\theta, Y_n)} g(\theta|Y_n)d\theta.$$

Under model (2.33), $w(\boldsymbol{x}, Y_n) = \frac{p(\boldsymbol{x}, Y_n)}{g(\boldsymbol{x}, Y_n)} = \frac{p(Y_n|\theta)}{g(Y_n|\theta)} = w(\theta, Y_n)$. Then,

$$L_p = L_g \int w(\theta, Y_n)g(\theta|Y_n)d\theta = L_g E_{g(\theta|Y_n)}[w(\theta, Y_n)].$$

The above integral needs to be solved numerically. From Section 2.5, the samples $\check{\theta}$ from $g(\check{\theta}|Y_n)$ can be obtained based on Section 2.5. Therefore L_p can be approximated using these simulation samples by

$$\log L_p \approx \log L_g + \log \bar{w}, \tag{2.46}$$

where

$$\bar{w} = \sum_{i=1}^{N} w(\check{\theta}^{(i)}, Y_n) = \sum_{i=1}^{N} \prod_{t=1}^{n} \frac{p(y_t|\check{\theta}_t^{(i)})}{g(y_t|\check{\theta}_t^{(i)})} = \sum_{i=1}^{N} \prod_{t=1}^{n} \frac{p(y_t - \check{\theta}_t^{(i)})}{g(y_t - \check{\theta}_t^{(i)})}, \tag{2.47}$$

according to (2.35) and (2.36).

However, Durbin and Koopman (1997) point out that the likelihood estimation in (2.46) is biased after taking log function. An approximated unbiased estimator has been proposed.

$$\log L_p \approx \log L_g + \log \bar{w} + \frac{s_w^2}{2N\bar{w}^2}, \tag{2.48}$$

where

$$s_w^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\prod_{t=1}^{n} \frac{p(y_t|\check{\theta}_t^{(i)})}{g(y_t|\check{\theta}_t^{(i)})} - \bar{w})^2.$$

Next we incorporate the antithetic variables into the calculation of the likelihood value. In (2.47), only $\check{\theta}$ is used to calculate w. With antithetic variables, w_i is updated as

$$w_i = \frac{1}{4}[w(\check{\theta}^{(i)}, Y_n) + w(\check{\theta}^{A1,(i)}, Y_n) + w(\check{\theta}^{A2,(i)}, Y_n) + w(\check{\theta}^{A3,(i)}, Y_n)], \tag{2.49}$$

where $\check{\theta}^{A1,(i)}, \check{\theta}^{A2,(i)}, \check{\theta}^{A3,(i)}$ are calculated from (2.41), (2.42) and (2.43). Now we use the updated w_i to calculate the $\bar{w}$ and s_w^2, then substitute into (2.48) to get the approximated unbiased likelihood estimation.

2.7 Diagnostic Test

Diagnostic tests for the Gaussian SSM are performed using standardized one-step ahead prediction errors

$$e_t = \frac{v_t}{\sqrt{F_t}}, \tag{2.50}$$

where v_t and F_t are from the Kalman filter in Section 2.3.2. If the model is properly calibrated, e_t should follow a normal distribution and be uncorrelated, based on (2.44).

The property that e_t is uncorrelated can be derived through Cholesky decomposition (Durbin and Koopman, 2001, for more details). Therefore, e_t should be a standard normal and uncorrelated series under proper model specifications. The normality assumption can be tested with a QQ-plot and a Shapiro-Wilk normality test. The correlation could be tested using an ACF graph and a Ljung-Box Test.

If the error terms are not Gaussian, the estimated error terms should follow the assumed non-Gaussian distribution. As shown in Section 3.3.3, a QQ-plot and a Kolmogorov-Smirnov test would be used for diagnostic test.

Chapter 3

Mortality Rate Models

In this chapter, state space models will be used to fit and predict the mortality index. The state space model would be chosen based on the AIC or BIC criterion. Other two mortality models in the literature will be discussed and compared to the proposed state space model.

The key issue when modeling the mortality index is how to handle shocks. Contrary to Li and Chan (2005, 2007), the mortality shocks are critical components of the mortality index. Li and Chan (2005) mention that the outlier detection technique cannot distinguish the extreme values whether they are outliers or from a fat-tailed distribution. Incorporating a fat-tailed distribution into our modeling of the mortality index is more reasonable in terms of the nature of our problem. Therefore, a model based on time series for the baseline mortality that includes a fat-tailed distribution to model the shocks is proposed.

The state space model is an ideal choice. The classical state space model can handle various linear connections between latent variables and observations. Extending to non-Gaussian state space models, we can use different distributions and linear transformations to model error terms. Moreover, generalized SSMs are

useful tools and model various relations (univariate or multidimensional, linear or non-linear, Gaussian or non-Gaussian, etc.) between latent variables, error terms, and observations.

According to the observations, the shocks occurred several times over the last one hundred years with a the huge shock in 1918. Since most of these shocks occurred over a one-year period, we assume that the mortality shocks are transient. Moreover, for better projecting mortality shocks, we also assume that the occurrence and scale of shocks would not be affected by the improvement of mortality. In other words, we believe that shocks would happen in the same behavior as what they did in the past (i.e. independent of time).

3.1 ARIMA Models

In this section, we rewrite the ARIMA model in the form of SSMs, that is

$$\begin{aligned} y_t &= Z_t \, \boldsymbol{x}_t, \\ \boldsymbol{x}_{t+1} &= T_t \, \boldsymbol{x}_t + R_t \, \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2), \end{aligned} \tag{3.1}$$

where

$$\boldsymbol{x}_t = \begin{bmatrix} \dot{y}_{t-1} \\ \Delta^1 \dot{y}_{t-1} \\ \vdots \\ \Delta^d \dot{y}_t \\ \phi_2 \Delta^d \dot{y}_{t-1} + \cdots + \phi_r \Delta^d \dot{y}_{t-r+1} + \varphi_1 \eta_t + \cdots + \varphi_{r-1} \eta_{t-r+2} \\ \phi_3 \Delta^d \dot{y}_{t-1} + \cdots + \phi_r \Delta^d \dot{y}_{t-r+2} + \varphi_2 \eta_t + \cdots + \varphi_{r-1} \eta_{t-r+3} \\ \vdots \\ \phi_r \Delta^d \dot{y}_{t-1} + \varphi_{r-1} \eta_t \\ \mu_t \end{bmatrix}_{(d+r+1)\times 1}$$

$$Z_t = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{1\times(d+r+1)}$$

$$T_t = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \phi_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \phi_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \phi_r & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \beta \end{bmatrix}_{(d+r+1)\times(d+r+1)} \qquad R_t = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{r-1} \\ 0 \end{bmatrix}_{(d+r+1)\times 1} \tag{3.2}$$

$r = max(p, q+1)$, $\dot{y}_t = y_t - \mu_t$, and $\Delta^d \dot{y}_t$ is the n^{th} order differentiation of $\dot{y}_t$. The parameter μ_t represents the decreasing tendency with decreasing rate β and initial value μ_1. The variable $\dot{y}_t$ follows an ARIMA(p,d,q) and correspondingly $\Delta^d \dot{y}_t$ follows an ARMA(p,q).

3.2 Gaussian Mortality Models

This section tests if it is suitable to use normal error terms to model the mortality index. In fact, Section 1.2 concludes that normal error terms do not provide a good fit of the mortality index using the outlier detection technique. However, it would be interesting to investigate if similar conclusions could be drawn using the SSM analysis.

Using model (3.1), a normal error ε_t could be added to model shock effects,

$$\begin{aligned} y_t &= Z_t\,\boldsymbol{x}_t + \varepsilon_t, & \varepsilon_t &\sim N(0,\sigma_\varepsilon^2),\\ \boldsymbol{x}_{t+1} &= T_t\,\boldsymbol{x}_t + R_t\,\eta_t, & \eta_t &\sim N(0,\sigma_\eta^2), \end{aligned} \tag{3.3}$$

where $\boldsymbol{x}_t$, Z_t, T_t, and R_t are given in (3.2), and y_t is the logit transform observations given in Figure 1.2.

We apply the Kalman filter and smoother introduced in Section 2.3.2 to fit the model. To choose the best model, we calculate the likelihood function for a wide range of ARIMA models using (2.45) and evaluate the AIC and BIC criteria, which are given in (1.4). According to (2.45), we need to get F_t and v_t, which are calculated during the filtering process discussed in Section 2.3.2. Then the maximum likelihood value and the parameters can be estimated using a numerical optimization algorithm[1]. Table 3.1 shows a sample of the ARIMA models with different orders.

As we can see, the ARIMA$(1,0,0)$ model has the smallest AIC and BIC values. The ARIMA$(1,0,0)$ process for $\dot{y}_t$ is $\dot{y}_t = \phi_1\dot{y}_{t-1} + \eta_t$. The mortality index model

[1]Using the R package: *optim* to perform the optimization.

Table 3.1: Gaussian SSM models selection

ARIMA	Parameters	Likelihood	AIC	BIC
(0,0,0)	4	97.62	−187.25	-176.48
(1,0,0)	5	164.82	−319.63	-306.17
(0,0,1)	5	128.53	−247.06	-233.60
(1,0,1)	6	165.12	−318.25	-302.10
(2,0,1)	7	165.12	−316.24	-297.40

using a SSM with ARIMA$(1,0,0)$ baseline mortality is given by

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \dot{y}_t \\ \mu_t \end{bmatrix} + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2), \\ \begin{bmatrix} \dot{y}_{t+1} \\ \mu_{t+1} \end{bmatrix} &= \begin{bmatrix} \phi_1 & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} \dot{y}_t \\ \mu_t \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{t+1}, & \eta_t &\sim N(0, \sigma_\eta^2). \end{aligned} \tag{3.4}$$

In this model, there are three types of parameters: ϕ_1 is the parameter of the AR process for the latent variable $\dot{y}_t$, σ_ε^2 and σ_η^2 are the variances of error terms, and β controls the decreasing rate of μ_t. The estimated parameters are given in Table 3.2.

Table 3.2: Parameter estimation for ARIMA$(1,0,0)$

$\phi_1 = 0.9502$	$\sigma_\eta^2 = \exp(-6.4117)$	$\sigma_\varepsilon^2 = \exp(-7.2338)$
$\beta = 0.991$	$\mu_1 = 2.0605$	
$\log L = 164.82$	$AIC = -319.63$	$BIC = -306.17$

The estimated parameters are plugged into the model in (3.4). Apply the Kalman filter and smoother in Section 2.3.2 to get the smoothed $\tilde{x}_t$ as well as $\tilde{\eta}_t$ and $\tilde{\varepsilon}_t$. Figure 3.1 shows that the model fits the data well before diagnostic tests.

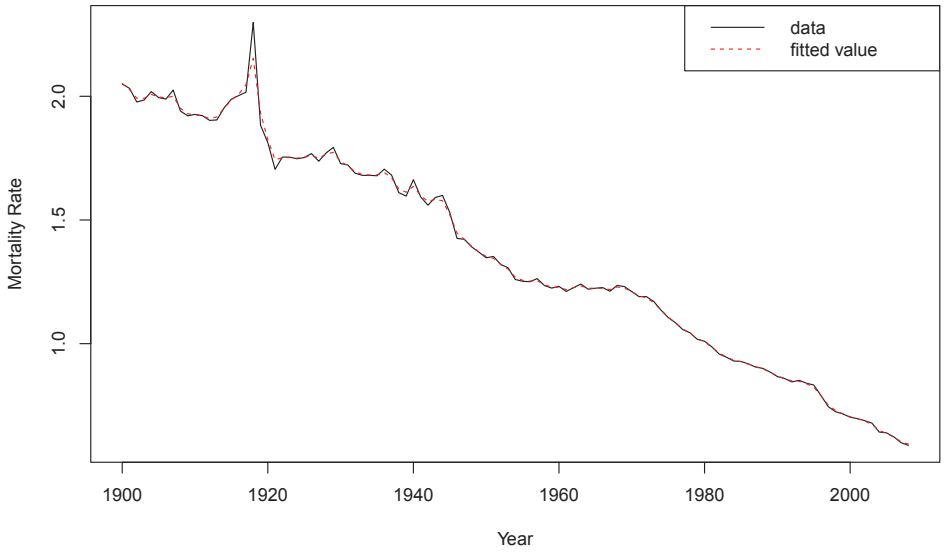

Figure 3.1: The observations y_t v.s. fitted value $\tilde{\boldsymbol{x}}_t$

However, Figure 3.2 indicates some problems for $\tilde{\eta}_t$ and $\tilde{\varepsilon}_t$. The diagnostic tests show that the two error terms are not normally distributed. According to the QQ-plot, both $\tilde{\eta}_t$ and $\tilde{\varepsilon}_t$ have fatter tails than normal distributions. In other words, normal error terms are not able to capture mortality shocks.

3.3 Non-Gaussian Mortality Models

Recall from the outlier detection technique in Section 1.2 that the ARIMA model could provide a good fit to the baseline mortality. In terms of the time series analysis, ARIMA models are flexible and easy to handle. This implies that the baseline process could be modeled using normal residuals after removing shock effects.

In state space models, the error term in the observation equation ε_t disturbs y_t without affecting the latent variable $\boldsymbol{x}_t$, which behaves in a similar way to AOs in

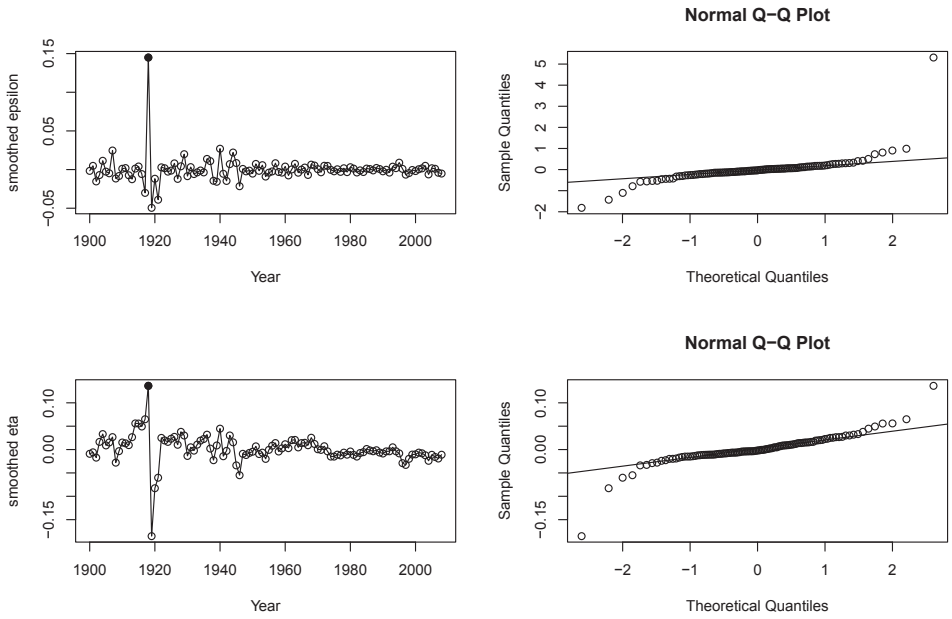

Figure 3.2: The smoothed values and normal QQ-plot of $\tilde{\eta}_t$ (lower) and $\tilde{\varepsilon}_t$ (upper)

outlier detection analysis. We mentioned that mortality shocks are supposed to be transient, so we expect that ε_t would handle the mortality shocks in y_t, which is reasonable because shocks are mainly detected as AOs (Table 1.3).

Since we want to incorporate the shocks into the state space model, there are two approaches that could be considered to model transient shocks. The shocks can be added into the deterministic part $Z_t \boldsymbol{x}_t$ or the stochastic part ε_t. Both of these approaches are practical for model calibration. For the former, indicator functions could be used to model the occurrence of shocks. Its prediction becomes the key issue. A natural way to model and predict indicator sequences is using Markov chains (Mächler and Bühlmann, 2004) or using binary sequence modeling, e.g, Keenan (1982) and Zhen and Basawa (2009). However this approach is more complex. On the other hand, letting the stochastic part ε_t capture the shocks leads to a simplified model with nice predictive properties.

Based on the above discussion, we propose a fat-tailed distribution or a mixed distribution to represent the error term ε_t which is used to model shocks. According to the outlier detection analysis in Section 1.2, the mortality index has five outliers; three of them have positive effects and two of them have negative effects. This implies that the domain of the proposed distribution should include positive and negative values. The mode and mean are expected to be around zero, since these error terms should not influence the average level of mortality rates, but just model the possible occurrence of shocks. The skewness of mortality shocks is expected to be positive since larger positive extreme mortality shocks were discovered. In fact, the concavity of the logit transform function (1.1) adds this effect on the skewness of the projected mortality index.

3.3.1 Model Calibration

We propose a general state space model where the state equation is an ARIMA(p, d, q) process and the observation equation includes a fat-tailed distribution for ε_t, that is

$$\begin{aligned} y_t &= Z_t\, \boldsymbol{x}_t + \varepsilon_t, \qquad && \varepsilon_t \sim \text{non-Gaussian distribution}, \\ \boldsymbol{x}_{t+1} &= T_t\, \boldsymbol{x}_t + R_t\, \eta_t, \qquad && \eta_t \sim N(0, \sigma_\eta^2), \end{aligned} \tag{3.5}$$

where $\boldsymbol{x}_t$, Z_t, T_t, and R_t are given in (3.2).

A possible choice for ε_t is the t-distribution. We will use this as an example to estimate parameters and the likelihood function given specific orders of ARIMA models. Later on, we will consider more possible error distributions. Here, we consider

$$\frac{\varepsilon_t}{\sigma_\varepsilon} \sim t_v \quad \text{and} \quad \dot{y}_t \sim \text{ARIMA}(2, 1, 0). \tag{3.6}$$

Compared with the Gaussian model in Section 3.2, there is one more parameter v in this model.

3.3.2 Estimation of Parameters

The importance sampling technique introduced in Chapter 2 is applied to estimate the SSM. The general idea is to use a classic SSM to approximate (3.6). In other words, we use a Gaussian distribution $g(\varepsilon_t)$ as importance density to approximate $p(\varepsilon_t)$. The approximated SSM is given by (3.6) with $\varepsilon_t \sim N(0, \tilde{H}_t)$, where

$$\tilde{H}_t = \frac{\sigma_\varepsilon^2 v + (y_t - \tilde{\theta}_t)^2}{v+1} = \frac{\sigma_\varepsilon^2 v + \tilde{\varepsilon}_t^2}{v+1}. \tag{3.7}$$

The variance $\tilde{H}_t$ is obtained using (2.38), where

$$p(\varepsilon_t) \propto \left(1 + \frac{\varepsilon_t^2}{v\sigma_\varepsilon^2}\right)^{-\frac{v+1}{2}} \frac{1}{\sigma_\varepsilon}. \tag{3.8}$$

Therefore,

$$\tilde{H}_t^{-1} = -\frac{1}{\varepsilon_t}\frac{d\log p(\varepsilon_t)}{d\varepsilon_t}\bigg|_{\varepsilon_t = y_t - \tilde{\theta}_t} = \frac{v+1}{\sigma_\varepsilon^2 v + (y_t - \tilde{\theta}_t)^2}.$$

According to Lemma A.1, $\tilde{\theta}_t + \tilde{\varepsilon}_t = y_t$, which leads to

$$\tilde{H}_t = \frac{\sigma_\varepsilon^2 v + (y_t - \tilde{\theta}_t)^2}{v+1} = \frac{\sigma_\varepsilon^2 v + \tilde{\varepsilon}_t^2}{v+1}.$$

Equation (3.7) indicates that we could choose either $\tilde{\theta}_t$ or $\tilde{\varepsilon}_t$ to calculate $\tilde{H}_t$, where $\tilde{\theta}_t$ or $\tilde{\varepsilon}_t$ are both smoothed values by the Kalman filter and smoother. Note that $\tilde{H}_t$ calculated based on the t-distribution satisfies (2.39) for any $\tilde{\varepsilon}_t$.

We apply the algorithm presented in Section 2.4.2. With the Kalman filter and smoother as well as the convergent $\tilde{H}_t$, we are able to calculate the log-likelihood value of the approximated Gaussian model using (2.45). To calculate the log-likelihood of our original model (3.6) simulation techniques are required based on Section 2.6.2. All the steps for the Kalman filter and smoother are not necessary. In fact, only v_t, $\hat{\boldsymbol{x}}_t$, and r_t are required. Then $\tilde{\varepsilon}_t$ can be calculated directly using the Kalman smoother.

For comparison, several distributions are used to model the error term ε_t. The spliced t-distribution has different t-distributions with different degrees of freedom,

defined over positive and negative domains. The second spliced distribution includes normal distribution for the negative domain and the t-distribution for the positive domain. Therefore, there is one more parameter compared to the single t-distribution. Then we choose the corresponding orders for ARIMA models according to the AIC criterion. The results are shown in Table 3.3. The t-distribution ranked first using both the AIC and BIC criteria. The AIC and BIC are much lower for normal distributions which cannot capture the shocks properly.

Table 3.3: SSM models comparison

ε_t	ARIMA	Parameters	Likelihood	AIC	BIC
t-distribution	(2,1,0)	7	245.99	−477.97	-459.13
Normal	(1,0,0)	6	164.61	−319.63	-306.17
spliced t-dist	(2,1,0)	8	246.70	−477.40	-455.87
spliced normal and t-dist	(3,0,0)	7	235.05	−456.10	-437.26

MLEs are presented in Table 3.4 and are obtained using numerical optimization in R[2]. The parameters σ_η^2 and σ_ε^2 are quite close, meaning that the volatilities caused by these scaling parameters are equivalent. This implies that both of the state and observation error terms share the volatilities of the mortality index. The parameter β indicates the mortality index has a decreasing tendency. The degrees of freedom of the t-distribution v, the most important parameter in this model, is estimated to be 1.3 which means that the second moment does not exist and more emphasis is put on fat-tailed distributions. Compared to the results in model (3.4), the likelihood as well as AIC and BIC are improved significantly, indicating the superiority of applying fat-tail distributions.

[2]We can use *optim*, *DEoptim* or *genoud* in R to perform the optimization.

Table 3.4: Estimated parameters for non-Gaussian SSM

$\phi_1 = 1.2892$	$\phi_2 = -0.6397$	$\sigma_\eta^2 = \exp(-9.6201)$
$\sigma_\varepsilon^2 = \exp(-10.2831)$	$\beta = 0.9916$	$\mu_1 = 2.0564$
$v = 1.3049$		
$\log L = 245.99$	$AIC = -477.97$	$BIC = -459.13$

3.3.3 Model Diagnostic

This section discusses two diagnostic tests for the approximated model and the error term ε_t, respectively. We begin and analyze the standardized one-step ahead prediction error e_t obtained from the approximated SSM. First we test its normality. We can see from the QQ-plot of standardized one-step ahead prediction errors e_t (Figure 3.3) that the normal distribution fits well. The Shapiro-Wilk Normality statistic is $W = 0.9314$[3] which gives a p-value of 0.2008 above the confidence level 0.05.

Then we test its independence. From the ACF and PACF graphs in Figure 3.3 we can conclude that there is no autocorrelation or partial correlation in e_t. This is also confirmed by the Ljung-Box test (Cryer and Chan, 2008), which shows that the corresponding statistic p-value is above the confidence level 0.05 for lags between 6 and 20.

Next, we will test whether ε_t follows a t-distribution. Since we are not able to separate ε_t from v_t because of the latent variable $\boldsymbol{x}_t$, we would use $\tilde{\varepsilon}_t$ as samples to do the diagnostic tests. Figure 3.4 shows that the t-distribution is able to capture the huge shock in 1918, $\tilde{\varepsilon}_{1918}^{t-dist} = 0.351$ while $\tilde{\varepsilon}_{1918}^{Normal-dist} = 0.16$ as discussed in Section 3.2, which reflects the difference of the ability to capture large shocks by the

[3]Use *shapiro.test* in R.

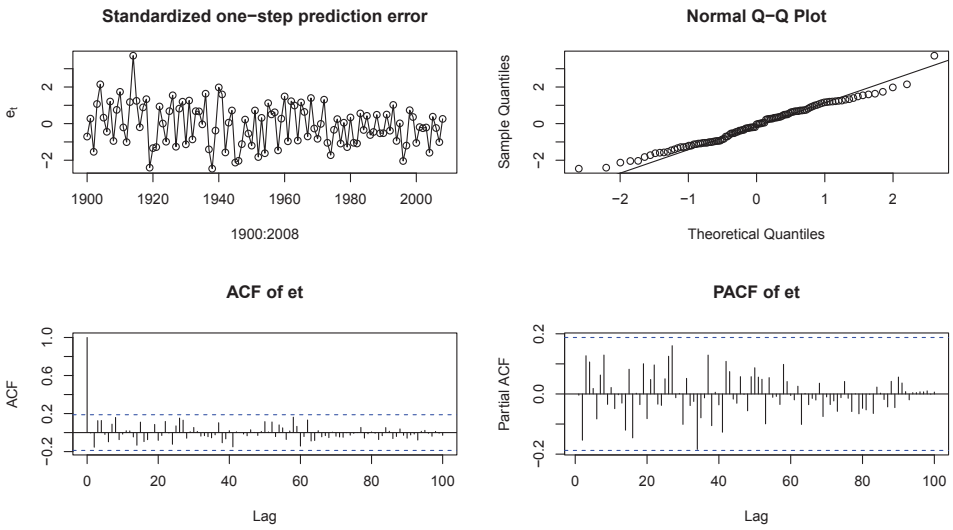

Figure 3.3: Diagnostic tests on standardized one-step ahead prediction error e_t

t-distribution and the normal distribution.

Figure 3.4 also presents the empirical fitted cumulative distribution functions of $\tilde{\varepsilon}_t$ as well as a t-distribution QQ-plot. Generally speaking, both of the *cdf* plot and QQ-plot show that the t-distribution provides a good fit. The QQ-plot shows that the shock in 1918 is an extreme value.

We also apply non-parametric diagnostic tests to verify the t-distribution. The Kolmogorov-Smirnov and Cramer-von Mises tests do not reject that the $\tilde{\varepsilon}_t$ is from a t-distribution with p-value of 0.2533 and 0.2033, respectively. According to these diagnostic tests, the t-distribution is a reasonable choice to model shocks.

3.3.4 Model Fitting and Forecast

Figure 3.5 shows four curves: the transformed data y_t, the filtering, smoothing and one-step ahead forecast of θ_t based on the approximated model. The one-step ahead

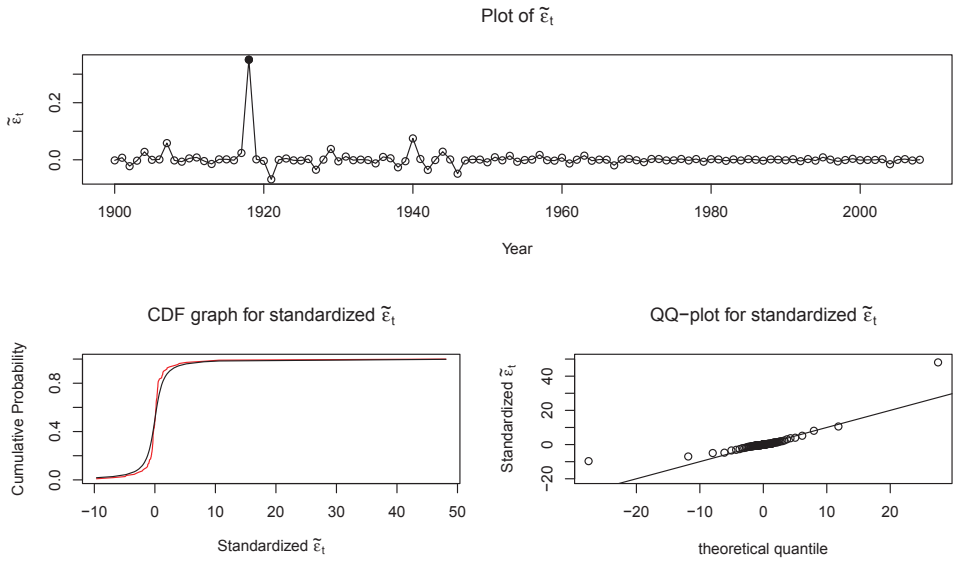

Figure 3.4: Diagnostic tests on standardized smoothed observation error $\tilde{\varepsilon}_t$

forecast $\hat{\theta}_t$ and filtering values $\bar{\theta}_t$ are calculated using,

$$\hat{\theta}_t = Z_t \hat{\boldsymbol{x}}_t = y_t - v_t, \qquad \bar{\theta}_t = Z_t\, \bar{\boldsymbol{x}}_t = Z_t(\hat{\boldsymbol{x}}_t + \hat{V}_t\, Z_t'\, F_t^{-1}\, v_t),$$

and smoothed value $\tilde{\theta}_t$ is illustrated in Section 2.3.2. While one-step ahead forecast values reflect the tendency based on up-to-date information, the smoothed value is closer to the mean level since it takes into account the future information.

Next, we forecast the mortality index for thirty years. With the estimated parameters under model (3.6), we generate future paths by simulating samples from the normal and t-distributions. Figure 3.6 shows the prediction for thirty years of y_t using $100,000$ simulated paths. Since the estimated degrees of freedom $v > 1$, the theoretical mean for y_t exists. Figure 3.6 also shows the 1% confidence interval, which is symmetric around the mean value.

Figure 3.7 shows the actual prediction of the mortality index q_t with mean and 1% confidence intervals.

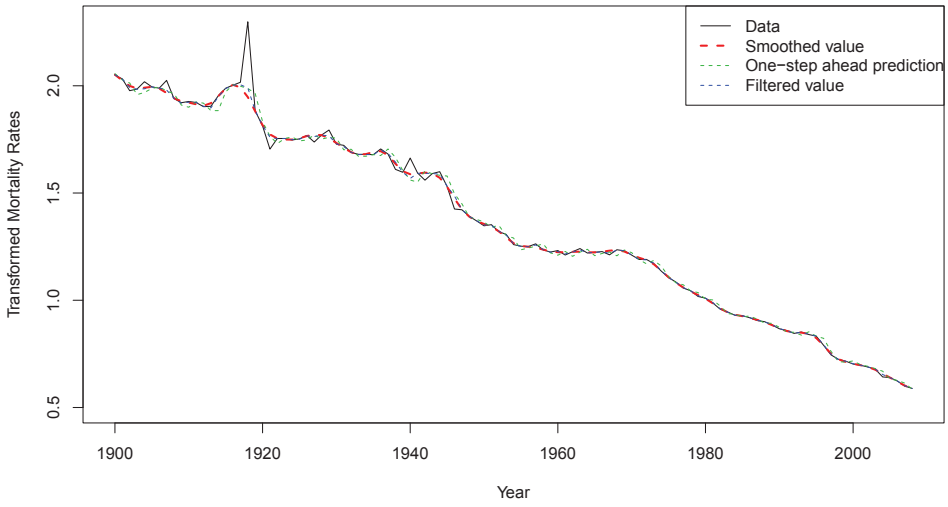

Figure 3.5: The plot of data y_t, one-step ahead forecast $\hat{\theta}_t$, filtered values $\bar{\theta}_t$ and smoothed values $\tilde{\theta}_t$

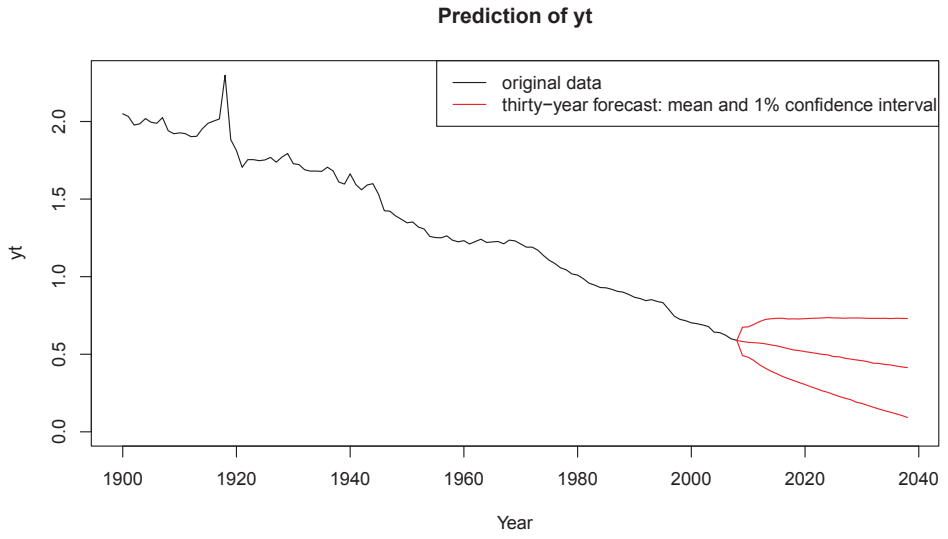

Figure 3.6: Thirty-year prediction of y_t: mean values of simulated paths and 1% confidence interval.

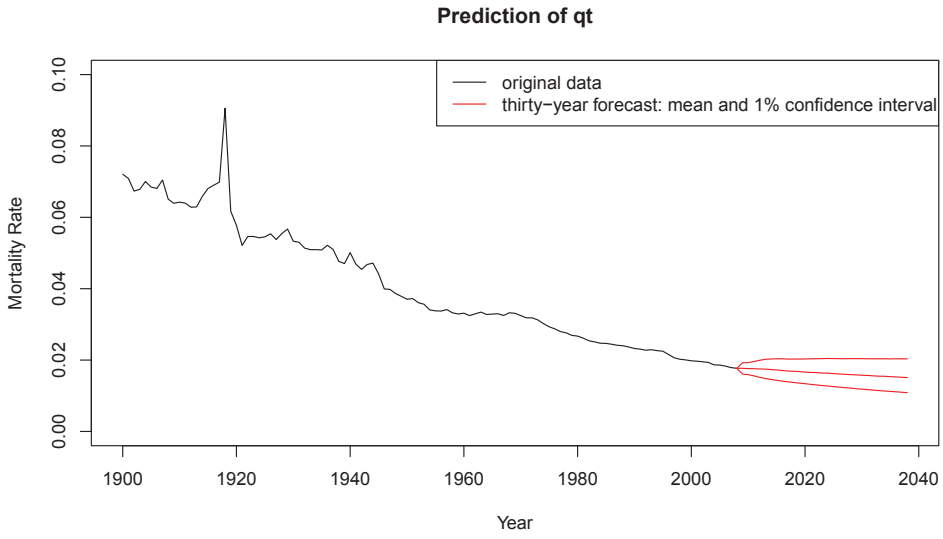

Figure 3.7: Thirty-year prediction of q_t: mean values of simulated paths and 1% confidence interval.

3.4 The Outlier Analysis and State Space Models

In the previous two sections, we use the normal and student distributions to model the shock effects. Figure 3.8 shows the graphs of $\tilde{\varepsilon}_t$ for these two models. Table 3.5 gives the five solid points which represent the outliers detected in Section 1.2. The first two lines are detected shock effects from Table 1.3. Five of the largest shock effects, according to $\tilde{\varepsilon}_t^{t-dist}$, are consistent with the detected shocks and their values are close to δ_T. On the other hand, $\tilde{\varepsilon}_t^{Normal-dist}$ under the Gaussian SSM model does not capture shocks properly.

Previously we mentioned that it was not important for the mortality index to distinguish between IOs and AOs. It is interesting to note that our model captures the shock effect in 1946, which was detected as IO in Section 1.2. In fact, if we apply the outlier detection technique based on $\tilde{\theta}_t$, which excludes shock effects $\tilde{\varepsilon}_t$, neither

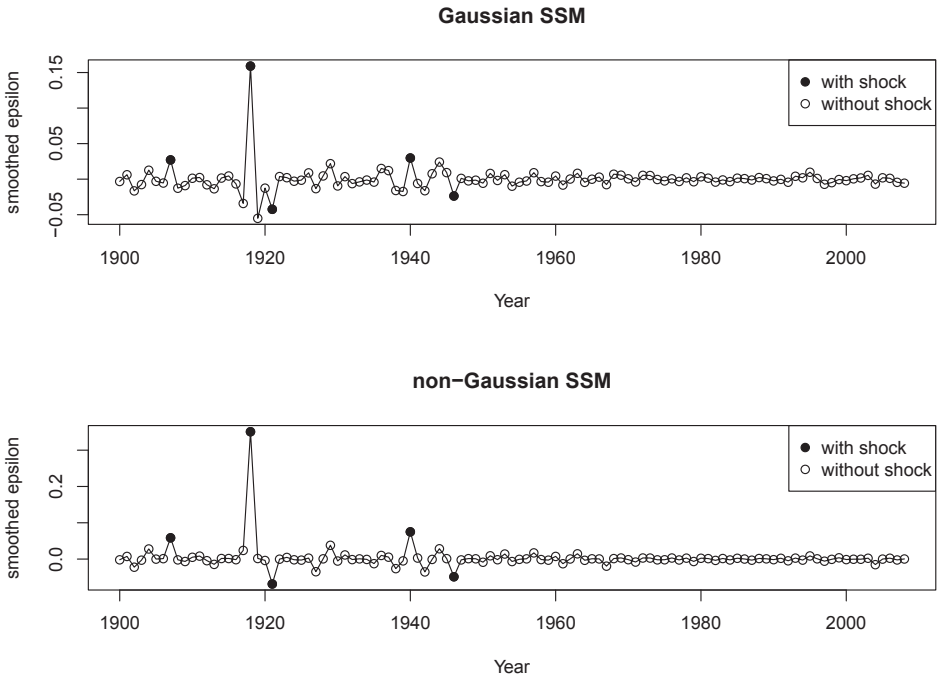

Figure 3.8: $\tilde{\varepsilon}_t$ marked with detected outliers: $\tilde{\varepsilon}_t^{Normal-dist}$ (upper) and $\tilde{\varepsilon}_t^{t-dist}$ (lower)

Table 3.5: Comparison of shock effects

Year T	1907	1918	1921	1940	1946
δ_T with AIC	0.05465	0.3219	−0.0738	0.055	−0.07327
δ_T with BIC	0.05872	0.356	−0.07922	0.0684	−0.08466
$\tilde{\varepsilon}_t^{Normal-dist}$	0.02718	0.1592	−0.04229	0.02986	−0.02359
$\tilde{\varepsilon}_t^{t-dist}$	0.05871	0.3510	−0.06847	0.07518	−0.04859

IOs nor AOs can be detected. This means that ε_t is actually responsible for capturing all of shock effects so that the baseline model does not contain any shocks.

3.5 Comparison of Mortality Rate Models

In this section, we compare our model with other mortality index models, which have already been discussed briefly in the Introduction.

3.5.1 Lin and Cox (2008)

Lin and Cox (2008) model the mortality index using two log-normal processes, one for baseline mortality rates and the other for modeling shocks, that is

$$\begin{aligned}
&\text{baseline model:} && \bar{q}_t = \bar{q}_{t-1} e^{(\alpha-\sigma^2/2)+\sigma Z_{1,t}}, \\
&\text{shock model:} && Y_t = \begin{cases} e^{m+sZ_{2,t}}, & \text{with probability } p, \\ 1, & \text{with probability } 1-p, \end{cases} \\
&\text{mortality rate model:} && q_t = \bar{q}_t Y_t,
\end{aligned}$$

where $Z_{1,t}$ and $Z_{2,t}$ are two independent standard normal variables. The baseline process is a geometric Brownian motion with drift parameter α and scale parameter σ. In terms of shocks, they use a Bernoulli distribution to model the probability of shocks and thus a piece-wise defined function to model the shock effects with shift parameter m and scale parameter s. They assume that these two processes are independent, which means that the shocks do not affect the baseline mortality. The mortality index is the synthesis of these two processes. In the case that a shock occurs, a positive value m is added to the mean and the variance is increased by s. But the shock at time t would not affect the next period; in other words, the mortality rate would go back to its baseline unless there is another shock that happens at time $t+1$. It is worth mentioning that q_t is observable and $\bar{q}_t$ is a latent variable.

Using $y_t = \ln(q_t)$, we can rewrite the above model in terms of SSM,

$$\begin{aligned} &\text{observation equation:} && y_t = \bar{y}_t + (m + sZ_2)B_t, \\ &\text{state equation:} && \bar{y}_t = \bar{y}_{t-1} + (\alpha - \sigma^2/2) + \sigma Z_1, \end{aligned} \tag{3.9}$$

where B_t are $i.i.d$ random variables that follow a Bernoulli distribution with parameter p. The baseline mortality is an ARIMA(0,1,0) process. Therefore, this model is a general form of Lee and Carter (1992) where modeling shocks is much simpler. The main difference with (3.6) is that they use two random variables (Z_2 and B_t) to model the shocks while our model uses the t-distribution. The Lin and Cox (2008) model is much easier to fit since the combination of geometric Brownian motion and normal error terms in the observation equation lead to an analytical form for the likelihood function.

The main drawback is the difficulty to verify whether the baseline model follows geometric Brownian motion and the shock effects can be modeled by a log-normal distribution. Moreover, Table 1.3 and Figure 3.8 show that there is a negative shock in 1921, but Lin and Cox (2008) focus on positive shocks and assumed that the baseline model can take this negative shock into account. However, Figure 3.2 indicates that normal error term is not able to capture both positive and negative shocks. In addition, they also simplify the calculation of the likelihood function, which would not be accurate.

3.5.2 Milidonis et al. (2011)

Milidonis et al. (2011) use a log-normal regime switching process to model the

increment in the mortality index. They use

$$\text{Two-Regime Model:} \quad Y_t = \begin{cases} \mu_1 t + \sigma_1 W_t, & \text{Volatile Regime,} \\ \mu_2 t + \sigma_2 W_t, & \text{Stable Regime,} \end{cases}$$

$$\text{Transition Matrix:} \quad P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix},$$

where $Y_t = \ln\left(\frac{q_t}{q_{t-1}}\right)$.

Based on their estimation results, the drifts of both regimes fluctuate around the decreasing tendency rate[4]. Therefore, this model uses scale parameters to distinguish between stable regime and volatile regime where most shocks appeared. Based on Figure 1.1, the mortality index has a high volatility during the first half of the twentieth century and a lower volatility in the past fifty years. Moreover, the volatility may be overestimated because mortality rates would go back to baseline once shocks happen. The main difference between these two models is that Lin and Cox (2008) use an add-on variable Y_t to model shocks so they do not allow this volatile regime to last over one-year. Milidonis et al. (2011) model this using a transition probability instead.

The main difference between our model and other two models is the way to model shocks. While Lin and Cox (2008) use a shift parameter to control shock effects and Milidonis et al. (2011) use a scale parameter to represent the volatile regime, we use a shape parameter in t-distributions to represent the shock effects. In addition, our model does not define the occurrence of shocks.

[4]It could represent the parameter redundancy.

3.6 Pricing Mortality Bond

This section prices the Swiss Re mortality bonds issued in 2003. These bonds are securities used by insurance companies to hedge their catastrophic mortality. The cash flows of this bond include coupons and a face amount linked with the mortality index at maturity. Coupons vary with the LIBOR rate while the payment at maturity would be determined based on the realized mortality index. For instance, if there is mortality shock during this period, the payment at maturity would be decreased.

This thesis mainly discusses the Swiss Re bond, the first mortality bond issued in 2003. Two more mortality bonds were issued by Swiss Re Company in 2004. The same method can be applied to get the present value. We do not consider those bonds since they are based on a different mortality index (see Milidonis et al., 2011). We will first introduce the payoff structure of this bond and find the actuarial present value. We will compare our model with Lin and Cox (2008) and Milidonis et al. (2011) as well as the industrial empirical result given by Wang (2004).

3.6.1 Design of the Swiss Re Bond

In December 2003, Swiss Re issued a catastrophic mortality bond with a principle of $400 million that would mature in three years (from 2004 to 2006) with quarterly coupons (totally 12 coupons) and a mortality-linked face amount at maturity.

The discounted cash flow (DCF) of the Swiss Re bond can be calculated as follows,

$$DCF = \sum_{j=1}^{12} \frac{C_j}{(1+\frac{r}{4})^j} + \frac{P \times F}{(1+\frac{r}{4})^{12}}, \tag{3.10}$$

where the quarterly coupon value at the j^{th} quarter is $C_j = \frac{(S+L_j)F}{4}$, $F = \$400$ million is the face value of the bond, $S = 1.35\%$ is the spread, the LIBOR rate $L_j = 1\%$ is assumed, the nominal discounting rate $r = 1\%$ is also assumed, and the percentage

of face value at maturity P is defined by the Swiss Re bond as follows,

$$P = max\left(100\% - \sum_{t=2004}^{2006} loss_t, 0\right), \tag{3.11}$$

where

$$loss_t = \begin{cases} 0, & \text{if } q_t \leq 1.3q_0, \\ 1 - \dfrac{1.5q_0 - q_t}{0.2q_0}, & \text{if } 1.3q_0 < q_t \leq 1.5q_0, \\ 1, & \text{if } q_t > 1.5q_0, \end{cases} \qquad t = 2004, 2005, 2006,$$

and $q_0 = q_{2003}$.

The probability of getting two mortality shocks or more within a three-year period is approximately

$$\begin{aligned} &1 - [P(q_t < 1.3q_0)]^3 - 3[P(q_t < 1.3q_0)]^2 P(q_t > 1.3q_0) \\ \approx &1 - \left[P\left(\frac{\varepsilon_t}{\sigma_\varepsilon} < \frac{\ln 1.3}{\sigma_\varepsilon}\right)\right]^3 - 3\left[P\left(\frac{\varepsilon_t}{\sigma_\varepsilon} < \frac{\ln 1.3}{\sigma_\varepsilon}\right)\right]^2 \left[P\left(\frac{\varepsilon_t}{\sigma_\varepsilon} > \frac{\ln 1.3}{\sigma_\varepsilon}\right)\right] \\ = &9.5717 \times 10^{-6}. \end{aligned}$$

For this reason, Lin and Cox (2008) approximate $\sum_{t=2004}^{2006} loss_t$ in (3.11) by $max_{t=2004,2005,2006}(loss_t)$[5]. Then (3.11) can be written as

$$P = \begin{cases} 1, & \text{if } q_{max} \leq 1.3q_0, \\ \dfrac{1.5q_0 - q_{max}}{0.2q_0}, & \text{if } 1.3q_0 < q_{max} \leq 1.5q_0, \\ 0, & \text{if } q_{max} > 1.5q_0, \end{cases} \tag{3.12}$$

where $q_{max} = max(q_{2004}, q_{2005}, q_{2006})$. We will also use (3.12) since it is easier to handle and leads to the same result.

3.6.2 Market Price of Risk

Obtaining that the actuarial present value of the discounted cash flow will not match the actual face value of the bond, Lin and Cox (2008) explain this discrepancy by

[5]This approximation may not be accurate when the duration of the bond is long.

introducing the mortality market price of risk, which is introduced through the Wang (2002) transform. Generally speaking, this mortality risk premium is a surcharge on the observed mortality imposed by investors in exchange of taking the mortality risk. Hence, the cdf of the observed mortality index is transformed using the two-factor Wang transform:

$$F^*_{q_{max}}(x) = Q[\Phi^{-1}(F_{q_{max}}(x)) - \lambda] \tag{3.13}$$

where $\Phi(x)$ is standard normal cdf, $Q(x)$ is t-distribution, $F_{q_{max}}(x)$ is the cdf of q_{max}, λ is the mortality risk premium. Wang (2004) concludes that the degrees of freedom are five according to the industrial empirical result for catastrophic bonds.

Since we cannot represent the $F_{q_{max}}(x)$ analytically, we use the empirical *cdf* of q_{max}, which is presented in Figure 3.9. We first set λ with an initial value λ_0. With the empirical *cdf* $F_{q_{max}}(x)$, the $F^*_{q_{max}}(x)$ can be calculated using (3.13). Then the corresponding transformed *pdf* $f^*_{q_{max}}(x)$ can be calculated. As shown in Figure 3.9, the Wang transformed $f^*_{q_{max}}(x)$ shifts to the right with more weights on mortality shocks, specially for the extreme values. The lower graph is with the whole domain, while the upper graph is with up to 99.8% quantile for a better view.

With N simulation paths for q_{max}, the percentages $P^{(1)}, P^{(2)}, \cdots, P^{(N)}$ can be calculated using (3.12). Then using (3.10) the discounted cash flow can be obtained for each path: $DCF^{(1)}, DCF^{(2)}, \cdots, DCF^{(N)}$. Then the transformed APV of the mortality bond can be calculated using

$$APV(\lambda_0) = \sum_{i=1}^{N} DCF^{(i)} f^*_X(x_i).$$

Finally, set $APV(\lambda) = F$ to search for the numerical solution of λ. The market price of risk for our model is estimated to be 0.4085.

Table 3.6 compares the market price of risk of our model with other models

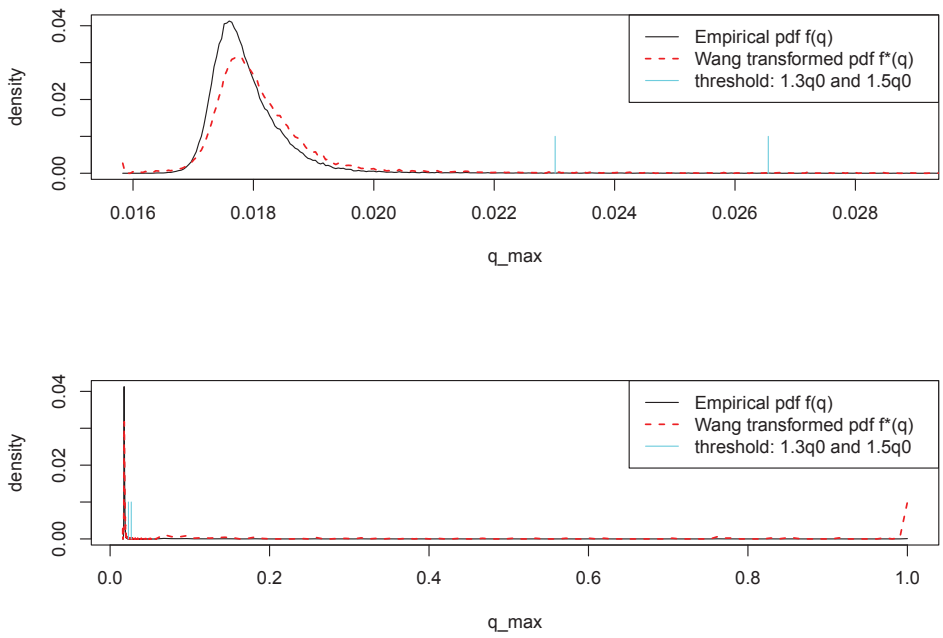

Figure 3.9: Empirical probability distribution $f_X(x)$ and Wang transformed probability distribution $f_X^*(x)$ with 50000 simulation paths

discussed in Section 3.5. Our result is close to Wang's empirical result which is based on twelve catastrophic bonds. On the other hand, Lin and Cox (2008) and Milidonis et al. (2011)[6] both have a larger market price of risk by underestimating shocks, since these two models are based on the normal distribution to control shock effects.

[6]Use R package *RHmm*.

Table 3.6: Comparison of market price of risk λ among different models

Model	Market Price of Risk λ
SSM mortality rate model	0.4085
Lin and Cox (2008)	1.3603
Milidonis et al. (2011)	0.84
Wang (2004)	0.45

Conclusion

This thesis applies non-Gaussian state space models to fit the mortality index of the Swiss Re bond. The best model includes an ARIMA(2,1,0) process in the state equation and a t-distribution error term in the observation equation. The degrees of freedom of the t-distribution are 1.3, which is a fat-tailed distribution and implies the occurrence of mortality shocks. This can also be confirmed based on the outlier detection, from which five mortality shocks are detected, including the huge shock caused by the Spanish flu pandemic in 1918. Our model provides a good fit, passes diagnostic tests, and has nicer properties compared with other mortality models in the literature. The market price of risk under this model is estimated to be 0.4 by Wang's transform, close to the industrial empirical result based on twelve catastrophic bonds given by Wang (2004).

Bibliography

Anderson, T. W. *An introduction to Multivariate Statistics Analysis (2nd Edition)*. Wiley, 1984.

Chen, C. and Liu, L. M. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297, 1993.

Cryer, J. D. and Chan, K. S. *Time Series Analysis with Application in R (2nd Edition)*. Springer, 2008.

De Jong, P. and Shephard, N. The simulation smoother for time series models. *Biometrika*, 82(2):339–350, 1995.

Deng, Y., Brockett, P. L., and MacMinn, R. D. Longevity/mortality risk modeling and securities pricing. *Journal of Risk and Insurance*, 79(3):697–721, 2012.

Durbin, J. and Koopman, S. J. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, 1997.

Durbin, J. and Koopman, S. J. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

Durbin, J. and Koopman, S. J. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616, 2002.

Girosi, F. and King, G. Understanding the Lee-Carter mortality forecasting method. 2007.

Huynh, A., Browne, B., and Bruhn, A. Catastrophic mortality bonds: Analysing basis risk and hedge effectiveness. 2012.

Jungbacker, B. and Koopman, S. J. Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika*, 94(4):827–839, 2007.

Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Keenan, D. M. A time series analysis of binary data. *Journal of the American Statistical Association*, 77(380):816–821, 1982.

Kitagawa, G. Non-Gaussian state space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987.

Kitagawa, G. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

Krutov, A. *Investing in Insurance Risk: Insurance-Linked Securities - A Practitioner's Perspective*. Riskbook, 2010.

Lee, R. D. and Carter, L. R. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

Li, J. S. H., Chan, W. S., and Cheung, S. H. Structural changes in the Lee-Carter mortality indexes: Detection and implications. *North American Actuarial Journal*, 15(1):13–31, 2011.

Li, S. H. and Chan, W. S. Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. *Scandinavian Actuarial Journal*, 2005(3): 187–211, 2005.

Li, S. H. and Chan, W. S. The Lee-Carter model for forecasting mortality, revisited. *North American Actuarial Journal*, 11(1):68–89, 2007.

Lin, Y. and Cox, S. H. Securitization of catastrophe mortality risks. *Insurance: Mathematics and Economics*, 42(2):628–637, 2008.

Mächler, M. and Bühlmann, P. Variable length Markov chains: Methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13 (2):435–455, 2004.

Milidonis, A., Lin, Y., and Cox, S. H. Mortality regimes and pricing. *North American Actuarial Journal*, 15(2):266–289, 2011.

Morens, D. M., Taubenberger, J. K., Folkers, G. K., and Fauci, A. S. Pandemic influenza's 500th anniversary. *Clinical Infectious Diseases*, 51(12):1442–1444, 2010.

Wang, S. S. A universal framework for pricing financial and insurance risks. *ASTIN Bulletin*, 32(2):213–234, 2002.

Wang, S. S. Cat bond pricing using probability transforms. *Geneva Papers*, 278: 19–29, 2004.

Ward, E. J., Hilborn, R., Towell, R. G., and Gerber, L. A state-space mixture approach for estimating catastrophic events in time series data. *Canadian Journal of Fisheries and Aquatic Sciences*, 64(6):899–910, 2007.

Zhen, X. and Basawa, V. Observation-driven generalized state space models for categorical time series. *Statistics and Probability Letters*, 79(24):2462–2468, 2009.

Appendix A

The Simulation Smoother

Lemma A.1. *Under the model (2.21), $Z_t\,\tilde{\boldsymbol{x}}_t + \tilde{\varepsilon}_t = y_t$ for $t = 1, 2, \ldots, n$.*

Proof. Using the Kalman smoother,

$$\tilde{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_t + \hat{V}_t(Z_t'\,F_t^{-1}\,v_t + L_t'\,r_t), \qquad \tilde{\varepsilon}_t = H_t(F_t^{-1}\,v_t - K_t'\,r_t),$$

then

$$\begin{aligned} Z_t\tilde{\boldsymbol{x}}_t + \tilde{\varepsilon}_t &= Z_t\,\hat{\boldsymbol{x}}_t + (Z_t\,\hat{V}_t\,Z_t' + H_t)F_t^{-1}\,v_t + (Z_t\,\hat{V}_t\,L_t' - H_t\,K_t')r_t \\ &= (Z_t\,\hat{\boldsymbol{x}}_t + v_t) + (Z_t\,\hat{V}_t\,L_t' - H_t\,K_t')r_t. \end{aligned}$$

Using the Kalman filter,

$$Z_t\,\hat{\boldsymbol{x}}_t + v_t = y_t,$$

and

$$\begin{aligned} Z_t\,\hat{V}_t\,L_t' - H_t\,K_t' &= Z_t\,\hat{V}_t(T_t' - Z_t'\,K_t') - H_t\,K_t \\ &= Z_t\,\hat{V}_t\,T_t' - (Z_t\,\hat{V}_t\,Z_t' + H_t)K_t' \\ &= Z_t\,\hat{V}_t\,T_t' - F_t[(F_t^{-1})'\,Z_t\,\hat{V}_t'\,T_t']. \end{aligned}$$

Since F_t and $\hat{V}_t$ are both covariance matrices, we have $\hat{V}_t' = \hat{V}_t$ and $(F_t^{-1})' = F_t^{-1}$. Then,

$$Z_t\,\hat{V}_t\,T_t' - F_t[(F_t^{-1})'Z_t\,\hat{V}_t'\,T_t'] = 0.$$

To sum up,

$$Z_t\,\tilde{\boldsymbol{x}}_t + \tilde{\varepsilon}_t = y_t.$$

□

Lemma A.2. *Under the model in* (2.21), $y_t = Z_t\,\check{\boldsymbol{x}}_t + \check{\varepsilon}_t = \check{\theta}_t + \check{\varepsilon}_t$, *for* $t = 1, 2, \ldots, n$.

Proof. By the definition of $\check{\theta}_t$ shown in Algorithm 2.1, as well as the definition of $\check{\varepsilon}_t$, we have

$$\begin{aligned}\check{\theta}_t + \check{\varepsilon}_t &= [\tilde{\theta}_t + (\theta_t^+ - \tilde{\theta}_t^+)] + [\tilde{\varepsilon}_t + (\varepsilon_t^+ - \tilde{\varepsilon}_t^+)] \\ &= (\tilde{\theta}_t + \tilde{\varepsilon}_t) + (\theta_t^+ + \varepsilon_t^+) - (\tilde{\theta}_t^+ + \tilde{\varepsilon}_t^+).\end{aligned}$$

By Lemma A.1, $\tilde{\theta}_t + \tilde{\varepsilon}_t = y_t$. According to the second step in Algorithm 2.1, $\theta_t^+ + \varepsilon_t^+ = y_t^+$. And also according to the Lemma A.1, saying that the sum of the smoothing values of θ_t and ε_t is equal to y_t, we have

$$\tilde{\theta}_t^+ + \tilde{\varepsilon}_t^+ = y_t^+.$$

Therefore,

$$\check{\theta}_t + \check{\varepsilon}_t = y_t + y_t^+ - y_t^+ = y_t.$$

□