

# Construction an Informative Prior Distribution of Noise in Seasonal Adjustment

Linyi Guo

Department of Mathematics and Statistics  
Faculty of Science  
University of Ottawa

# Abstract

Time series data is very common in our daily life. Since they are related to time, most of them show a periodicity more or less. The existence of this periodic influence leads to our research problem, named seasonal adjustment. Seasonal adjustment is generally applied around us, especially in areas of economy and finance. There are many researches and methods aiming at this problem, and two of the most widely used are X-13ARIMA-SEATS and TRAMO-SEATS, which are both mainly based on ARIMA models. Simultaneously, state-space modelling is also a popular method and some researchers, like Durbin, Koopman and Harvey, have contributed a lot of work in this area. In this paper, we are going to apply state-space modelling to some simulated time series data at first and then explore methods to improve the accuracy of the results from state-space models.

# Dedication

To mum and dad


# Declaration

I declare that..

# Acknowledgements

I want to thank...

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
<b>2</b>		<b>10</b>
<b>3</b>	<b>State space modelling and the Kalman filter</b>	<b>11</b>
3.1	Introduction to state space modelling . . . . .	11
3.2	Common state space models . . . . .	13
3.2.1	Structural time series models . . . . .	13
 3.2.2	ARIMA models . . . . .	15
3.2.3	Regression models . . . . .	16
3.3	The Kalman filter . . . . .	17
3.3.1	Filtering process . . . . .	17
3.3.2	Smoothing process . . . . .	18
3.4	Comparison of SSMs and conventional methodologies . . . . .	18
<b>4</b>	<b>Bayesian analysis</b>	<b>19</b>
4.1	Introduction . . . . .	19
4.2	Behaviour of maximum likelihood estimators . . . . .	20
4.3	Loss function and the derivative-free optimization algorithm . . . . .	21
4.4	Weakly-informative prior distribution . . . . .	24
4.5	Empirical prior distribution . . . . .	26
4.6	Transition of the empirical prior . . . . .	29
<b>5</b>	<b>Conclusion</b>	<b>31</b>
<b>A</b>	<b>Kalman filter</b>	<b>32</b>
A.1	Filtering process . . . . .	32
A.2	Smoothing process . . . . .	33
<b>B</b>	<b>Other supplement</b>	<b>34</b>

# List of Figures

1.1	Observed Data Distribution . . . . .	8
3.1	State space models . . . . .	12
4.1	Decomposition comparison between X-11 and SSM(MLE) . . . . .	21
4.2	Decomposition comparison among X-11, SSM(MLE) and SSM(1,1,1) . . . . .	21
4.3	Decomposition comparison between X-11 and SSMs . . . . .	23
4.4	Decomposition comparison between X-11 and SSMs from 2000 to 2004 . . . . .	23
4.5	Distributions of variance estimators . . . . .	24
4.6	Comparison of variance distributions . . . . .	25
4.7	Boxplots of decomposition errors . . . . .	26
4.8	Densities of decomposition errors . . . . .	26
4.9	Comparison of variance distribution(2) . . . . .	27
4.10	One failed case of the empirical prior . . . . .	27
4.11	Empirical distribution of variances . . . . .	28
4.12	Improvement of the last failed case . . . . .	28
4.13	Posterior distributions with different k . . . . .	29
4.14	Log likelihood under different length and variances . . . . .	30
B.1	Comparison of decomposition(Unemployment from 1990 to 2016 in U.S.) . . . . .	34
B.2	Boxplot comparison of decomposition errors . . . . .	35
B.3	Empirical distributions of the irregular variance from 8 groups . . . . .	36
B.4	Empirical distributions of the trend variance from 8 groups . . . . .	36

# List of Tables

3.1	Dimensions of notations . . . . .	12
4.1	Trend and Seasonal components' error(unemployment) . . . . .	22
4.2	Information of decomposition error . . . . .	25
B.1	Information of decomposition error(2) . . . . .	34
B.2	Information of SSMs used for simulation . . . . .	36



# Chapter 1

## Introduction

Seasonal adjustment is widely applied around us. In the area of economics, people need to deal with numerous time series data almost everyday. Generally, one time series data could be decomposed into different components, such as the trend, the seasonal and the irregular series. In signal processing of engineering, we also call them signals. Due to the influence of seasonal movements and some other calendar effects like the Christmas, the Easter and the Chinese New Year, the raw data is usually hard to utilize for analysis directly. Therefore, removing those undesired signals is important for our analysis. Meanwhile, to obtain a good forecasting, ensuring an accurate decomposition of the data is also necessary.

To illustrate the significance of the decomposition, we suppose the distribution of the observed data is

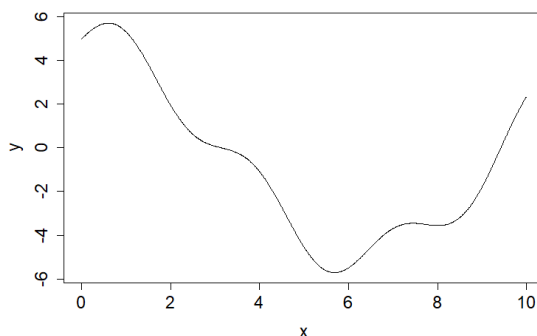


Figure 1.1: Observed Data Distribution

It seems that there is barely no pattern behind it, and the only reasonable prediction we could make is that it would increase later. But actually the raw data is simulated from the function

$$y = \sin 2x + 5\cos \frac{x}{2} \quad (1.1)$$

which means if we could find the expression of these two parts, then our prediction would be perfect! Therefore, if we could find the accurate expression of our components, it would not only help us with analysis but also for the future prediction.

The history of seasonal adjustment problem could be traced back to 1960s, at which time the first method X-11 was proposed by Statistics Canada.[reference, ]



Then U.S. Census Bureau developed X-12-ARIMA[reference, ] based on the previous one. Almost at the same time, the Bank of Spain came up with an ARIMA model-based method called TRAMO-SEATS [reference, ], which is used widely in official statistics. In 2007, U.S. Census Bureau brought up X-13ARIMA-SEATS [reference], which combined the previous work together and is another method applied widely in official statistics.

Generally speaking, because of the existence of outliers, calendar effects and other factors, we usually need to preprocess our raw data at first when dealing with seasonal adjustment problem. In both widely used methods TRAMO-SEATS and X-13ARIMA-SEATS, they use ARIMA model-based methods, TRAMO and RegARIMA, to achieve the preprocessing process separately. The next step is to decompose the time series obtained in the last step. For TRAMO-SEATS, this is handled by SEATS. For X-13ARIMA-SEATS, you can either choose X-11 or SEATS to decompose. But users need to notice that X-11 is a non-parametric universal method which use linear filters to decompose the time series data. The theories behind these methods are given in Chapter 2.

However, state space modelling is also an efficient way to solve seasonal adjustment problem. We have mentioned the components in seasonal adjustment are also referred to as signals, so basically the method used for signal extraction could be applied for seasonal adjustment as well. Specifically, the Kalman filter, which is based on the state-space model is one of the commonest methods and we shall mainly utilize it in this paper. Compared with the filter-based method and the ARIMA model-based method, state-space modelling is more general and allows more underlying improvement in the future.[reference, Bell, Himmer, 1984] And this is the reason why we explore it instead of sticking with the X-11 or other ARIMA model-based methods. But due to its complexity, it hasn't been applied widely in the area of official statistics and industries. Chapter 3 will give a general introduction of SSM and the Kalman filter, where we will also explain the advantages of SSMs in the end.

In Chapter 4, we shall mainly talk about our contribution to how to generate the satisfying decomposition results by SSMs. Over the last few decades, methods such as X-11 and SEATS have been used widely in government departments and statistics agencies to deal with seasonal adjustment problem and the result is good and convincing. Based on this fact, the main purpose of this paper is to explore how to use SSMs to generate the similar decomposition result compared with X-11's. Specifically, we will first show the deficiency of MLEs and then introduce a method about how to obtain a satisfying decomposition result with the help of X-11. Nevertheless, since our goal is to apply state space models instead of X-11 or other ARIMA-model-based methods, we come up with an empirical-Bayesian-based method to get rid of X-11 and talk about the generalization problem in the end.



## Chapter 2

# Chapter 3

## State space modelling and the Kalman filter

State space modelling was originated in 1960s. And in 1960, R.E.Kalman brought up the Kalman filter for the first time.[reference, kalman 1960] The state space model, also known as the hidden Markov model [reference, Rabiner 1989], is a powerful modelling method and applied widely in engineering, statistics, economics and etc. We shall introduce this model and explain it with some examples in section 3.1.

For SMMs or HMMs, there are many methods to extract hidden states from our observations and the Kalman filter is the most widely used one. The first two sections will introduce state space modelling briefly and give some examples. Section 3.3 shows the theory of how the Kalman filter works given a common state space model, which is also what we used in our research. Different from the Kalman filter, the particle filter is a more general method for SMMs, which will not be covered in this paper, refer to [reference, Robert, Casella, 2004] if interested. Section 3.4 will give a detailed comparison of different methods, which is also to explain why we would like to try SSMs.

### 3.1 Introduction to state space modelling

State space modelling was first proposed to solve the problems in the area of control theory in 1960s. Then in 1980s and 1990s, with the gradual development of related theories, this model became more and more popular.

For a state space model, the observation is usually composed by one or more components, which is called state in our model. For each state space model, both of the observation and the state could be multivariate or univariate. But in practice, at least in seasonal adjustment, we usually deal with cases in which the observation is univariate and the state space is multivariate. In SSMs, states are usually unobserved, and this is the reason why we call SSMs as hidden Markov models. In general cases, what we know about the whole system are our observed measurements, the relation between observations and states, and the relation of two adjacent states.

Figure 3.1 illustrates the pattern of SMMs vividly. In this figure,  $y_{0:T}$  is the observation and  $x_{0:T}$  is our hidden state, which behaves as a Markov chain, that is, the current state only depends on the last state. Generally, we use two equations as

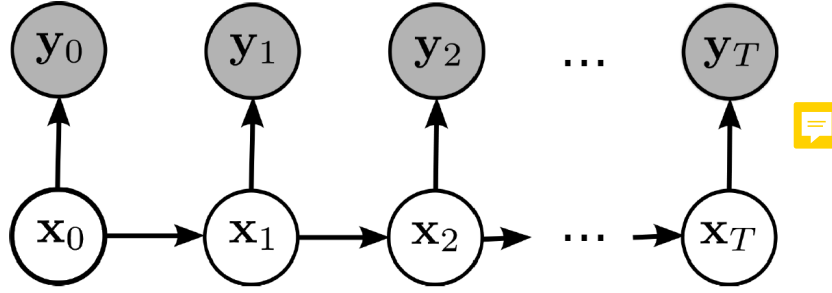


Figure 3.1: State space models

following to express a linear gaussian state space model:

$$y_t = Z_t X_t + \epsilon_t \quad \epsilon_t \sim N(0, H_t) \quad (3.1)$$

$$X_{t+1} = T_t X_t + R_t \eta_t \quad \eta_t \sim N(0, Q_t) \quad (3.2)$$

where  $t = 1, \dots, n$ , and  $X_1 \sim N(a_1, P_1)$ . Equation 3.1 is called the *measurement equation* and equation 3.2 is called the *transition equation*.  $Z_t$  reflects the relation between observations and states at time  $t$ ,  $T_t$  is the transition matrix of states from time  $t$  to  $t+1$ . Suppose the dimension of our observation is  $p \times 1$  and state is  $m \times 1$ , then dimensions of above matrices are given in the table 3.1.

Vector	Dimension	Matrix	Dimension
$y_t$	$p \times 1$	$Z_t$	$p \times m$
$X_t$	$m \times 1$	$T_t$	$m \times m$
$\epsilon_t$	$p \times 1$	$R_t$	$m \times r$
$\eta_t$	$r \times 1$	$H_t$	$p \times p$
		$Q_t$	$r \times r$

Table 3.1: Dimensions of notations

**Example 3.1.1.** In chapter 2, we have talked about the ARIMA models. Here we will show how to transform a  $AR(2)$  model to a state space form at first and then introduce the SSM form for  $AR(p)$  models.

Suppose our model is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \quad (3.3)$$

where  $\epsilon_t \sim NID(0, \sigma^2)$ , then we may find a new observation is related to the previous two values, therefore when defining this state space model, the transition equation 3.2 should have at least two states to achieve iterations.

Based on this, we will get the following result

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} x_t \quad (3.4)$$

$$x_t = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} x_{t-1} + \omega_t \quad (3.5)$$

where  $x_t = [y_t \ y_{t-1}]^T$  and  $\omega_t = [\epsilon_t \ 0]^T$ .

More generally, suppose our model is  $AR(p)$ , that is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad (3.6)$$

where  $\epsilon_t \sim NID(0, \sigma^2)$ . Then the state space form will be

$$y_t = [1 \ 0 \ \cdots \ 0] x_t \quad (3.7)$$

$$x_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} x_{t-1} + \omega_t \quad (3.8)$$

where  $x_t = [y_t \ y_{t-1} \ \cdots \ y_{t-p+1}]_{1 \times p}^T$  and  $\omega_t = [\epsilon_t \ 0 \ \cdots \ 0]_{1 \times p}^T$ .

## 3.2 Common state space models


Since state space modelling is a general method, many different models could be transformed into state space forms. Durbin and Koopman have detailedly showed in their book[reference, book 2012], so here we shall briefly introduce three common models.

### 3.2.1 Structural time series models

The structural time series model is the main one we used in SSM due to its structural characteristic. For one time series data, if we model it as a combination of the trend, seasonal, cycle and irregular components, then we call it a structural time series model. But in many research papers, the cycle component is combined with the trend, and we shall take the same strategy in this paper. The structural time series model is usually written in two ways:

$$y_t = T_t + S_t + I_t \quad (3.9)$$

$$y_t = T_t \times S_t \times I_t \quad (3.10)$$

where  $T_t$ ,  $S_t$  and  $I_t$  stand for the *trend*, *seasonal* and *irregular* components. The series without seasonal part is called *seasonally adjusted series*. If the model is multiplicative, then we usually take the *log* thm before transforming.

The simplest case is the *local level model*, where we do not have any seasonal or other explanatory variables:

$$\begin{aligned} y_t &= T_t + \varepsilon_t \\ T_{t+1} &= T_t + \eta_t \end{aligned} \quad (3.11)$$

where  $\varepsilon_t \sim NID(0, \sigma_y^2)$  and  $\eta_t \sim NID(0, \sigma_T^2)$ . If we add a slope to the trend component, the model will be:

$$\begin{aligned} y_t &= T_t + \varepsilon_t \\ T_{t+1} &= T_t + v_t + \eta_t \\ v_{t+1} &= v_t + \zeta_t \end{aligned} \quad (3.12)$$

we call it *the local linear trend model*, which is what we will use to replace the trend in equation 3.9 later. As for the seasonal component, the simple way to model it is:

$$S_{t+1} = - \sum_{j=1}^{s-1} S_{t+1-j} + \omega_t \quad (3.13)$$

where  $\omega_t \sim NID(0, \sigma_S^2)$  and  $s$  is the seasonal frequency of our data, that is, for weekly and monthly data,  $s = 7$  and  $12$  separately. But sometimes people prefer to use the trigonometric form to express seasonal components:

$$\begin{aligned} S_t &= \sum_{j=1}^{[s/2]} (\tilde{S}_{jt} \cos \lambda_j t + \tilde{S}_{jt}^* \sin \lambda_j t) \\ \tilde{S}_{j,t+1} &= \tilde{S}_{jt} + \tilde{\omega}_{jt} \\ \tilde{S}_{j,t+1}^* &= \tilde{S}_{jt}^* + \tilde{\omega}_{jt}^* \end{aligned} \quad (3.14)$$

where  $\lambda_j = \frac{2\pi j}{s}$ ,  $j = 1, \dots, [s/2]$  and  $\tilde{\omega}_{jt}, \tilde{\omega}_{jt}^*$  are normally and independently distributed variables with variance  $\sigma_\omega^2$ . [reference, Young, Lane, Ng and Palmer 1991] Generally the irregular component in equation 3.9 is treated as a normally-distributed noise directly.

Therefore, if we combine the local linear trend 3.12 and the seasonal 3.13, then we could obtain the following state space form: (we could achieve a similar but easier process for the local level model with seasonal, which is what we used in this paper and introduced in Chapter 4)

$$\begin{aligned} y_t &= T_t + S_t + I_t \\ T_{t+1} &= T_t + v_t + \eta_t \\ v_{t+1} &= v_t + \zeta_t \\ S_{t+1} &= - \sum_{j=1}^{s-1} S_{t+1-j} + \omega_t \end{aligned} \quad (3.15)$$

In terms of the equations 3.1 and 3.2, we could get the following expression:

$$\begin{aligned} X_t &= [T_t \quad v_t \quad S_t \quad S_{t-1} \quad \cdots \quad S_{t-s+2}]^T \\ Z_t &= [Z_{[T]} \quad Z_{[S]}] \\ T_t &= \text{diag} [T_{[T]} \quad T_{[S]}] \\ R_t &= \text{diag} [R_{[T]} \quad R_{[S]}] \\ Q_t &= \text{diag} [Q_{[T]} \quad Q_{[S]}] \end{aligned} \quad (3.16)$$

where

$$\begin{aligned}
 Z_{[T]} &= [1 \quad 0] & Z_{[S]} &= [1 \quad 0 \quad \cdots \quad 0]_{s-1} \\
 T_{[T]} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} & T_{[S]} &= \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \\
 R_{[T]} &= I_2 & R_{[S]} &= [1 \quad 0 \quad \cdots \quad 0] \\
 Q_{[T]} &= \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix} & Q_{[S]} &= \sigma_\omega^2
 \end{aligned}$$

### 3.2.2 ARIMA models

We have introduced the ARIMA model in [refer to arima section] and showed an AR(2) example in Section 3.1 above. In this section, we will show how to transform an arbitrary ARIMA model into state space form.

When encountering a stationary time series data, not only could we model it into an ARMA model but also into a state space model. Suppose we now have an ARMA(p,q) model

$$\begin{aligned}
 y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \\
 &= \sum_{i=1}^p \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \\
 &= \sum_{i=1}^r \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{r-1} \theta_j \varepsilon_{t-j}
 \end{aligned} \tag{3.17}$$

where  $r = \max(p, q + 1)$  and  $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$ . To transform it into state space form, we can define the measurement equation as

$$\begin{aligned}
 y_t &= [1 \quad 0 \quad \cdots \quad 0] x_t \\
 \text{where } x_t &= \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_r y_{t-r+1} + \theta_1 \varepsilon_t + \cdots + \theta_{r-1} \varepsilon_{t-r+2} \\ \phi_3 y_{t-1} + \cdots + \phi_r y_{t-r+2} + \theta_2 \varepsilon_t + \cdots + \theta_{r-1} \varepsilon_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \theta_{r-1} \varepsilon_t \end{pmatrix}
 \end{aligned} \tag{3.18}$$

The matrices in the transition equation are:

$$T_t = T = \begin{bmatrix} \phi_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \phi_{r-1} & 0 & \cdots & 1 \\ \phi_r & 0 & \cdots & 0 \end{bmatrix} \quad R_t = R = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} \tag{3.19}$$

By 3.18 and 3.19, we have the capacity to transform every known ARMA model to a corresponding state space model. Similarly, we could put any ARIMA model into a SSM, see [reference, Durbin and Koopman 2012 section 3.4].



Therefore, mathematically we are able to transform every ARIMA and ARMA model to a state space form, which just confirms that state space modelling is a more general and practical method. Simultaneously, with the development of techniques in SSM, we could handle these ARIMA models better. On the other hand, many but not all state space models have their corresponding ARIMA models. Example 3.2.1 is one simple case and more related work could be referred to [reference harvey 1989].

**Example 3.2.1.** In the local linear trend model 3.12, if we take two difference of observations, we shall get

$$\Delta^2 y_t = \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} + \eta_{t-1} - \eta_{t-2} + \zeta_{t-2}$$

It is not hard to notice only the first two autocorrelations are nonzero, so we can use a  $MA(2)$  model to express the right hand side equivalently, that is

$$\Delta^2 y_t = \delta_t + \theta_1^* \delta_{t-1} + \theta_2^* \delta_{t-2}$$

which is the expression of one  $ARIMA(0,2,2)$  model.

We have been aware of the relation between ARIMA modelling and state space modelling. In the example above, although we did transform the local linear trend model to an ARIMA model, the information with regard to the slope  $v_t$  and the level/trend  $T_t$  is lost in this process. And this is the reason why we would like to apply the structural time series SSM in our research instead of the ARIMA model-based methods.

### 3.2.3 Regression models

The regression model is one of the most fundamental concepts in statistics. The interesting thing is if we consider the *measurement equation* ignoring the subscript  $t$ , it is exactly a regression model, which means we could perhaps view a linear regression model as a SSM. Suppose we have a simple regression model for a univariate variable  $y$ :

$$y = X\beta + \varepsilon \quad \text{where } \varepsilon \sim N(0, H) \quad (3.20)$$

corresponding to 3.1, suppose  $t = 1, 2, \dots, n$  and  $n$ =number of measurements, then we have

$$Z_t = X_t \quad T_t = I_t \quad R_t = Q_t = 0 \quad (3.21)$$

If the coefficient  $\beta_t$  is changeable, then we could model it based on equation 3.2. For example, if each element in  $\beta$  follows a random walk, then it is the multivariate version of the transition equation in 3.11, that is

$$T_t = R_t = I_t \quad Q_t = \Sigma_t \quad (3.22)$$

where  $\Sigma$  is the diagonal variance matrix of coefficients. For regression problems, one of goals is to determine these coefficients, which is exactly what we shall compute with the above SSM. From this perspective, we can use the techniques in SSM to solve a regression problem.

### 3.3 The Kalman filter

We have talked what the state space model is and its classifications in the last two sections. The purpose of this section is to illustrate how Kalman filter works, which is one of the commonest techniques used to pick out these latent states.

In Section 3.3.1, we shall talk about how to use a forward recursive algorithm to extract states out of a series of measurements given the notation in 3.1 and 3.2. Then, Section 3.3.2 explains how to smooth the filtered states with a backward algorithm. As for the derivation details, please refer to Appendix A or [reference, durbin and koopman 2012].

#### 3.3.1 Filtering process

Before showing the filtering algorithm, we need to introduce some new notations to simplify our writing. Given a state space model as following,

$$\begin{aligned} y_t &= Z_t \alpha_t + \epsilon_t & \epsilon_t &\sim N(0, H_t) \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t & \eta_t &\sim N(0, Q_t) \end{aligned}$$

we define  $a_{t|t}$  is the expectation of  $\alpha_t$ , the state at time  $t$ , given the observations  $\{y_t\}$  until time  $t$ , and  $a_{t+1}$  is the expectation of the state at time  $t+1$ , given the information until time  $t$ ;  $P_{t|t}$  and  $P_{t+1}$  are variances of the state at time  $t$  and  $t+1$  separately given  $y_{1:t}$ ;  $v_t$  is the error between the observation  $y_t$  and the prediction  $Z_t a_t$  at time  $t$  and  $F_t$  is the variance of  $v_t$  given  $Y_{t-1}$ .

$$\begin{aligned} a_{t|t} &= E(\alpha_t | Y_t) & a_{t+1} &= E(\alpha_{t+1} | Y_t) \\ P_{t|t} &= Var(\alpha_t | Y_t) & P_{t+1} &= Var(\alpha_{t+1} | Y_t) \\ v_t &= y_t - Z_t a_t & F_t &= Var(v_t | Y_{t-1}) \end{aligned} \tag{3.23}$$

The filtering process is mainly composed by update and prediction two parts, which are as following

---

**Algorithm 1** Filtering process

---

**Require:**  $\alpha_1 \sim N(a_1, P_1)$ ,  $Y_T$  and matrices  $Z$ ,  $H$ ,  $T$ ,  $R$ ,  $Q$

**for**  $t \leftarrow 1, \dots, T$  **do**

$v_t \leftarrow y_t - Z_t a_t$

$F_t \leftarrow Z_t P_t Z_t^T + H_t$

$K_t \leftarrow T_t P_t Z_t^T F_t^{-1}$

UPDATE:

$a_{t|t} \leftarrow a_t + P_t Z_t^T F_t^{-1} v_t$

$P_{t|t} \leftarrow P_t - P_t Z_t^T F_t^{-1} Z_t P_t$

PREDICT:

$a_{t+1} \leftarrow T_t a_t + K_t v_t$

$P_{t+1} \leftarrow T_t P_t (T_t - K_t Z_t)^T + R_t Q_t R_t^T$

**end for**

---


### **3.3.2 Smoothing process**

## **3.4 Comparison of SSMs and conventional methodologies**

# Chapter 4


## Bayesian analysis

### 4.1 Introduction

In Section 3.1 and Section 3.2, we have introduced the general expression of a state space model is 

$$\begin{aligned} y_t &= Z_t X_t + \epsilon_t & \epsilon_t &\sim N(0, H_t) \\ X_{t+1} &= T_t X_t + R_t \eta_t & \eta_t &\sim N(0, Q_t) \end{aligned} \quad (4.1)$$

where  $t = 1, \dots, n$ , and  $X_1 \sim N(a_1, P_1)$ . In this chapter, we let

$$\begin{aligned} Z_t &= \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix} & T_t &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & -1 & \dots & -1 & -1 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \\ R_t &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} & Q_t &= \begin{bmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_S^2 \end{bmatrix} & H_t &= \sigma_I^2 \end{aligned} \quad \text{$$

Then we could derive the state space model applied in our research:

$$\begin{aligned} y_t &= T_t + S_t + I_t \\ T_{t+1} &= T_t + \eta_t \quad \text{$$
 \\ S\_{t+1} &= - \sum\_{j=1}^{s-1} S\_{t+1-j} + \omega\_t \end{aligned} \quad (4.2)

where  $I_t$ ,  $\eta_t$  and  $\omega_t$  are independent and identically distributed gaussian noises with mean 0 and variances  $\sigma_I^2$ ,  $\sigma_T^2$ ,  $\sigma_S^2$ . *Note:* we used two upper cases T here, the first one stands for the transition matrix in Equation 3.2, and another in Equation 4.2 is the trend component. We only use the first notation when referring to the general form of SSMs, so readers only need to remember the second meaning.

As what we have mentioned in Chapter 1, in this chapter, we shall talk about our main contributions to the seasonal adjustment problem, especially to the decomposition part. Section 4.2 will introduce the deficiency of maximum likelihood

estimators(abbreviated as MLEs) by comparing its decomposition results with the X-11's. After defining some penalty functions to reproduce the X-11's decomposition in Section 4.3, we utilize some weakly-informative priors to compute the posterior estimators and compare its result with MLEs' in Section 4.4. And then we shall make use of the prior knowledge gained from Section 4.3 to build an empirical prior for some particular data in Section 4.5 and compare this empirical posterior estimators with others. Meanwhile, we will try to approximate the empirical prior with a parametric distribution and introduce the motivation behind Section 4.6, in which we talk about the transformation of the existing empirical prior when we meet different data. In the last section, we shall compare all these methods' prediction abilities.

## 4.2 Behaviour of maximum likelihood estimators

In Chapter 3 we have seen that there are some parameters in the state space model. To apply the Kalman filter in Section 3.3, either we know these parameters' values or we estimate them. One common estimation method is the maximum likelihood estimation, but it turns out the decomposition results are not convenient for statistics agencies to analyse.

Given the notations in Section 3.3, suppose the initial distribution  $N(a_1, P_1)$  is known, the log-likelihood is expressed as

$$\log L(Y_n) = \log(p(y_1, y_2, \dots, y_n)) = \log(p(y_1) \prod_{t=2}^n p(y_t | Y_{t-1})) = \log(\sum_{t=1}^n p(y_t | Y_{t-1})) \quad (4.3)$$

where  $Y_n = y_1, \dots, y_n$  and  $p(y_1 | Y_0) = p(y_1)$ . We also know  $y_t | Y_{t-1} \sim N(Z_t a_t, F_t)$  and  $v_t = y_t - Z_t a_t$ , thus Equation 4.3 could be written as

$$\log L(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^n (\log |F_t| + v_t' F_t^{-1} v_t) \quad (4.4)$$

where  $p$  is the dimension of our state. For a univariate problem, it could be expressed as

$$\log L(Y_n) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^n \log(F_t + v_t^2 F_t^{-1}) \quad (4.5)$$

Since there are hidden variables in our problem, EM algorithm will be applied if we want to compute MLEs. We won't talk about the technical details in this process, [reference] have shown how to derive and apply it. Let's take the unemployment data of the United States from 1990 to 2016 as an example. The following figure 4.1 is the comparison of decomposition results from X-11 and the SSM with MLE:

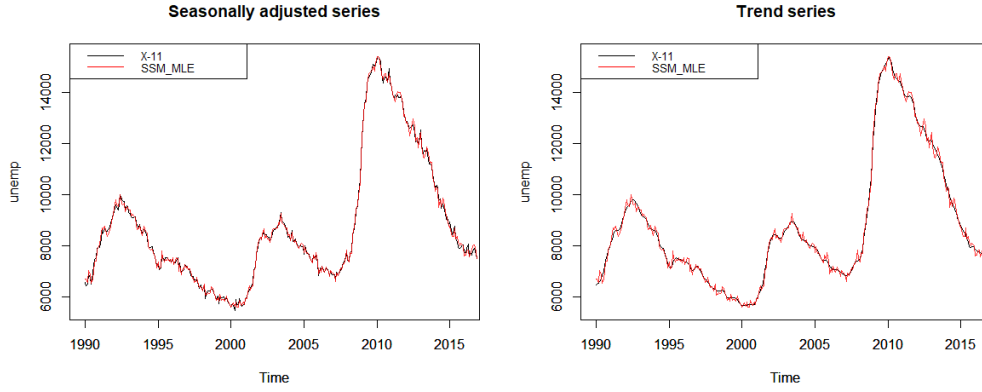


Figure 4.1: Decomposition comparison between X-11 and SSM(MLE)

As we can see, the difference between X-11 and SSM(MLE) is obvious especially for the trend component (the red line is much spikier). In economics, people would like to believe and see a relatively smooth trend instead of a spiky one, and the spiky fluctuation should be raised by seasons, holidays or other factors. So, when analysing one time series data, people mainly care about the seasonally adjusted and trend series, but now the decomposition based on maximum likelihood estimation is apparently not good enough. Actually if we let  $\sigma^2 = c(1, 1, 1)$ , decomposition results would be closer to those from X-11. See Figure 4.2.

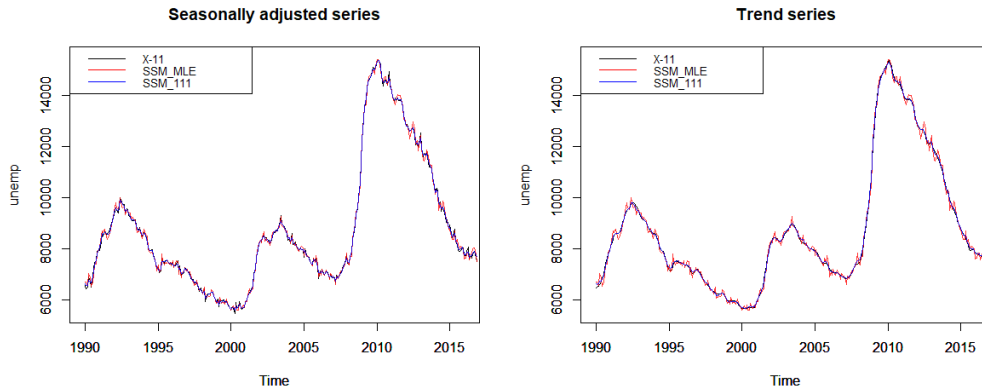


Figure 4.2: Decomposition comparison among X-11, SSM(MLE) and SSM(1,1,1)

Consequently, if we want to obtain a similar decomposition result by SSM w.r.t those from X-11, maximum likelihood estimator may not be a good choice. Hence we need to find other estimators to replace MLEs in our problem.

### 4.3 Loss function and the derivative-free optimization algorithm

Since our goal is to find parameters whose decomposition results, seasonally adjusted and trend series, are the closest to X-11 decomposition, we propose to define some loss functions based on our need and then check their real effects visually. We have mentioned that we mainly care about the seasonally adjusted and trend series in

our real life, but this argument is for the outliers-free time series data. If one time series contains some outliers and calendar effects, then we need to remove them at first, which is not we would cover in this paper. As the seasonally adjusted series is the original data minus the seasonal component, the penalty on the seasonal series equals to penalizing the seasonally adjusted series. Then we define our first loss function as:

$$L_1(\sigma^2) = \|T_{X11} - T_{SSM(\sigma^2)}\|_2^2 + \|S_{X11} - S_{SSM(\sigma^2)}\|_2^2 \quad (4.6)$$

where  $\sigma^2 = (\sigma_I^2, \sigma_T^2, \sigma_S^2)$ ,  $T_{SSM(\sigma^2)}$ ,  $S_{SSM(\sigma^2)}$  are the trend and seasonal series we obtained from the state space model with corresponding variance  $\sigma^2$ , and  $T_{X11}$ ,  $S_{X11}$  are those from X-11.

As what we proved in Section 3.3, the trend and seasonal series from the Kalman filter is obtained from a recursive process, it is hard to express the loss function with  $\sigma_I^2$ ,  $\sigma_T^2$  and  $\sigma_S^2$  explicitly, which means we can not calculate its derivative and thereby the usual optimization algorithm such as gradient descent doesn't work here. One naive way is to use the grid search to find the best value, which is also what was used at the beginning, but this method is too time-consuming. To accelerate our operation, we adopt one derivative-free optimization algorithm, *Hooke-Jeeves* algorithm [reference, 'dfoptim' package] to solve this black-box optimization problem.

After calculating regarding to the unemployment dataset, the values of parameters  $\sigma_I^2$ ,  $\sigma_T^2$  and  $\sigma_S^2$  with the lowest loss are 3.93750, 2.90625 and 1.87500, whereas the MLEs are 2.664035, 64895.19 and 0.01197881 separately if we set three initial values at zeros. As we will see in Figure 4.3, the effect of the former one is better. To force the smoothness of the trend series from SSM to be similar to that from X-11, we introduce another term with regard to the derivative of two trend series:

$$L_2(\sigma^2) = \|T_{X11} - T_{SSM(\sigma^2)}\|_2^2 + \|S_{X11} - S_{SSM(\sigma^2)}\|_2^2 + \|D(T_{X11}) - D(T_{SSM(\sigma^2)})\|_2^2 \quad (4.7)$$

where operation  $D$  takes the difference between two adjacent values. And the final best parameter values are 4.46875, 3.00000 and 2.31250. Table 4.1 gives the error between components from these three estimators and those from X-11, and Figure 4.3 shows the decomposition results visually and Figure 4.4 is the subset of components from 2000 to 2004, which is used to supply a better comparison:

	MLE	LOSS1	LOSS2
Trend	7525182	822060	823149
Seasonal	3333002	1010155	1022638

Table 4.1: Trend and Seasonal components' error(unemployment)

Table 4.1 and Figure 4.3 show that both our current decomposition fit better compared with the result from MLE, and the distinction between two loss functions is not obvious either. In the following text, we adopted  $L_2$  as our default loss function if not specified. So far we have realized how to reproduce the classical decomposition result but we also notice that our loss function still needs to rely on first fitting the dataset to X-11 or other conventional methods. We will try to use the empirical Bayesian analysis to avoid this in Section 4.5.

To accelerate and simplify our computation, we also utilize the following property to reduce the number of variance parameters from 3 to 2 in Equation 4.2:

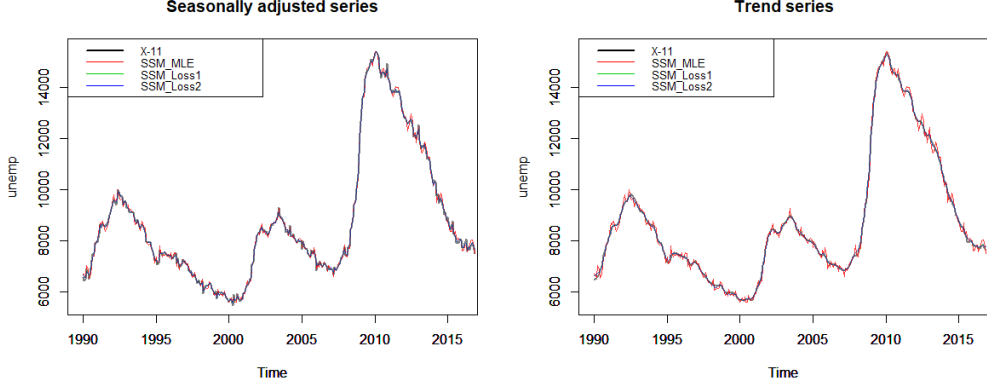


Figure 4.3: Decomposition comparison between X-11 and SSMs

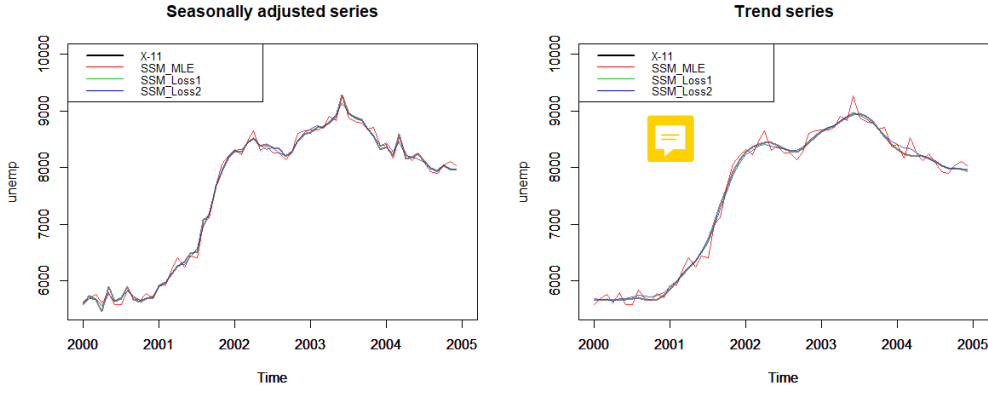


Figure 4.4: Decomposition comparison between X-11 and SSMs from 2000 to 2004

**Lemma 4.3.1.** Given the notations in Section 3.3, if  $Z_t$ ,  $T_t$ ,  $H_t$ ,  $R_t$  and  $Q_t$  are time-invariant, then the variance matrix  $P_t$  converges to a constant matrix  $\bar{P}$ , which is the solution to

$$\bar{P} = T\bar{P}T' - T\bar{P}Z'F^{-1}Z\bar{P}T' + RQR' \quad (4.8)$$

where  $\bar{F} = Z\bar{P}Z' + H$

**Lemma 4.3.2.** For the same time series data,  $\forall \lambda > 0$ , if  $(\sigma_I^2, \sigma_T^2, \sigma_S^2) = \lambda(\sigma_I^{2*}, \sigma_T^{2*}, \sigma_S^{2*})$ , then  $\text{SSM}(\sigma_I^2, \sigma_T^2, \sigma_S^2)$  and  $\text{SSM}(\sigma_I^{2*}, \sigma_T^{2*}, \sigma_S^{2*})$  will have the same decomposition results after applying the Kalman filter.

*Proof.* Take the local level model as an example: suppose we have such a local level model

$$\begin{aligned} y_t &= T_t + \varepsilon_t & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2) \\ T_{t+1} &= T_t + \eta_t & \eta_t &\sim N(0, \sigma_\eta^2) \end{aligned} \quad (4.9)$$

Let  $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}$ , corresponding to Equation 4.8, we could have

$$\bar{P} = \frac{1 + \sqrt{1 + 4\lambda}}{2} \sigma_\eta^2 \quad \bar{F} = \frac{1 + 2\lambda + \sqrt{1 + 4\lambda}}{2} \lambda \sigma_\eta^2 \quad (4.10)$$

At time  $t$ , when we update the state by equation  $x_{t|t} = x_t + \bar{P}_t Z_t^T F_t^{-1} v_t$ , we could find no matter what the values of  $\sigma_\eta^2$  and  $\sigma_\varepsilon^2$  are, as long as their ratio doesn't change,



then the result after updating stay the same. Similar things also happened in the prediction step.  $\square$

Under this lemma, we only need to care about their ratios so we could reduce our parameters by one if we are looking for the decomposition results. Without loss of generality we would let  $\sigma_S^2 = 1$  in the following analysis if we don't specify in particular.

When only dealing with one single dataset, it's not convincing to say the difference between MLEs and estimators from the loss function exists. Thus we could simulated numerous datasets from SSMs, compute the MLEs and loss-based estimators for each dataset, and then compare them. If the difference between two groups is obvious, then we have reason to believe that the MLE is truly not a good choice if we want to obtain the decomposition result similar to X-11; meanwhile, the loss-based estimators could be more like what we want.

In the following figure, we simulated 1000 monthly time series data sets at length 180(15 years) from the same state space model with variances  $\sigma_I^2 = 20$ ,  $\sigma_T^2 = 10$  and  $\sigma_S^2 = 1$  and computed the MLEs and optimal values of  $\sigma_I^2$  and  $\sigma_T^2$  w.r.t the  $L_2$ . Then we obtained their distributions:

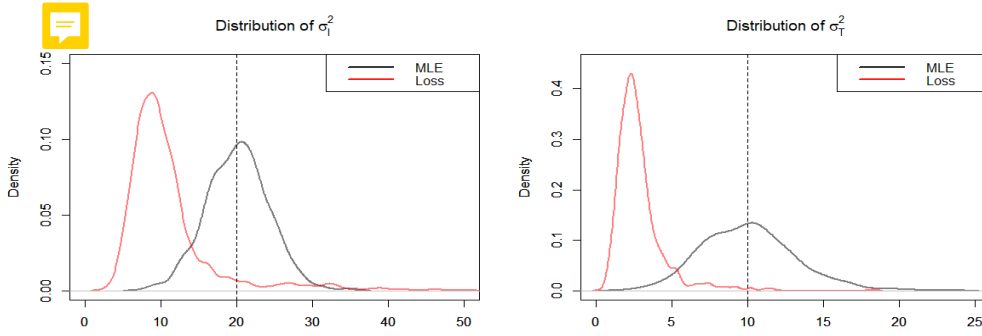


Figure 4.5: Distributions of variance estimators

As we can see, differences of estimators from two methods are prominent in this case. In fact, after a lot of simulations, we found the distributions of optimal parameters w.r.t our loss function do not change too much(See Section 4.5 and Appendix B) even though variance values  $\sigma_I^2, \sigma_T^2, \sigma_S^2$  used for simulation are very large or small, which is unlike the distributions of MLEs. Since our goal is to use SSMs to obtain the similar decomposition in terms of X-11, the next two sections basically talk about how we push the black lines(MLE) to the red line(optimal) without actually fitting an X11 model to each data.

## 4.4 Weakly-informative prior distribution

After the discussion above, we will talk about the weakly-informative prior distribution in this section. Let's briefly introduce Bayesian analysis first: we denote the prior distribution as  $g(\theta)$  and the likelihood function is  $f(Y_n|\theta)$ , then the posterior distribution is proportional to the product of them, that is:

$$g(\theta|Y_n) \propto g(\theta)f(Y_n|\theta) \quad (4.11)$$

which equals

$$\log(g(\theta|Y_n)) \propto \log(g(\theta)) + \log(f(Y_n|\theta)) \quad (4.12)$$

In this paper, we will use the maximum a posterior estimator(MAP) as the posterior estimator, whose definition is similar to the maximum likelihood estimator:

$$\begin{aligned} \theta_{MAP} &= \arg \max_{\theta} g(\theta|Y_n) \\ &= \arg \max_{\theta} g(\theta)f(Y_n|\theta) \\ &= \arg \max_{\theta} [\log(g(\theta)) + \log(f(Y_n|\theta))] \end{aligned} \quad (4.13)$$

Speaking of prior distributions, one naive thought is to define a uniform distribution on the interval with high probability for each variance such as  $[0,40]$  and  $[0,10]$ . However, this is not a wise choice since by this way we would directly abandon every point greater than upper limits and then move numerous posterior estimators(whose MLEs are greater than upper limits) to 40 and 10. The uniform distribution is not recommended. Motivated by *Gelman (2006)*, we shall use the half-normal distribution as priors for standard deviations  $\sigma_I$  and  $\sigma_T$ . The reason why we didn't adopt the half-Cauchy is that distributions of variance estimators from the loss function do not have a heavy tail. To better display the effect of the prior distribution, we used the simulated 1000 data sets before and then computed and plotted the distributions of MLEs, posterior estimators and optimal values, which is showed in Figure 4.6.

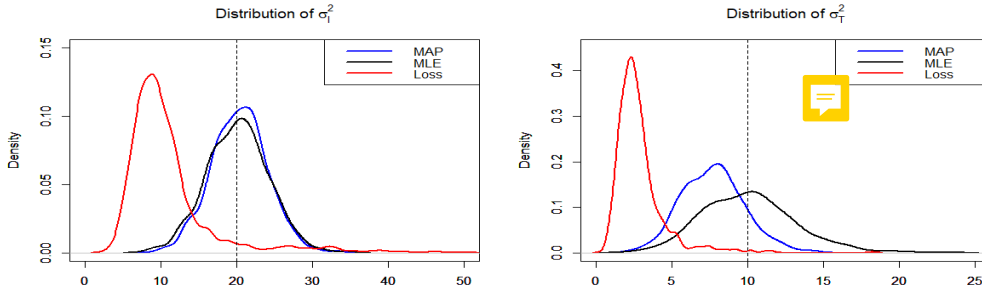


Figure 4.6: Comparison of variance distributions

The parameters we set up for half-normal distributions are  $\frac{\sqrt{40}}{3}$  and  $\frac{\sqrt{10}}{3}$  because the red line tells us the variances are less than 40 and 10 with high probability. And this is one reason why we call it a weakly-informative prior, because we only used the prior information that the variance estimators are mainly distributed over  $[0,40]$  and  $[0,10]$ . By defining the decomposition error below, we are able to compare the methods above, as is shown in Table 4.2 and Figure 4.7 and 4.8.

$$Er(\sigma^2) = \|T_{X11} - T_{SSM(\sigma^2)}\|_2^2 + \|S_{X11} - S_{SSM(\sigma^2)}\|_2^2 \quad (4.14)$$

	MLE	Loss	MAP
Median	761.9	645.2	715.2
Mean	785.2	657.1	733.3
sd	207.11	150.54	179.84

Table 4.2: Information of decomposition error

Figure 4.7: Boxplots of decomposition errors

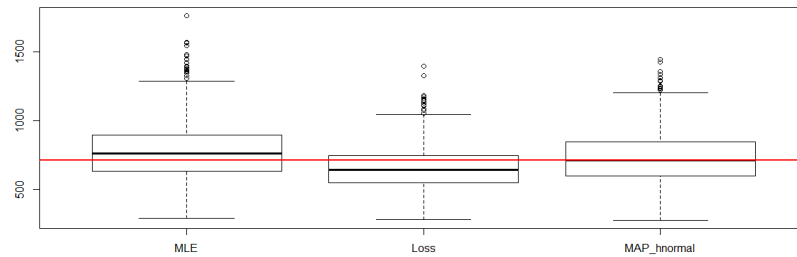
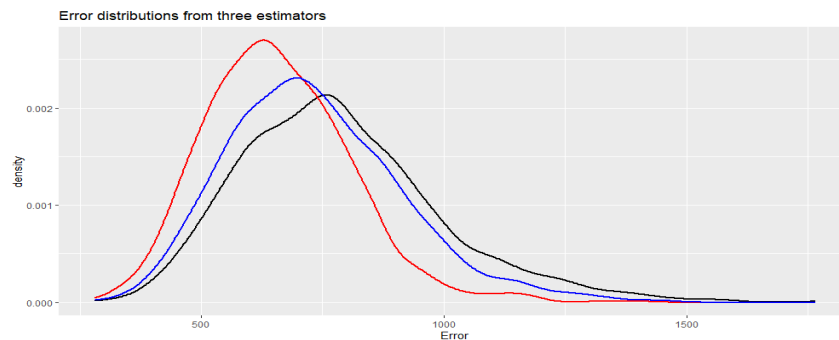


Figure 4.8: Densities of decomposition errors



Black, red and blue curves stand for MLE, Loss and MAP separately


We could tell that the maximum a posterior estimators does behave better than MLEs, and to back up our argument here we also used the *Friedman* test and *Mann-Whitney U* test to check whether the difference of errors from MLE and MAP are prominent or not. Both results showed that the difference is prominent (See Appendix B). Combined with the information above, we are confident to say our weakly-informative prior, half-normal distribution, does work.



## 4.5 Empirical prior distribution

In the last section, we have shown the weakly-informative prior does help us to have a closer decomposition to that from X-11 compared with using MLE directly. In economics, the same categorical data usually have the similar seasonal pattern like different brands of electronic products usually achieve sales peak in the winter and ice-cream manufacturers need to produce more ice-cream in the summer. Thus we have reasons to believe these similar data should share the same parameter distribution, which is called the prior distribution in Bayesian analysis. Based on this, we will treat the parameter distribution generated from loss function as an empirical prior for new datasets from a similar data-generating process. In this section, we shall first use 70% of the same data sets we used above to build the empirical prior distribution and then compare its results with other methods above with regard to the rest datasets.

Figure 4.9 shows the variance distribution comparison of different methods. And

we could tell that both distributions of posterior estimators are really closer to the variance distributions from the loss function. To make our guess more convincing, we calculated the mean and standard deviation of different decomposition error and did the same hypothesis test in Section 4.4. And all results have shown that after applying the empirical prior, generally speaking, the decomposition error would become less(see Appendix B) compared with the MLEs and posterior estimators from half-normal distributions. 

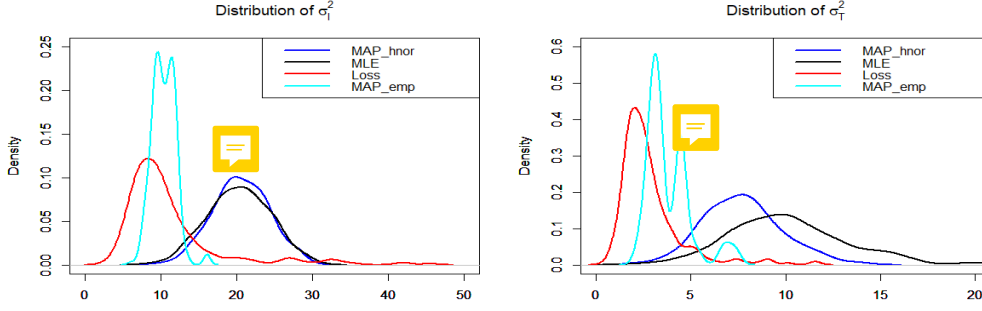



Figure 4.9: Comparison of variance distribution(2)

However, in some cases, if we treat the distribution of optimal values as the empirical prior *directly* it would collapse. As we can see in Figure 4.10, we simulated 1000 data sets at length 15 years from the SSM with variances  $\sigma_I^2 = 100$ ,  $\sigma_T^2 = 25$  and  $\sigma_S^2 = 1$ , and repeated the same process above, but the posterior estimators from the empirical prior are unsatisfying, mainly because (1) empirical priors' domains are limited instead of the whole non-negative real number; (2) they are not smooth. Besides these, the log likelihood is related to the magnitude of data and its length, so if the magnitude of the log likelihood is too large then our empirical/weakly informative prior barely change the posterior estimators. To fix this, the easiest way is to smooth the existing distribution and add a weak tail on the right, which we will show later. Another method is to use a parametric distribution to approximate the empirical distribution. 

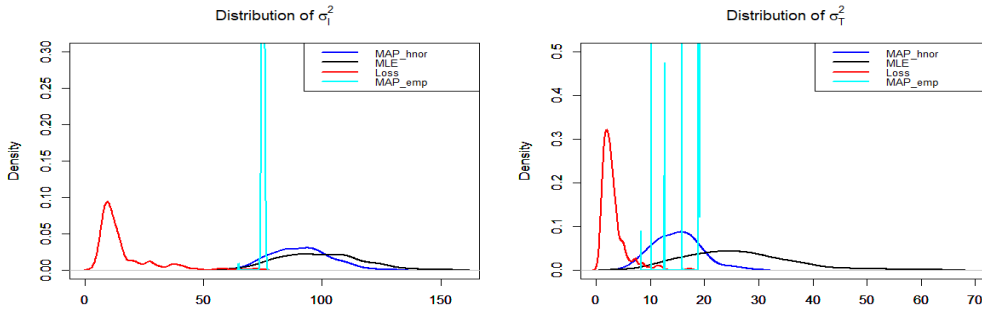


Figure 4.10: One failed case of the empirical prior

Based on Figure 4.9 and Figure 4.10, it is easy to find the distributions of maximum likelihood estimators has a close relationship with data itself, that is, different data could have totally different MLEs unless they belong to the similar categories.

On the contrary, after a bunch of simulation and analysis, we found the distributions of optimal  $\sigma_I^2$  and  $\sigma_T^2$  w.r.t the loss function defined above mainly center on intervals  $[0,50]$  and  $[0,15]$  approximately (see Appendix B). Figure 4.11 is the distributions of loss-based estimators from 8 different groups (each group contains 1000 data sets, see Appendix B for details) and their smoothed versions, which are designed to optimize the posterior distribution in Figure 4.10:

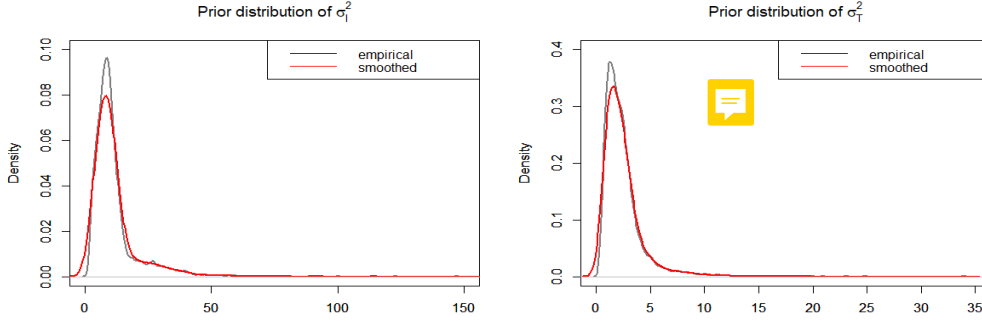


Figure 4.11: Empirical distribution of variances

In addition, we add two tails from  $N(0, \frac{50}{3})$  and  $N(0, 5)$  to both empirical priors' right side separately and use the combined distributions as empirical priors to get the corresponding posterior estimators:

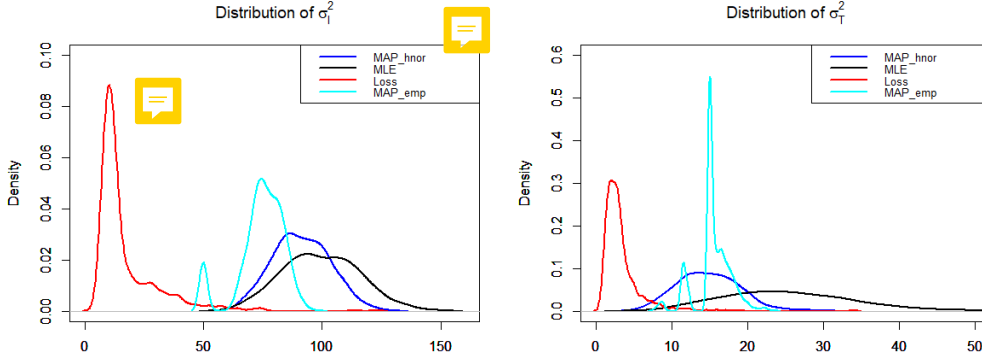


Figure 4.12: Improvement of the last failed case

The problem left now is that our MAP estimator is mainly controlled by the log likelihood since their distribution is very close to the MLE's. To force it to be closer to our prior, we tried different weights of the log prior when computing MAP estimators, that is

$$\theta_{MAP}^* = \arg \max_{\theta} [k \cdot \log(g(\theta)) + \log(f(Y_n|\theta))] \quad (4.15)$$

where  $k > 1$ . Figure 4.13 shows the distributions of MAP estimators under different  $k$ . As we shall see, with  $k$  increasing, the posterior estimators get closer to the distribution of loss-based estimators but more concentrated. We will talk more about these in Section 4.6.

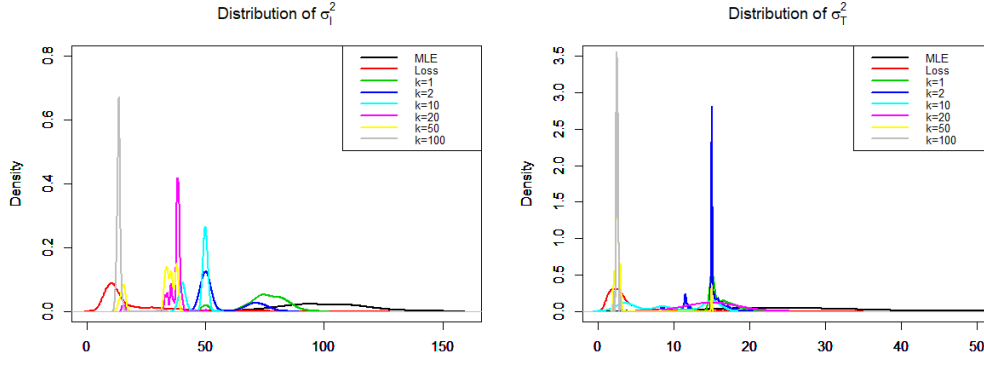


Figure 4.13: Posterior distributions with different k

## 4.6 Transition of the empirical prior

In Section 4.2, we have showed the log-likelihood of a univariate SSM could be expressed as

$$\log L(Y_n) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^n \log(F_t + v_t^2 F_t^{-1})$$

Intuitively, the log-likelihood should have an approximately linear relationship with the number of observations  $n$ . As  $n$  increases, the absolute value of log-likelihood will be greater. And we also know

$$\theta_{MAP} = \arg \max_{\theta} [\log(g(\theta)) + \log(f(Y_n|\theta))] \quad (4.16)$$

because our prior is unrelated to the sample size, as  $n \rightarrow \infty$ , the log-likelihood would dominate our choice of the posterior estimator. An illustrative example follow.

**Example 4.6.1.** Suppose  $x_1, x_2, \dots, x_n \sim B(1, \theta)$ , and the prior on  $\theta$  is  $Beta(\alpha, \beta)$ , where  $\alpha$  and  $\beta$  are constants, then the posterior distribution on  $\theta$  is  $Beta(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$ , since

$$\begin{aligned} g(\theta|X_n) &\propto g(\theta)f(X_n|\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &\propto \theta^{\alpha+\sum_{i=1}^n x_i-1}(1-\theta)^{\beta+n-\sum_{i=1}^n x_i-1} \end{aligned}$$

then after computing the derivative of  $\log(g(\theta|X_n))$ , the maximum a posterior estimator is  $\theta_{MAP} = \frac{\sum_{i=1}^n x_i + \alpha - 1}{n + \alpha + \beta - 2}$  whereas the maximum likelihood estimator is  $\theta_{MLE} = \frac{\sum x_i}{n}$ . As we could see, as  $n \rightarrow \infty$ ,  $\theta_{MAP} \rightarrow \theta_{MLE}$ .

However, as we showed in Section 4.2, maximum likelihood estimators don't fit our decomposition problem here. On the contrary, we hope our prior could play a more important role when computing the MAP estimators. Similar to the last part of Section 4.5, we add a coefficient  $k > 1$  to the log prior probability and compute the expression of the MAP estimator:

**Example 4.6.2.** Continued with Example 4.6.1, we already know

$$\theta_{MAP} = \arg \max_{\theta} [\log(g(\theta) + \log(f(X_n|\theta)))] = \frac{\sum_{i=1}^n x_i + \alpha - 1}{n + \alpha + \beta - 2}$$

Now

$$\begin{aligned} \theta_{MAP}^* &= \arg \max_{\theta} [k \cdot \log(g(\theta) + \log(f(X_n|\theta)))] \\ &= \arg \max_{\theta} [k \cdot \log(\theta^{\alpha-1}(1-\theta)^{\beta-1}) + \log(\theta^{\sum x_i}(1-\theta)^{n-\sum x_i})] \\ &= \frac{\sum_{i=1}^n x_i + k(\alpha - 1)}{n + k(\alpha + \beta - 2)} \end{aligned}$$

Because our prior is  $Beta(\alpha, \beta)$ , the mode of it is just  $\frac{\alpha-1}{\alpha+\beta-2}$ . Therefore, the more *weights* we put on the log prior, the closer our posterior estimate would be to the mode of the prior distribution if observations are fixed.

To verify our guess that the log-likelihood is a linear relation with the number of observations in SSM, we drew the plot of log likelihoods of one SSM as observations increase one by one. As a contrast, we also drew log-likelihood plots of other SSMs with proportional variances. The dataset is simulated from a SSM with variances  $\sigma_I^2 = 20$ ,  $\sigma_T^2 = 10$  and  $\sigma_S^2 = 1$ . The corresponding variances  $\sigma_I^2$ ,  $\sigma_T^2$  and  $\sigma_S^2$  of lines in Figure 4.14 from top to bottom are (20,10,1), (40,20,2), (60,30,3), (80,40,4) and (100,50,5):

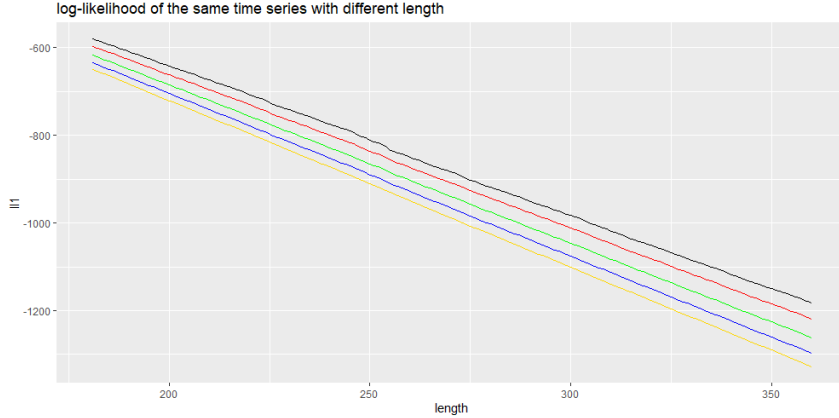


Figure 4.14: Log likelihood under different length and variances

From Figure 4.14, our intuition also tells us: if variances  $\sigma_I^2$ ,  $\sigma_T^2$  and  $\sigma_S^2$  of SSMs have the same ratio  $\lambda$ , then the log likelihood may have a linear relation with the ratio  $\lambda$ , but we won't cover this aspect in this paper. In the real life, we may have various datasets from the same category, but due to the different start date recorded in history, they often don't have the exactly same length. We have seen posterior estimators of the longer dataset will be closer to the MLEs, but the prior is one and only. Therefore, if we want to keep the posterior close to our prior for different length datasets, then the *weight*  $k$  should also be linearly dependent with the length  $n$ .

# Chapter 5

## Conclusion

conclusion



# Appendix A

## Kalman filter

Given the content in Section 3.3, we shall show how to derive the Kalman filtering step by step based on the general expression of a state space model below. The whole process could also be found in Durbin and Koopman, 2012.

$$y_t = Z_t \alpha_t + \epsilon_t \quad \epsilon_t \sim NID(0, H_t) \quad (\text{A.1})$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \quad \eta_t \sim NID(0, Q_t) \quad (\text{A.2})$$

Before giving the derivation procedure, we post a known conclusion from multivariate analysis:

**Lemma A.0.1.** Suppose X and Y are jointly normally distributed as following,

$$E[(x \ y)^T] = (\mu_x \ \mu_y)^T \quad Var \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{pmatrix} \quad (\text{A.3})$$

then the conditional distribution of X given Y is also normal with mean

$$E[x|y] = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y) \quad (\text{A.4})$$

and variance matrix

$$Var[x|y] = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{xy}^T \quad (\text{A.5})$$

### A.1 Filtering process

It's not hard to show the expectation of  $v_t$  given  $Y_{t-1}$  is 0, then with Lemma A.0.1 applying on  $\alpha_t$  and  $v_t$  given  $Y_{t-1}$ , we could show

$$a_{t|t} = E(\alpha_t | Y_{t-1}) + Cov(\alpha_t, v_t) Var(v_t)^{-1} v_t$$

where

$$\begin{aligned} Cov(\alpha_t, v_t) &= E(\alpha_t (Z_t \alpha_t + \epsilon_t - Z_t a_t)' | Y_{t-1}) \\ &= E(\alpha_t (\alpha_t - a_t)' Z_t' | Y_{t-1}) \\ &= P_t Z_t' \\ Var(v_t | Y_{t-1}) &= Var(Z_t \alpha_t + \epsilon_t - Z_t a_t | Y_{t-1}) \\ &= Z_t P_t Z_t' + H_t \\ &= F_t \end{aligned}$$

thereby,

$$a_{t|t} = a_t + P_t Z_t' F_t^{-1} v_t$$

Similarly, by Lemma A.0.1 we derive another update equation

$$\begin{aligned} P_{t|t} &= \text{Var}(\alpha_t | Y_t) \\ &= \text{Var}(\alpha_t | Y_{t-1}, v_t) \\ &= \text{Var}(\alpha_t | Y_{t-1}) - \text{Cov}(\alpha_t, v_t) \text{Var}(v_t)^{-1} \text{Cov}(\alpha_t, v_t)' \\ &= P_t - P_t Z_t' F_t^{-1} Z_t P_t \end{aligned}$$

Now let's look at how to predict the state at time  $t+1$ :

$$\begin{aligned} a_{t+1} &= E(\alpha_{t+1} | Y_t) \\ &= E(T_t \alpha_t + R_t \eta_t | Y_t) \\ &= T_t E(\alpha_t | Y_t) \\ &= T_t a_{t|t} \\ P_{t+1} &= \text{Var}(T_t \alpha_t + R_t \eta_t | Y_t) \\ &= T_t \text{Var}(\alpha_t | Y_t) T_t' + R_t Q_t R_t' \\ &= T_t P_{t|t} T_t' + R_t Q_t R_t' \end{aligned}$$

With update equations we obtained above and the Kalman gain  $K_t = T_t P_t Z_t' F_t^{-1}$ , we could have the final version of our prediction equation:

$$\begin{aligned} a_{t+1} &= T_t a_t + K_t v_t \\ P_{t+1} &= T_t P_t (T_t - K_t Z_t)' + R_t Q_t R_t' \end{aligned}$$

Sometimes  $Z_t$ ,  $T_t$ ,  $H_t$ ,  $R_t$  and  $Q_t$  are time-invariant, then we can show that the variance matrix  $P_t$  converges to a constant matrix  $\bar{P}$ , which is the solution to

$$\bar{P} = T \bar{P} T' - T \bar{P} Z' \bar{F}^{-1} Z \bar{P} T' + R Q R' \quad (\text{A.6})$$

where  $\bar{F} = Z \bar{P} Z' + H$ .

## A.2 Smoothing process

# Appendix B

## Other supplement

This is the comparison of the decomposition results from MLEs and random small variances at 1:

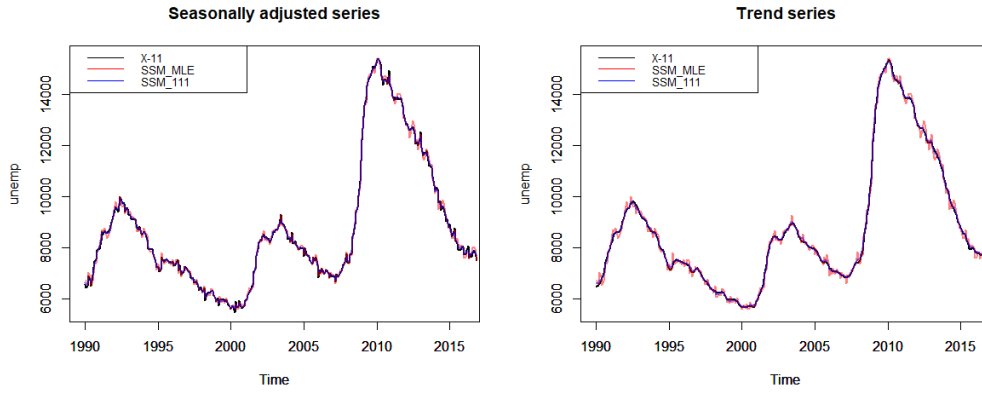


Figure B.1: Comparison of decomposition(Unemployment from 1990 to 2016 in U.S.)

The following is the hypothesis test results between the weakly-informative prior and MLEs:

*Friedman rank sum test*

*Friedman chi-squared = 583.7, df = 1, p-value < 2.2e-16*

*Wilcoxon rank sum test with continuity correction*

*W = 570603, p-value = 4.565e-08*

*alternative hypothesis: true location shift is not equal to 0*

The decomposition error comparison among MLEs, the posterior estimators from weakly-informative and empirical priors is:

	MLE	Loss	MAP(hnormal)	MAP(empirical)
Median	775.7	651.7	716.1	690.8
Mean	796.8	662.3	743.8	719.6
sd	213.4	150.9	185.5	171.4

Table B.1: Information of decomposition error(2)

This is the box plot of their decomposition errors:

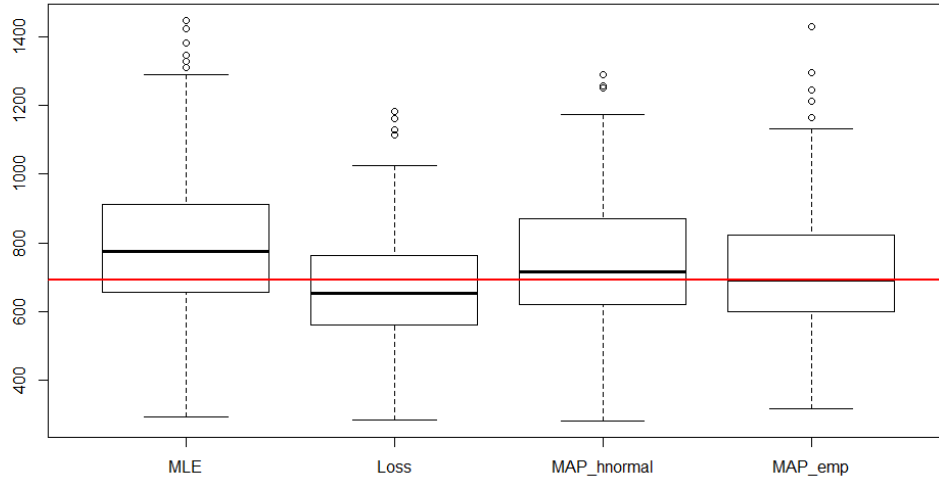


Figure B.2: Boxplot comparison of decomposition errors

Part of testing result w.r.t the posterior estimators from the empirical prior:

*Friedman rank sum test*

*data: MLE,MAP(hnormal),MAP(empirical)*

*Friedman chi-squared = 186.91, df = 2, p-value < 2.2e-16*

*Friedman rank sum test*

*data: MAP(hnormal),MAP(empirical)*

*Friedman chi-squared = 34.68, df = 1, p-value = 3.886e-09*

*Friedman rank sum test*

*data: MLE,MAP(empirical)*

*Friedman chi-squared = 75, df = 1, p-value < 2.2e-16*

*Wilcoxon signed rank test with continuity correction*

*data: MLE,MAP(empirical)*

*V = 500500, p-value < 2.2e-16*

*alternative hypothesis: true location is not equal to 0*

*Wilcoxon signed rank test with continuity correction*

*data: MAP(hnormal),MAP(empirical)*

*V = 500500, p-value < 2.2e-16*

*alternative hypothesis: true location is not equal to 0*

The following figure is the empirical distribution of variances from 8 different groups:

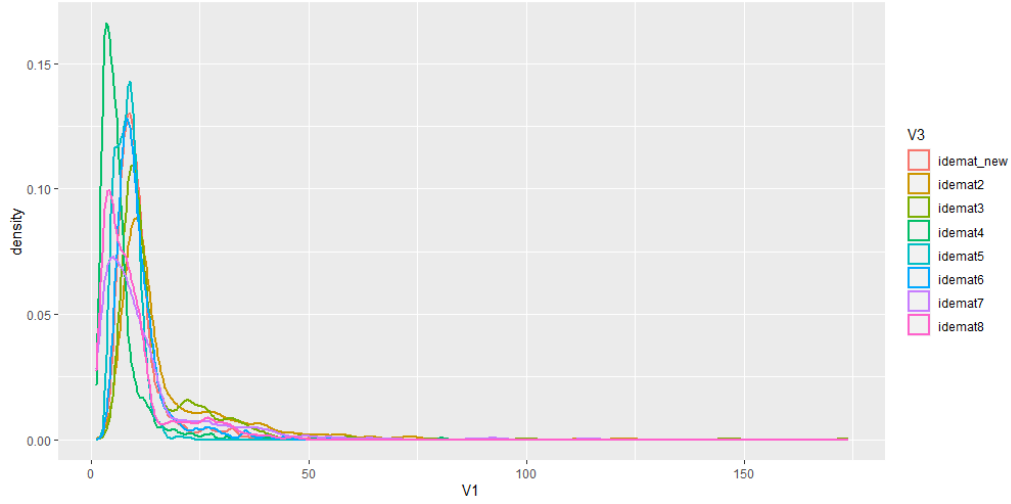


Figure B.3: Empirical distributions of the irregular variance from 8 groups

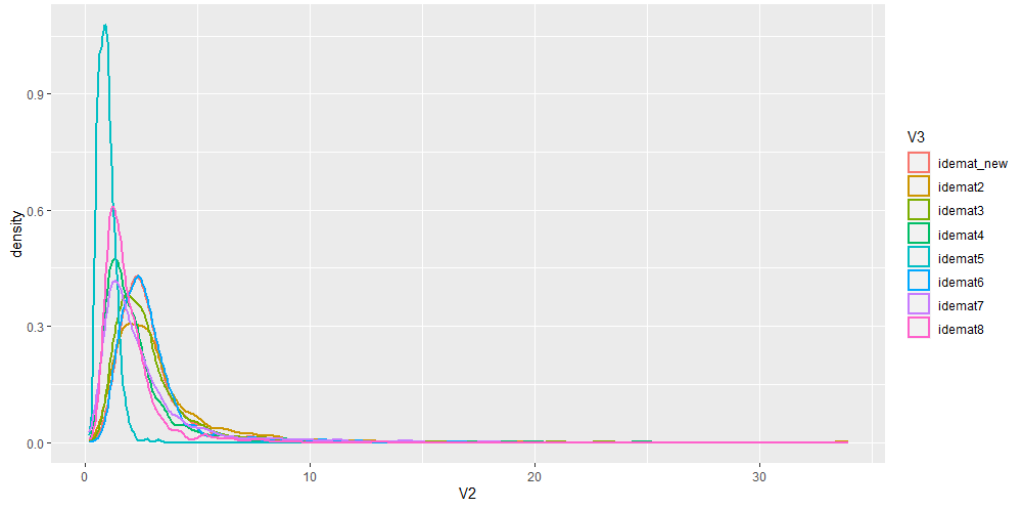


Figure B.4: Empirical distributions of the trend variance from 8 groups

*Note: the information of the SSMs used to simulate is contained in Table B.2*

Name	Length(yrs)	$\sigma_I^2$	$\sigma_T^2$	$\sigma_S^2$
simlist_new	15	20	10	1
simlist2	15	100	25	1
simlist3	20	100	25	1
simlist4	15	25	100	1
simlist5	15	1	0.25	1
simlist6	15	200	100	10
simlist7	15	$(N(0, 10))^2$	$(N(0, 10))^2$	1
simlist8	30	$(N(0, 10))^2$	$(N(0, 10))^2$	1

Table B.2: Information of SSMs used for simulation