Short Master Thesis

Linyi Guo

March 2020

1 Introduction

Review the general expression of a state space model is

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

And the state space model applied in our research is

$$\begin{aligned} y_t &= T_t + S_t + I_t \\ T_{t+1} &= T_t + \eta_t \\ S_{t+1} &= -\sum_{j=1}^{s-1} S_{t+1-j} + \omega_t \end{aligned} \tag{2}$$

where I_t, η_t and ω_t are IID Gaussian noises with mean 0 and variances σ_I^2, σ_T^2, σ_S^2.

In this chapter we shall talk about our main contributions to the seasonal adjustment problem, especially to the decomposition part. Section 2 will introduce the motivation behind our research by comparing the results from MLEs and X-11. After defining some penalty functions to reproduce the conventional methods' decomposition in Section 3, we utilize some weakly-informative priors to optimize the decomposition from MLEs in Section 4. And then we shall make use of the prior knowledge to build an empirical prior in Section 5 and compare its effect with other models. Meanwhile, we will try to approximate it with a parametric distribution and explain the motivation for Section 6, which talks 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.

2 Behaviour of maximum likelihood estimation

We have seen that there are some parameters in the state space model so to apply the Kalman filter we need to estimate them at first. One natural way is to use the maximum likelihood estimation but it turns out the decomposition results are not convenient for us to analysis even compared with those from random small numbers.

Let's take the unemployment data of the United States from 1990 to 2016 as an example. The following figure 1 is the comparison of results from X-11 and SSM with MLE:

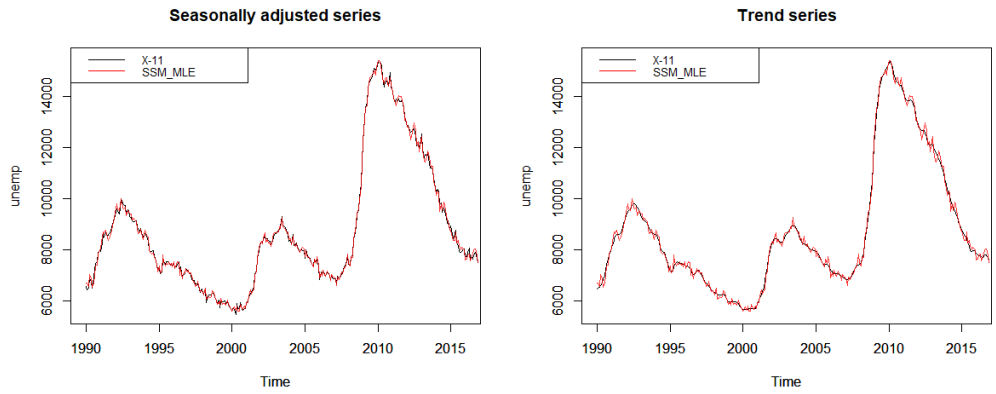

Figure 1: Decomposition comparison between X-11 and SSM(MLE)

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, 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 Appendix B.

Therefore, given the frequency analysis is not consistent with our demand, an alternative method is to consider Bayesian analysis, which we are going to talk about in Section 4 and Section 5.

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 shall 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 penalize the seasonally adjusted series. Then we define our first loss function as:

$$Loss = \|Trend_{X11} - Trend_{SSM}\|_2^2 + \|Seasonal_{X11} - Seasonal_{SSM}\|_2^2 \qquad (3)$$

Because it is hard to write the expression of the loss function explicitly, 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 I used at the beginning, but this method is too untechnical and time-consuming. To accelerate our operation, we adopt the derivative-free optimization algorithms such as *Hooke-Jeeves* and *Nelder-Mead* algorithms to solve this black-box optimization problem.

After calculating, the values of parameters σ_I^2, σ_T^2 and σ_S^2 with the lowest loss are 3.93750, 0.90625 and 1.87500. To further improve our accuracy, especially to control the smoothness of the trend component, we introduce another term with regard to the derivative of the trend:

$$Loss2 = \|Trend_{X11} - Trend_{SSM}\|_2^2 + \|Seasonal_{X11} - Seasonal_{SSM}\|_2^2 + \|D(Trend_{X11}) - D(Trend_{SSM})\|_2^2 \quad (4)$$

where operation D takes the difference between two adjacent values. And the final best parameter values are 4.46875, 3.00000 and 2.31250. Their decomposition result is as following

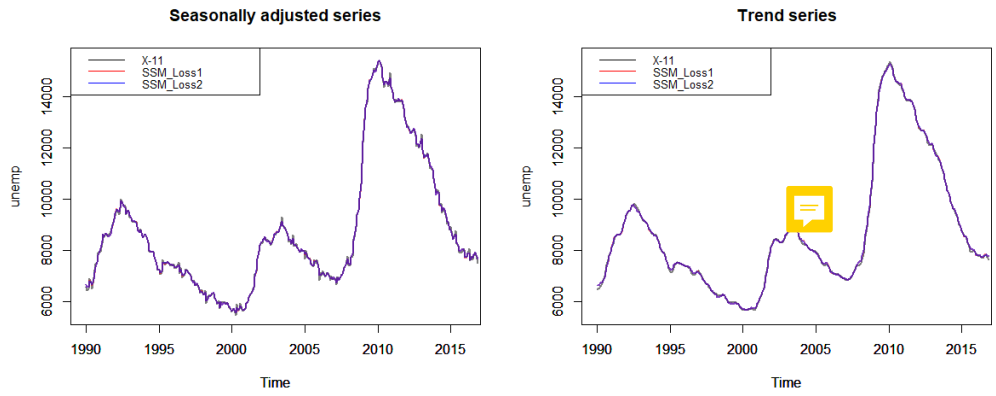

Figure 2: Decomposition comparison between X-11 and SSMs(Loss)

The figure above shows that both our current decomposition fit better compared with the result from MLE and the distinction between two loss functions is not obvious either. So far we have realized how to reproduce the classical decomposition result but we also notice that our loss function relies on X-11 or other conventional methods, which is not what we expect. We shall use empirical Bayesian analysis to fix this in Section 5.

To accelerate and simplify our computation, we also utilize the following property to reduce the number of parameters from 3 to 2:(refer to Appendix A for the proof details)

Lemma 3.1. For the same time series, if two state space models share the same ratios of three variances in Equation 2, then their decomposition stay the same.

Therefore we mainly need to care about their ratios. Without loss of generality we would let $\sigma_S^2 = 1$ in the following analysis if we don't specify in particular. 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 σ_I^2 and σ_T^2 w.r.t the *Loss2*. Then we obtained their distributions:

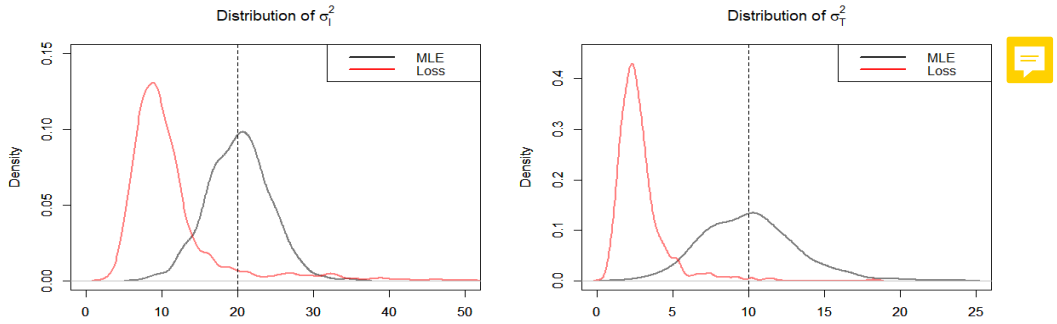

Figure 3: Distributions of variance estimators

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 5 and Appendix B) even though parameter values 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).

4 Weakly-informative prior distribution

After the discussion above, we will talk about the weakly-informative prior distribution in this section. The reason why we call them weakly-informative is due to the information they contain is weaker than whatever the real prior knowledge is. 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 useless 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 σ_I and σ_T. The reason why we didn't adopt the half-Cauchy distribution is 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.

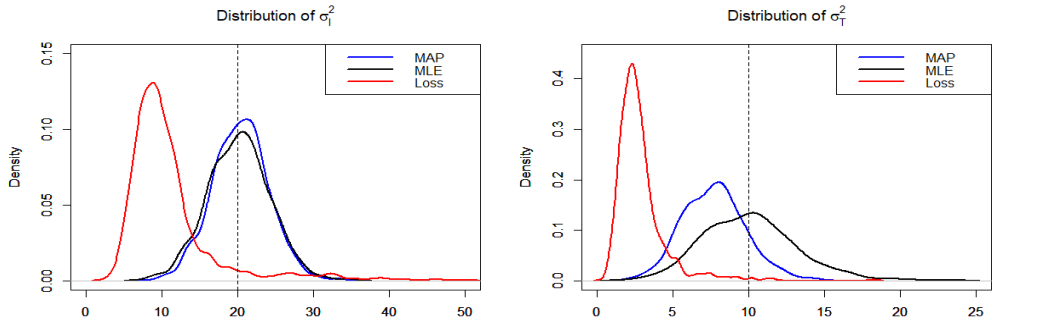

Figure 4: Comparison of variance distributions

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. By defining the decomposition error below, we are able to compare the methods above, as is shown in Table 1 and Figure 5.

$$Error = \|Trend_{X11} - Trend_{SSM}\|_2^2 + \|Seasonal_{X11} - Seasonal_{SSM}\|_2^2 \quad (5)$$

	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 1: Information of decomposition error

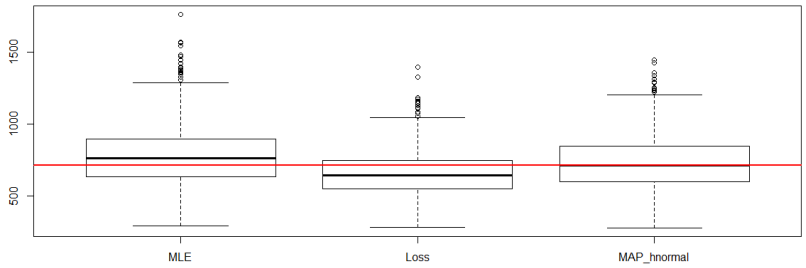

Figure 5: Boxplot of decomposition errors

We could tell that the state space model from maximum a posterior estimators does behave better than that from MLEs, and to back up our argument here we also used the *Friedman* test and *Mann–Whitney U* test to do the hypothesis test. Both results showed that the difference between MLE and MAP is prominent(See Appendix B). Combined with the information above, we are confident to say our weakly-informative prior, half-normal distribution does work.

5 Empirical prior distribution

In the last section, we have shown the weakly-informative prior does help us to have a better decomposition that we expect. The motivation of this section is part from the current background, big data 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 prior distribution in Bayesian analysis. Based on this, we will treat the parameter distribution generated from loss function as our prior distribution. 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 data sets.

Figure 6 shows the variance distribution comparison of different methods. Although the shape of maximum a posterior estimators from empirical prior is wi[illegible], it is really closer to the variance distribution that we expect. To make our guess more convincing, we calculated the mean and standard deviation of different decomposition error and did the same hypothesis test mentioned above. 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.

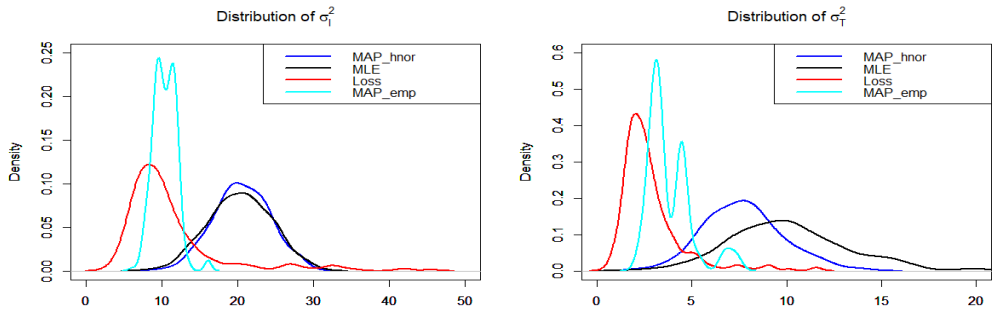

Figure 6: Comparison of variance distribution(2)

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 the following figure, 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 is very bad, mainly because its domain is limited instead of the whole non-negative real number. Besides this point, 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 add a very weak tail to each distribution, which we will show later. Another method is to use a parametric distribution to approximate the empirical distribution. As for the transformation problem, we would talk about it in the next section.

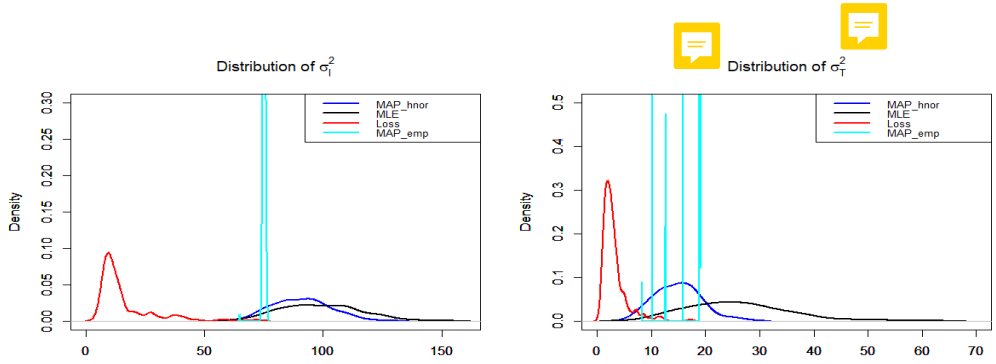

Figure 7: One failed case of the empirical prior

Based on Figure 6 and Figure 7, it is easy to find the distributions of maximum likelihood estimators has a close relationship with data itself, that is different data could own entirely different MLEs. However, after a bunch of simulation and analysis, we found the distributions of optimal σ_I^2 and σ_T^2 w.r.t the loss function defined above mainly center on intervals [0,50] and [0,15] approximately. The figure below is the distributions of optimal values from 8 different groups(each group contains 1000 data sets, see Appendix B for details):

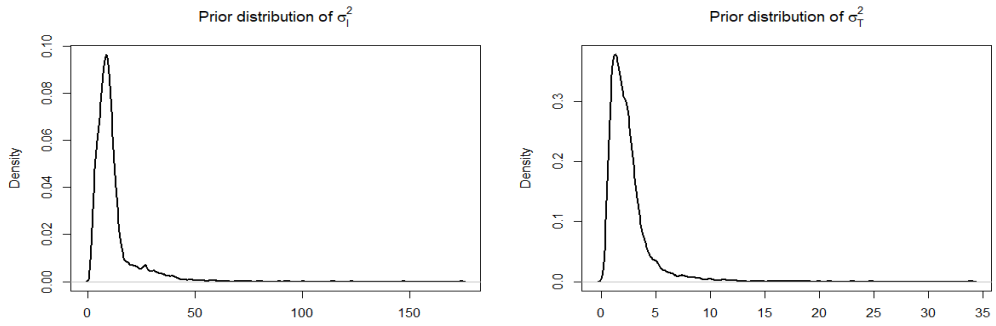

Figure 8: Empirical distribution of variances

To avoid previous weird posterior distributions, we add two tails from $N(0,\frac{50}{3})$ and N(0,5) to their right side separately and use them as the empirical priors to get the corresponding posterior estimators:

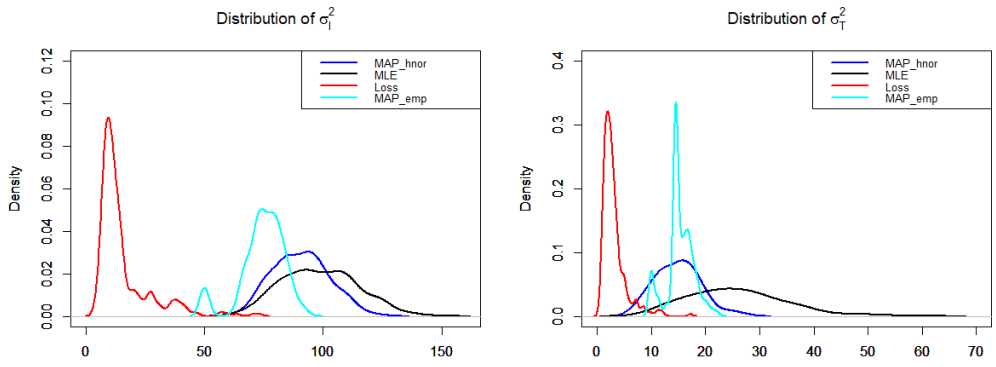

Figure 9: Improvement of the last failed case

The problem left now is that the MAP estimator is not close enough to the value that we want. Section 6 will talk about the method we used to make them closer.