Handout 5

Estimation of trend and seasonality.[Chapter 1.5 in Brockwell and Davis]

Consider the data "Electricity Sales to the Residential Sector", Jan 1990 - Dec 2012. The data has a total of $n = 276$ months of observations. We will denote the electricty sale at time t by Y_t, $t = 1, ..., n = 276$, with the understanding that time 1 corresponds to Jan 1990 and time 276 corresponds to Dec 2012. The plot of the series (Handout 1) shows that Y_t has three components: a trend part, a monthly component (seasonality) and a rough part. Seasonality means this: January sales tend to be similar in consecutive years, February sales tend to be similar in consecutive years an so on. The model is thus

$$Y_t = m_t + s_t + X_t, t = 1, ..., n = 276, \tag{1}$$

where m_t is the trend (smooth part), s_t is the seasonal component, and X_t is the rough part. [Note: In the text, electricity sale is denoted by X_t and the rough part is denoted by Y_t. Here the notations are exactly the opposite.]

In order to forecast Y_{281}, the electricity sale for May, 2013 (i.e., $t = 281$), we will need to get estimates for m_{281}, s_{281} and X_{281}. If we get these estimates then our forecast of Y_{281} is

$$\hat{Y}_{281} = \hat{m}_{281} + \hat{s}_{281} + \hat{X}_{281},$$

where $\hat{m}_{281}$, $\hat{s}_{281}$ and $\hat{X}_{281}$ are the estimates of m_{281}, s_{281} and X_{281}, respectively. In general, the forecast of Y_{n+h} (with $h = 5$, say) is given by

$$\hat{Y}_{n+h} = \hat{m}_{n+h} + \hat{s}_{n+h} + \hat{X}_{n+h}.$$

In order to be able to estimate m_{n+h}, s_{n+h} and X_{n+h}, we will need to have estimates of the trend, seasonality and the rough part for $t = 1, ..., n$, i.e., we need to obtain reliable estimates of m_t, s_t and X_t for $t = 1, ..., n = 276$. We will now discuss how to obtain estimates of m_t, s_t and X_t.

If look at the data plot, we can get an idea of how the trend looks like. Note that the trend increases over time and the fluctuations around the trend also seem to increase with time. There are many plausible reasons for the increase in the trend as well as the increase in the overall fluctuations:

i) increase in the number of households,

ii) changes in types of energy sources used for different purposes.

One way to handle such data sets is to transform the data so that the fluctuations about the trend do not change with time for the trnsformed series. For instance, if you look at the original plot and the natural logarithm of the plot, you can see that the fluctuations about the trend seem to change very slowly over time in the latter plot. Is there something even better? Well it turns out when you plot

$1/\sqrt{Y_t}$, the fluctuations seem to be almost the same over time. We will discuss how to select a reasonable transformation. But at this point of time, let us first see how to estimate the trend and seasonality for the transformed data.

Look at the plot of the series $Y_t' = 1/\sqrt{Y_t}$. From this plot it seems that the seasonal components are now the same across the years, i.e., January components seems to be the same over the years, February components seem to be the same over the years and so on. Instead of modeling the original series Y_t we will now model the transformed series Y_t' as

$$Y_t' = m_t + s_t + X_t, \tag{2}$$

where the trend, seasonality and the rough part are denoted by m_t, s_t and X_t, respectively. It is important to note that the three components of Y_t' as given in (2) are not the same as those given in (1), even though we have used the same notations.

Here is a strategy for modeling. We will use a regression method for estimating the trend and the seasonality. We can model the trend by a polynomial. Here we will use a cubic polynomial for the trend, i.e.,

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3. \tag{3}$$

Note that the numbers for $\{t^2\}$ and $\{t^3\}$ will be rather very large. So, for numerical stability we can use $t/n = t/276$ instead of t on the right hand side of (3).

There are many ways to create a model for the seasonal effects and here is one such. Create variable $I_{t,1}, ..., I_{t,11}$ as

$$I_{t,1} = \begin{cases} 1 & \text{if time t is Jan} \\ -1 & \text{if time t is Dec} \\ 0 & \text{otherwise} \end{cases}, I_{t,2} = \begin{cases} 1 & \text{if time t Feb} \\ -1 & \text{if time t is Dec} \\ 0 & \text{otherwise} \end{cases}, I_{t,11} = \begin{cases} 1 & \text{if time t Nov} \\ -1 & \text{if time t is Dec} \\ 0 & \text{otherwise} \end{cases}$$

Model s_t as a linear combination of $I_{t1}, ..., I_{t,11}$, i.e.,

$$s_t = \beta_4 I_{t,1} + \beta_5 I_{t,2} + \cdots + \beta_{14} I_{t,11}.$$

So the model we will fit is

$$Y_t' = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 I_{t,1} + \beta_5 I_{t,2} + \cdots + \beta_{14} I_{t,11} + X_t. \tag{4}$$

Note that this is linear regression model with 14 independent variables $\{t, t^2, t^3, I_{t,1}, ..., I_{t,11}\}$ and there are 15 beta parameters in this model. Use the standard least squares method to estimate the beta parameters. Once we estimate the parameters, then the estimated trend, seasonal components, fitted Y_t' values and X_t

are

$$\hat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 + \hat{\beta}_3 t^3, \quad \hat{s}_t = \hat{\beta}_4 I_{t,1} + \hat{\beta}_5 I_{t,2} + \cdots + \hat{\beta}_{14} I_{t,11},$$
$$\hat{Y}'_t = \hat{m}_t + \hat{s}_t, \ \hat{X}_t = Y'_t - \hat{Y}'_t = Y'_t - (\hat{m}_t + \hat{s}_t).$$

For the Electricity sales data, we have used $t' = t/n$ instead of t for numerical stability, i.e., we have used the representation $m_t = \beta_0 + \beta'_1 t' + \beta'_2 t'^2 + \beta'_3 t'^3$, where $\beta'_1 = n\beta_1$, $\beta'_2 = n^2\beta_2$ and $\beta'_3 = n^3\beta_3$. Here we have

$$\hat{\beta}_0 = 0.0623, \hat{\beta}'_1 = -0.0108, \hat{\beta}'_2 = -0.0203, \hat{\beta}'_3 = 0.0198,$$
$$\hat{\beta}_4 = -0.0046, \hat{\beta}_5 = -0.0005, \hat{\beta}_6 = 0.0016, \hat{\beta}_7 = 0.0057,$$
$$\hat{\beta}_8 = 0.0052, \hat{\beta}_9 = -0.0006, \hat{\beta}_{10} = -0.0059, \hat{\beta}_{11} = -0.0059,$$
$$\hat{\beta}_{12} = -0.0020, \hat{\beta}_{13} = 0.0037, \hat{\beta}_{14} = 0.0046$$

Estimates of X_t.

Once you have estimates of m_t and s_t, estimates of X_t are

$$\hat{X}_t = Y'_t - \hat{m}_t - \hat{s}_t.$$

How to estimate seasonality in the absence of trend

If the trend is absent, simply fit $m_t = \beta_0$ without the terms t, t^2 etc. and the rest of the method is the same.

How do we select the transformation?

We will now discuss the issue of selecting transformation for the data. This is achieved via Box-Cox (or power) transformations. Examples of power transformation are $\sqrt{Y_t}$, $\sqrt[3]{Y_t}$, $1/Y$ etc. As a general rule we can write the power transformation as Y_t^λ where λ can be positive or negative. Note that when $\lambda = 0$, $Y_t^\lambda = 1$ always. So in order to define this properly, a slight modification is done

$$Y_t(\lambda) = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(Y_t) & \lambda = 0. \end{cases}.$$

Note that when $\lambda \neq 0$, $Y_t(\lambda)$ is a rescaled version of Y_t^λ. So one can use $Y'_t = Y_t(\lambda)$. However, for practical purposes, it is enough to use $Y'_t = Y_t^\lambda$ if $\lambda \neq 0$ and, $Y'_t = \log(Y_t)$ if $\lambda = 0$.

There are a number of ways to select λ, but we will use a reasonably simple method. For each λ, we run a regression model given in (4) and then calculate the R-square value. Since this value depends on λ, call it $R^2(\lambda)$. Plot $R^2(\lambda)$ against λ and find the value of λ where the $R^2(\lambda)$ is the largest. Ideally speaking, one should calculate $R^2(\lambda)$ for many values of λ, but it turns out that in most practical applications, it

is good enough to consider the five transformations: $Y_t, Y_t^{1/2}, \log(Y_t), Y_t^{-1/2}$ and Y_t^{-1}. [A more cautious person may consider a total of 9 transformations: $Y_t, Y_t^{3/4}, Y_t^{1/2}, Y_t^{1/4}, \log(Y_t), Y_t^{-1/4}, Y_t^{-1/2}, Y_t^{-3/4}$ and Y_t^{-1}].

For the Electricity sales data, the maximum value of R^2 is at $\lambda = -.65$ and we have taken the transformation $Y_t^{-1/2}$ (since $\lambda = -0.5$ is close to the optimal point of -0.65).

Appendix:

You will find below R commands that

a) plot the data and its various transformations,

b) use the supplied function 'trndseas' to obtain a trend (a cubic polynomial) and the seasonals.

Plot of the data and its various transformations.

We can plot the series, its square root, logarithm amd inverse square root in four subgraphs in one plot. Here is how we can do it.

```
par(mfrow=c(2,2))
plot.ts(y,ylab='y',main='Plot of y')
plot.ts(y^.5,ylab='sqrt(y)',main='Plot of sqrt(y)')
plot.ts(log(y),ylab='ln(y)',main='Plot of log(y)')
plot.ts(1/y^.5,ylab='1/sqrt(y)',main='Plot of 1/sqrt(y)')
```

If you want the same plots as above, but you want subscripted y's to be displayed in the graphs, then the commands have to be modified as follows.

```
par(mfrow=c(2,2))
plot.ts(y,ylab=expression(paste(y[t])),main=expression(paste("Plot of ",y[t])))
plot.ts(y^.5,ylab=expression(paste("sqrt(",y[t],")")),main=expression(paste("Plot of sqrt(",y[t],")")))
plot.ts(log(y),ylab=expression(paste("log(",y[t],")")),main=expression(paste("Plot of ln(",y[t],")")))
plot.ts(1/y^.5,ylab=expression(paste("1/sqrt(",y[t],")")),main=expression(paste("Plot of 1/sqrt(",y[t],")"))).
```

Using the supplied function 'trndseas' to obtain the trend (a cubic polynomial) and the seasonals.

For the data, we first run the function 'trndseas' with a vector of lambdas containing values -1, -.95, -.90,...,1. Then we determine the best value of lambda which turns out to be equal to about -0.5. Then we run the function 'trndseas' again with a value of lambda equal to -0.5. The following are the commands needed to run and get all the plots.

```
lam=seq(-1,1,by=0.05)
ff=trndseas(y,seas,lam,3)
rsq=ff$rsq
```

```
ff=trndseas(y,seas,-0.5,3)
trend=ff$trend
season=ff$season
fit=ff$fit
tm=1:276
Month=1:12
```

Here are the commands to obtain the plots of

i) R-square against lambda, ii) 1/sqrt(y), fit and trend, iii) seasonals, iv) y, fit and trend (transformed version of (ii) in the original scale).

```
par(mfrow=c(2,2))
plot(lam,rsq,type='l',xlab='Lambda',ylab='R-sq',main='Electric sales: R-square')
plot.ts(1/y^.5,ylab='',main='Plot: 1/sqrt(y), trend and fit')
points(tm,fit,type='l',lty=2)
points(tm,trend,type='l')
legend(175,0.070, c("data","trend","fit"), lty=c(1,1,2))
plot(Month,season,type='l',ylab='Seasonals',main='Seasonals for 1/sqrt(y)')
plot.ts(y,ylab='',main='Plot: y, trend and fit')
points(tm,1/fit^2,type='l',lty=2)
points(tm,1/trend^2,type='l')
legend(0,525, c("data","trend","fit"), lty=c(1,1,2)).
```

The commands given below are the same as above except now the subscripts of y are displayed in the graph.

```
par(mfrow=c(2,2))
plot(lam,rsq,type='l',xlab='Lambda',ylab='R-sq',main='Electric sales: R-square')
plot.ts(1/y^.5,ylab='',main=expression(paste("Plot: 1/sqrt(",y[t],")",", trend and fit")))
points(tm,fit,type='l',lty=2)
points(tm,trend,type='l')
legend(175,0.070, c("data","trend","fit"), lty=c(1,1,2))
plot(Month,season,type='l',ylab='Seasonals',main=expression(paste("Seasonals for 1/sqrt(",y[t],")")))
plot.ts(y,ylab='',main=expression(paste("Plot: ",y[t],", trend and fit")))
points(tm,1/fit^2,type='l',lty=2)
points(tm,1/trend^2,type='l')
legend(0,525, c("data","trend","fit"), lty=c(1,1,2)).
```

Figure 1: Electricity sales: Jan, 1990 - Dec, 2012

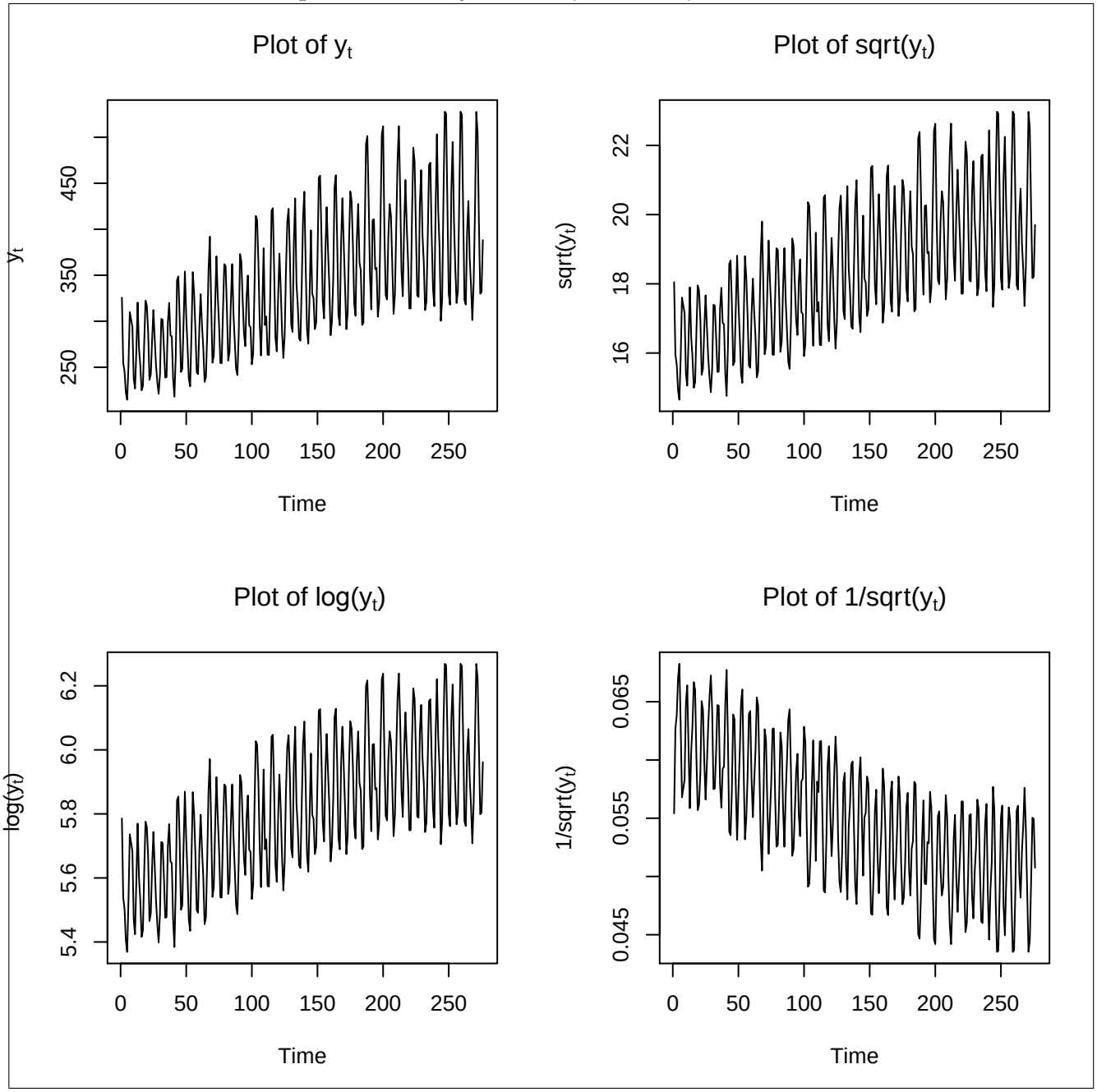

Figure 2: Electricity sales: Jan, 1990 - Dec, 2012