COMP6212 (2016/17): Computational Finance Assignment 4 Report

Ningning Liu (ln1u16@soton.ac.uk)
MSc Data Science
Wednesday 24 May 2017

This report was done to summarize the work for COMP6212 assignment 4.

Task 1: Implementation for White Noise Time Series Model

In this task, a white noise time series is generated based on process vector $\boldsymbol{a} = [a_1 \quad a_2 \quad a_0]^T$ and observation noise variance $v_t \sim \mathrm{N}\,(0, 1)$. So, for each observation z_t at time step t, such that $z_t = Y_t \boldsymbol{a} + v_t$ where Y_t is matrix for historical values and a_0 is weight for constant term in least square solution. Firstly, I define $\boldsymbol{a} = [0.01 \quad -0.2 \quad 0.5]^T$ and starting observations is one. Fig 1 shows observations from designed white noise time series.

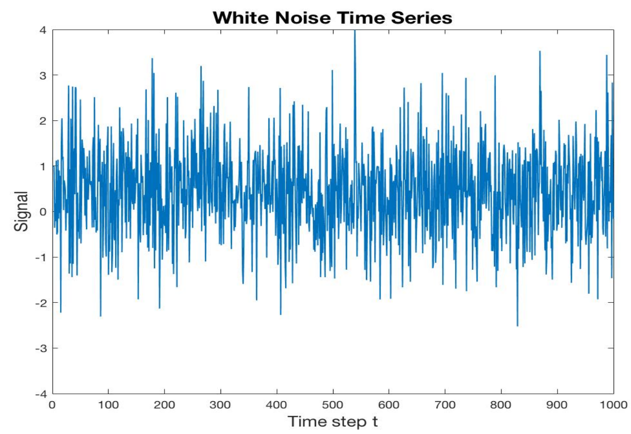


Figure 1: Generated observations in white noise time series with observation noise variance $v_t \sim N\ (0,1)$

Next, I use least square solution and Kalman filtering to estimate $\boldsymbol{a}$ respectively.

Least square solution

Initially, the design matrix Y has three columns. In the linear part, the first two columns store historical values from white noise time series using sliding window of two observations. The third column has ones multiplying a_0. Vector $\boldsymbol{f}$ is the true values. Hence, by least square, the estimated $\widehat{\boldsymbol{a}} = (Y^T Y)^{-1} Y \boldsymbol{f}$. It gives $\widehat{\boldsymbol{a}} = [0.041 \quad -0.086 \quad 0.42]^T$. We can see some considerable difference between $\widehat{\boldsymbol{a}}$ and $\boldsymbol{a}$.

Kalman filtering

In dynamic system in model **(1)**, $\boldsymbol{a}_t$ is updated by previous vector $\boldsymbol{a}_{t-1}$ and process variance w_t and observations f_t at time t can be predicated by set of past observations $\boldsymbol{y}_t$ and $\boldsymbol{a}_t$ with observation variance v_t. F_t is taken as an identity matrix and V_t is defined as one. In Kalman filtering [1], we should tune W_t to check if there is convergence in the system. For simplicity, we take W_t as fixed constant.

$$\begin{cases} f_t = \boldsymbol{y}_t \boldsymbol{a}_t + v_t & v_t \sim \mathrm{N}\,(0, V_t) \\ \boldsymbol{a}_t = F_t\, \boldsymbol{a}_{t-1} + w_t & w_t \sim \mathrm{N}\,(0, W_t) \end{cases} \qquad \textbf{(1)}$$

Kalman filter in model **(2)** follow below set of equations recursively. $P_{t|\tau}$ is associated error matrix where τ denoting the time step before and including τ. Fig 2 shows several experiments of estimating $\boldsymbol{a}_t$ with tuning different W_t and V_t. Particularly in Fig 2 (d), with smaller W_t and V_t, there is clear convergence of $\widehat{\boldsymbol{a}}_{t|t}$ towards true $\boldsymbol{a}$ whereas Fig 2 (a, b, c) shows no convergence when time step increasing.

$$\begin{cases} \widehat{\boldsymbol{a}}_{t|t-1} = F_t \widehat{\boldsymbol{a}}_{t-1|t-1} \\ P_{t|t-1} = F_t P_{t|t-1} F_t^T + W_{k-1} \\ r_t = f_t - \boldsymbol{y}_t \boldsymbol{a}_{t-1} \\ S_t = \boldsymbol{y}_t P_{t|t-1} \boldsymbol{y}_t^T + V_t \\ K_t = P_{t|t-1} \boldsymbol{y}_t^T S_t^{-1} \\ \widehat{\boldsymbol{a}}_{t|t} = \widehat{\boldsymbol{a}}_{t|t-1} + K_t r_t \\ P_{t|t} = (I - K_t \boldsymbol{y}_t) P_{t|t-1} \end{cases} \qquad \textbf{(2)}$$

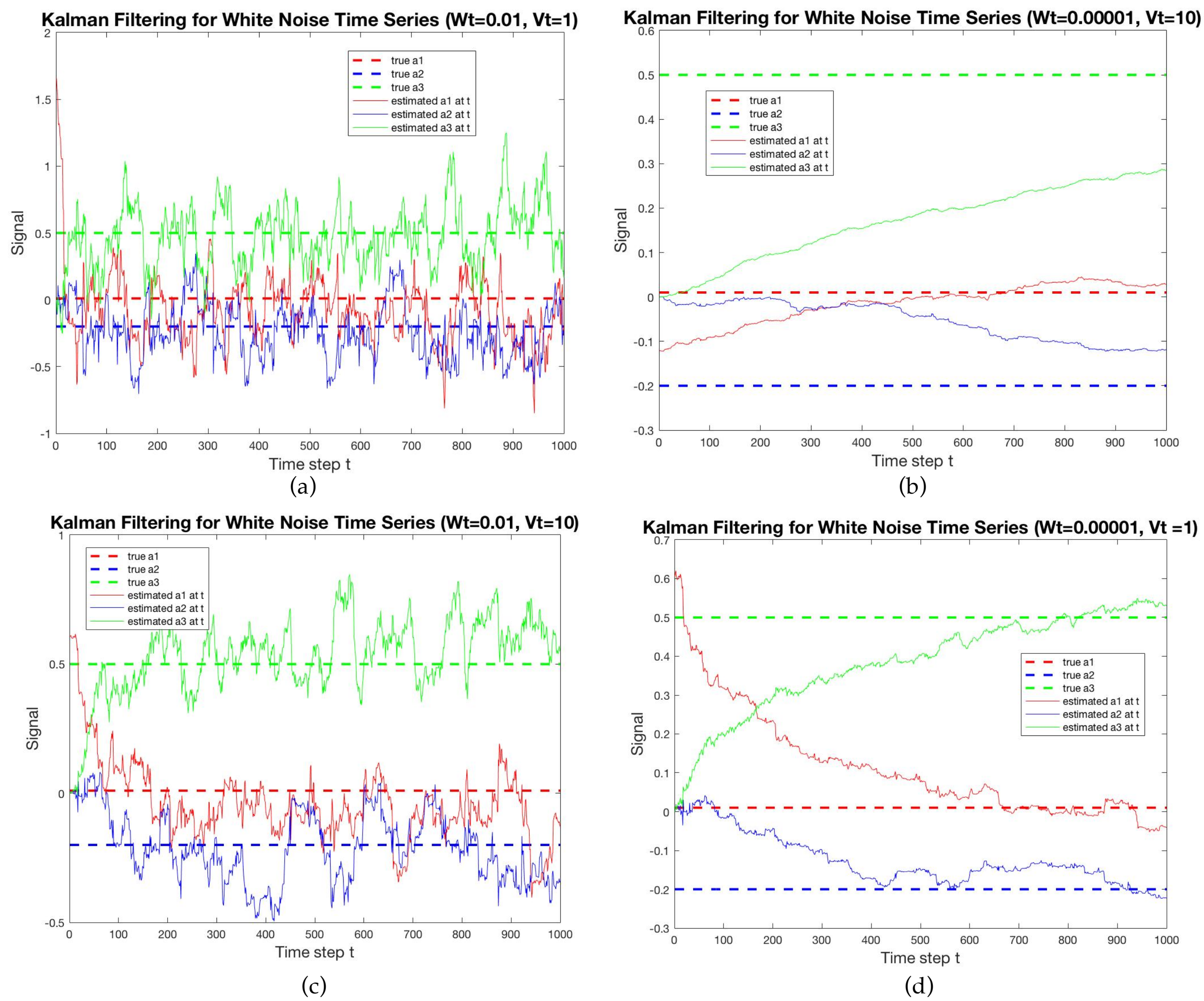
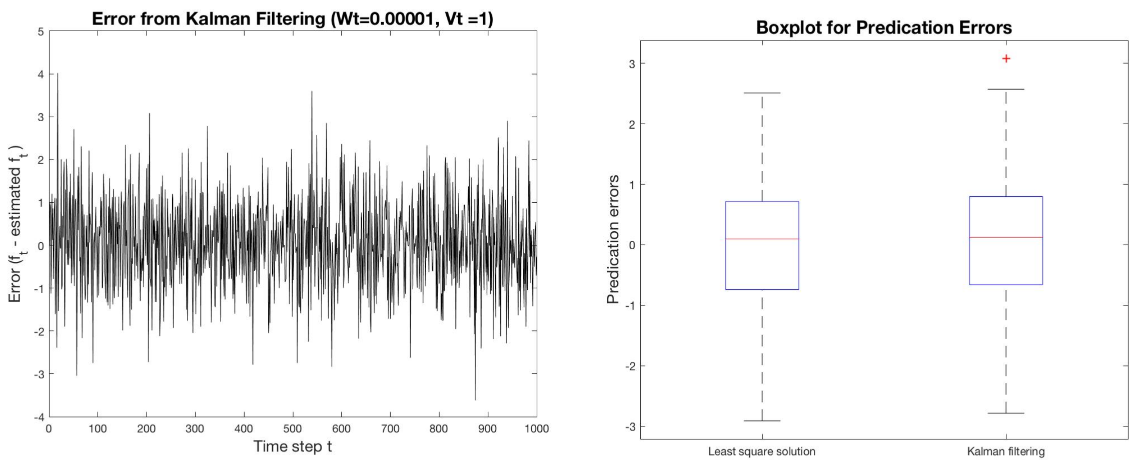


Figure 2: Experiments of estimating $\boldsymbol{a}_t$ using Kalman filter with tuning different W_t and V_t. If W_t takes 0.00001 and V_t takes its original value 1, (d) shows it can converge to its true vector $\boldsymbol{a}$ when time increases. If W_t and V_t are tuned to larger values, (a, b, c) shows that there is no convergence towards $\boldsymbol{a}$.

The predication error from Kalman filter in Fig 3 (a) approximately has Gaussian distribution with mean zeros and variance 1. Fig 3 (b) displays boxplot predication errors using least squares and Kalman filter respectively. We can see that both estimation methods give considerable small predication errors. However, $\widehat{\boldsymbol{a}}_{t|t}$ from Kalman filtering is $[0.071 \quad -0.191 \quad 0.457]^T$ which is very closed to true parameter $\boldsymbol{a}$.



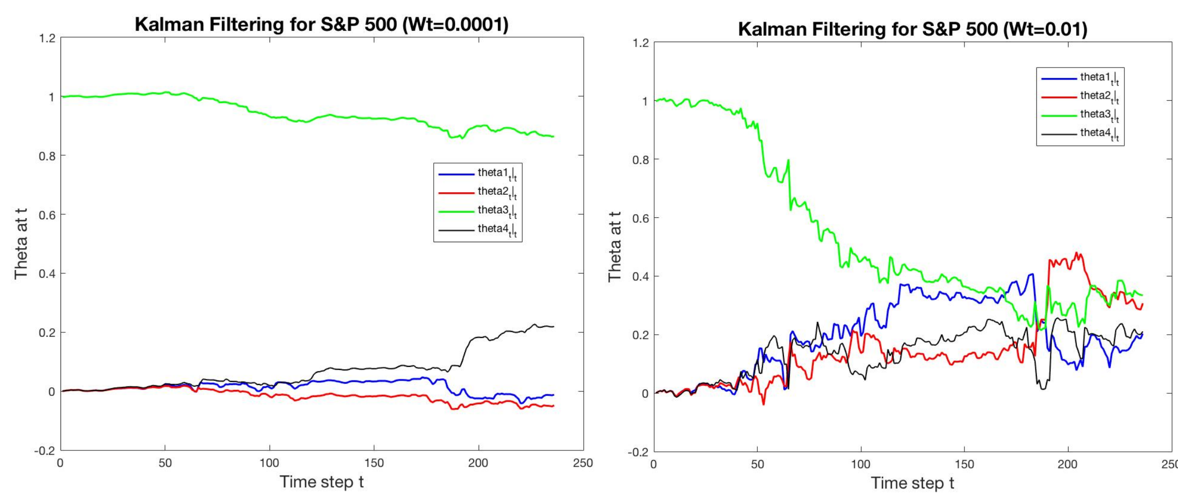
(a) (b)

Figure 3: (a) The prediction error $r_t = f_t - \boldsymbol{y}_t \boldsymbol{a}_{t-1}$ is illustrated in (d). We can see r_t approximately has Gaussian distribution with mean zero and variance one. (b) Boxplot for predication errors using least squares and Kalman filtering shows that Kalman filter has greater accuracy when predicting dynamic time series.

Task 2: LagLasso to Explain the Residual

Step 1: Kalman filtering

Initially, we implement Kalman filter to real data of S&P 500 index. Observation noise variance V_t is obtained from second order autoregressive model. It gives $V_t = 2800$. By tuning process noise variance W_t in Kalman filtering, we need to check if there is any convergence in this system. In this step, the sample period is 20 years' monthly data and sliding window of 4 months is taken as historical values [1]. Suppose the process vector is $\boldsymbol{\theta}_t$ with variance W_t. We estimate $\boldsymbol{\theta}_t$ using Kalman filter.



(a) (b)

Figure 4: By tuning process noise variance W_t (a) and (b) gives two patterns of $\widehat{\boldsymbol{\theta}}_t$ with $W_t = 0.0001$ and $W_t = 0.1$ respectively. (a) shows convergent of $\widehat{\boldsymbol{\theta}}_t$ when time increases whereas there is no convergence in (b).

Fig 4 displays patterns of $\widehat{\boldsymbol{\theta}}_t$ (i.e. estimated $\boldsymbol{\theta}_t$) with two different W_t. Fig 4 (a) shows that with smaller W_t, it has convergence for $\widehat{\boldsymbol{\theta}}_t$ when time increases. However, we need further model error signal using LagLasso to explain the residual.

Step 2: Selecting variables using LagLasso

Taking the residuals from the model in Fig 4 (a) as target, we select the following explanatory variables that may influence the residuals [1]: the S&P 500 Price-to-Earing (**PER**), the Spot Oil Price (**OIL**), the NANP ISM Manufacturing Purchasing Managers' Composite Index (**PMI**), the Disposal Personal Income (**INCOME**), the Corporate Profit After Tax (**COR PROFIT**), the US Population (**POPULATION**), the US Unemployment rate **(UNEMPOLYED**).

Algorithm 1 LagLasso Variable Selection

Input: **R**, $\boldsymbol{S}_w$, T, lag, τ

Output: W

Initialize: k= lag and $W_{k-1} = \emptyset$

While $k + 1 \leq T$ **do**

For $\mathrm{r}_{\mathrm{k}+1} \in \boldsymbol{R}$, compute $\mathbf{w}_{\mathrm{k}} = min_{\mathbf{w}_{\mathrm{k}}} \left\|\mathrm{r}_{\mathrm{k}+1}\text{-}\mathbf{S}_{\mathrm{w}_{\mathrm{k}}}\mathbf{w}_{\mathrm{k}}\right\|_2^2 + \tau\|\mathbf{w}_{\mathrm{k}}\|_1$

For all w_j in $\mathbf{w}_{\mathrm{k}}$, $j = 1, 2, \ldots, 7\times|lag|$, set $w_j = 0$ if $w_j < 0.0001$

Set $W_k = W_{k-1} \cup \mathbf{w}_{\mathrm{k}}$

Set $k = k + 1$

End while

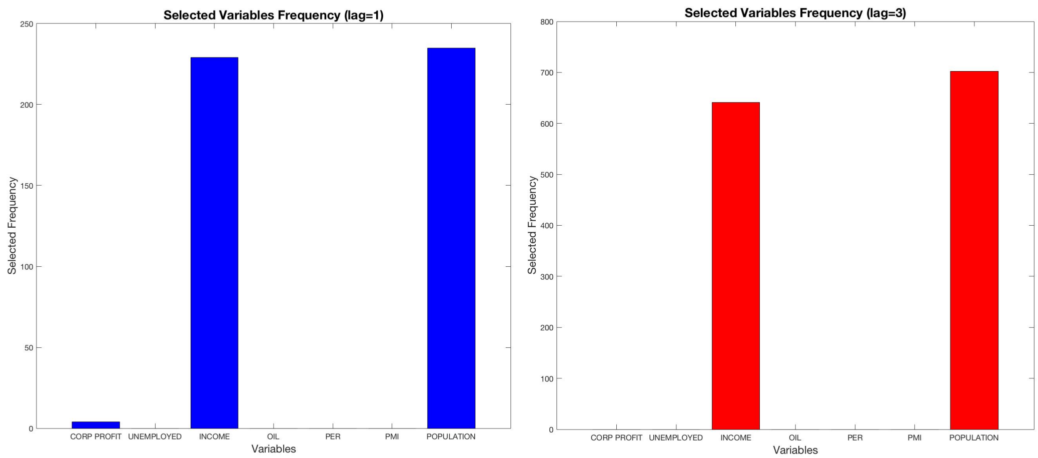
Set $W = W_{T-1}$

Return W

In **Algorithm 1**, we use LagLasso to select relevant variables combining with sliding window [1,2]. Let us denote above variables as $\boldsymbol{rs}_j, j = 1, 2 \ldots, 7$ accordingly and $\boldsymbol{S} = \{\boldsymbol{rs}_1, \boldsymbol{rs}_2, \ldots, \boldsymbol{rs}_7\}$. $\boldsymbol{S}_w$ is $1\times(|lag|\times 7)$-dimensional sub-vector of **S** where weights **w** is $(|lag|\times 7)\times 1$-dimensional vector and $|lag|$ is window size i.e. the past values. Suppose we takes 20 years' monthly data ($\mathbb{T} = 20\times 12 = 240$) for explanatory variables and synchronize the time frame with the residuals $\boldsymbol{R} = \{r_{lag+1}, r_{lag+2}, \ldots, r_T\}$. So, we use the lags (past values) of $\boldsymbol{S}$ to predict the $(lag+1)^{th}$ residual r_{lag+1} in **R**. In the end, we collect all estimated **w** at each computation into W where W is $|T - lag|\times 7$- dimensional matrix.

Explanatory variables	**number of selections**	
	Lag = 1	**lag = 3**
COR PROFIT	4	0
UNEMPOLYED	0	0
INCOME	229	641
OIL	0	0
PMI	0	0
POPULATION	235	702

Table 1: Number of selection for seven explanatory variables using Kalman LagLasso, Lag = 1 and Lag = 3



(a) (b)

Figure 5: Selected variables histogram for Kalman LagLasso. (a) Lag = 1. (b) Lag = 3. INCOME and POPULATION are selected in both figures.

In this step, we input lag = 1 and 3 into **Algorithm 1** to see how different lags affect variable selection. By tuning an appropriate value of τ, the frequency of non-zero coefficients for explanatory variable has been shown in Table 1 and visualized in Fig 5. INCOME and POPULATION are selected for both Fig 5 (a) and Fig 5 (b). Hence, these two variables have considerable relevance with error residual of S&P 500. This work shows that some external information in the environment can explain the error signal resulted from time-series predication model.

Reference:

[1] N. Mahler, "Modeling the S&P 500 index using Kalman filter and LagLasso," in *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop*, Sept 2009, pp. 1-6.

[2] R. Tibshirani, "Regression shrinkage and selective via the lasso," *Journal of Royal Statistical Society*, 1996, Series B, vol. 58, no.1, pp. 267-288.