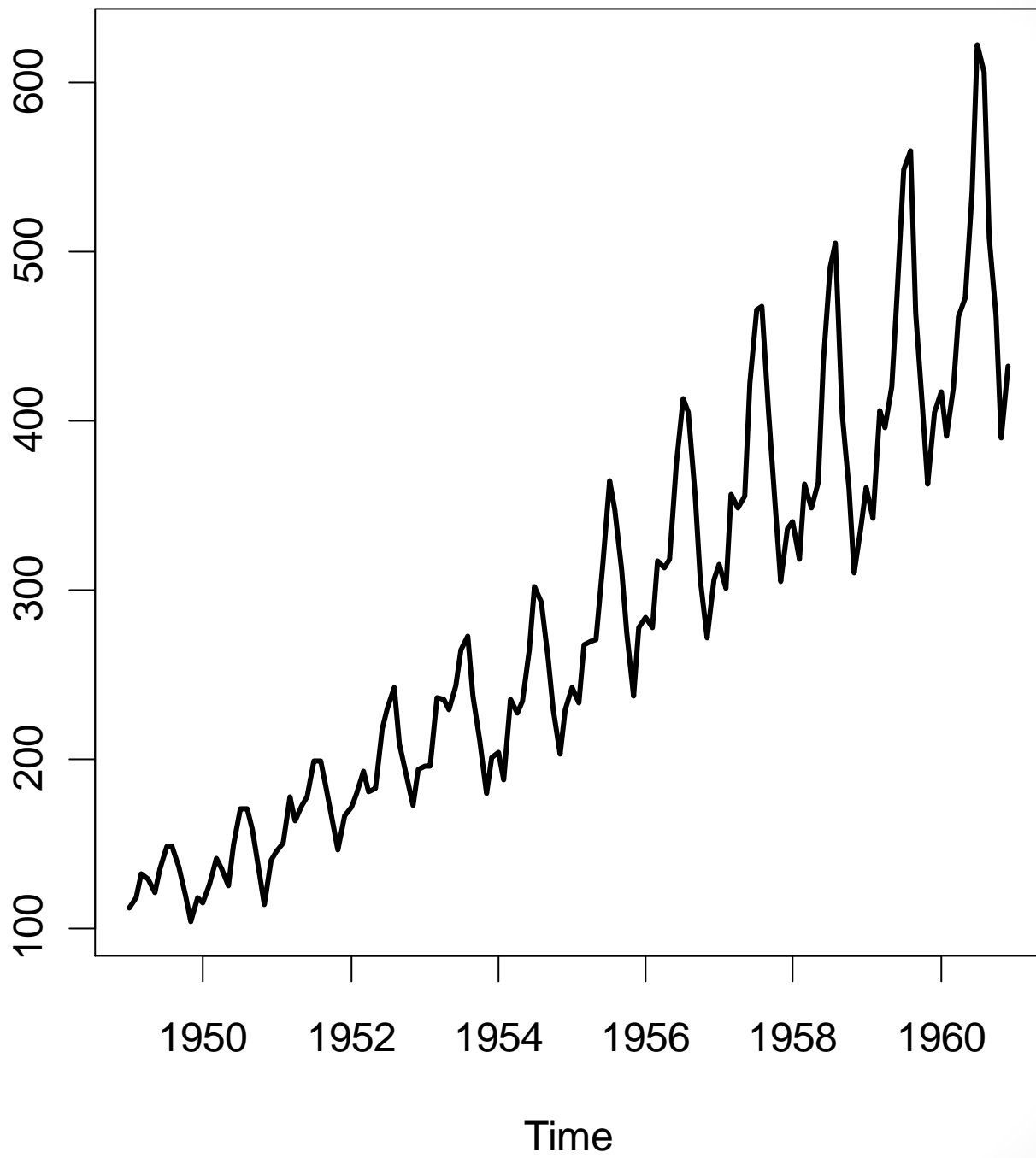


Выделение цикла, тренда и сезонности временного ряда

Машинное обучение — осень 2015

Количественная аналитика

AirPassengers



Содержание

- выделение сезонной и остаточной компонент
- разделение цикла и тренда

Описание модели

Пусть имеется процесс $\{y_t\}_{t=1}^N$

Нас интересует его разложение в виде

$y_t = c_t \cdot tr_t \cdot s_t \cdot i_t$, где

c_t — циклическая компонента;

tr_t — тренд;

s_t — сезонная компонента;

i_t — нерегулярная остаточная компонента

Для работы с аддитивной моделью, перейдём к логарифмам:

$$\ln y_t = \ln c_t + \ln tr_t + \ln s_t + \ln i_t$$

$$\bar{y}_t = \bar{c}_t + \bar{tr}_t + \bar{s}_t + \bar{i}_t$$

Выделение сезонной и остаточной компонент

Процедура Loess

Loess smoothing, LS

Сглаживание осуществляется с помощью полиномов с весовыми коэффициентами $v_t(x)$:

$$y_t \sim (x_t + x_t^2 + \dots + x_t^p) v_t(x),$$

$$v_t(x) = W\left(\frac{|x_t - x|}{d_q(x)}\right), \quad W(u) = \begin{cases} (1 - u^3)^3, & u \in [0; 1) \\ 0, & u \geq 1 \end{cases},$$

$d_q(x)$ — расстояние от x до q -го ближайшего x_t

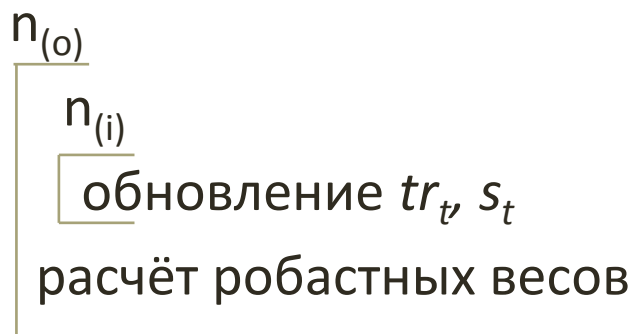
Выбирается $p = 1$, если зависимость $y(x)$ гладкая и $p = 2$ в противном случае

Если наблюдениям придаются разные веса $\rho_t = \frac{1}{\sigma_t}$, тогда в процедуре Loess используются веса $\rho_t v_t(x)$

Процедура STL

Seasonal-Trend Loess decomposition, STL

Процедура STL состоит из двух вложенных циклов:



$n_{(o)}$ — количество итераций внешнего цикла,

$n_{(i)}$ — количество итераций внутреннего цикла

Из исходного ряда выделяются $n_{(p)}$ сезонных подпоследовательностей (например, январские данные, февральские данные, ..., декабрьские данные)

Процедура STL

Пусть $s_t^{(k)}$ и $tr_t^{(k)}$ — оценки сезонной и трендовой компонент на k -м шаге внутреннего цикла

Схема расчёта $s_t^{(k+1)}$ и tr_t^{k+1} :

1. Удаление тренда: $i_t^* = y_t - tr_t^{(k)}$
2. Сглаживание сезонных подпоследовательностей:
 $s_t^{*(k+1)} = LS(i_t^*, q = n_{(s)}, p = 1), t \in \{0; \dots; T + 1\}$
3. Фильтрация $s_t^{*(k+1)}$:
 $MA(n_{(p)}) + MA(n_{(p)}) + MA(3) + LS(q = n_{(l)}, p = 1) \rightarrow$
 $\rightarrow l_t^{(k+1)}$
4. Расчёт сезонности: $s_t^{(k+1)} = s_t^{*(k+1)} - l_t^{k+1}$

Процедура STL

5. Удаление сезонности: $tr_t^{*(k+1)} = y_t - s_t^{k+1}$

6. Выделение тренда:

$$tr_t^{(k+1)} = LS \left(tr_t^{*(k+1)}, q = n_{(t)}, p = 1 \right)$$

Для самой первой итерации внутреннего цикла задаётся начальное значение $tr_t^{(0)} = 0$

Если исходный ряд не содержит сильных выбросов, тогда $n_{(o)} = 0$ и $n_{(i)} = 2$, иначе $n_{(o)} = 15$ и $n_{(i)} = 1$

Процедура STL

Параметры внутреннего цикла

- $n_{(l)}$ — наименьшее нечётное число, большее или равное $n_{(p)}$
- $n_{(s)}$ зависит от соотношения дисперсий сезонной и остаточной компонент; чем больше $n_{(s)}$ тем меньше дисперсия сезонности
- $n_{(t)}$ — наименьшее нечётное число, удовлетворяющее условию
$$n_{(t)} \geq \frac{1.5n_{(p)}}{1 - 1.5n_{(s)}^{-1}}$$

Процедура STL

Внешний цикл

$$i_t = y_t - tr_t - s_t$$

Расчёт робастных весов:

$$\rho_t = B\left(\frac{|i_t|}{6\text{median}(|i_t|)}\right), \quad B(u) = \begin{cases} (1 - u^2)^2, & u \in [0; 1) \\ 0, & u \geq 1 \end{cases}$$

Следующее выполнение внутреннего цикла происходит с весами $\rho_t v_t(x)$

Циклические подпоследовательности сглаживаются отдельно и объединяются в один ряд, получающийся неровным

Для его сглаживания применяют процедуру Loess:

$$LS(q = < small >, p = 2)$$

Процедура STL в R

для очистки ряда от сезонности используем методику STL

исходные данные

```
library(datasets)
```

```
y <- AirPassengers
```

процедура разложения временного ряда

```
library(stats)
```

```
stl.y <- stl(y,s.window="periodic",robust=FALSE)
```

Процедура STL в R

```
plot(stl.y)
```

Процедура **stl** разделяет
исходный ряд на три
составляющие:

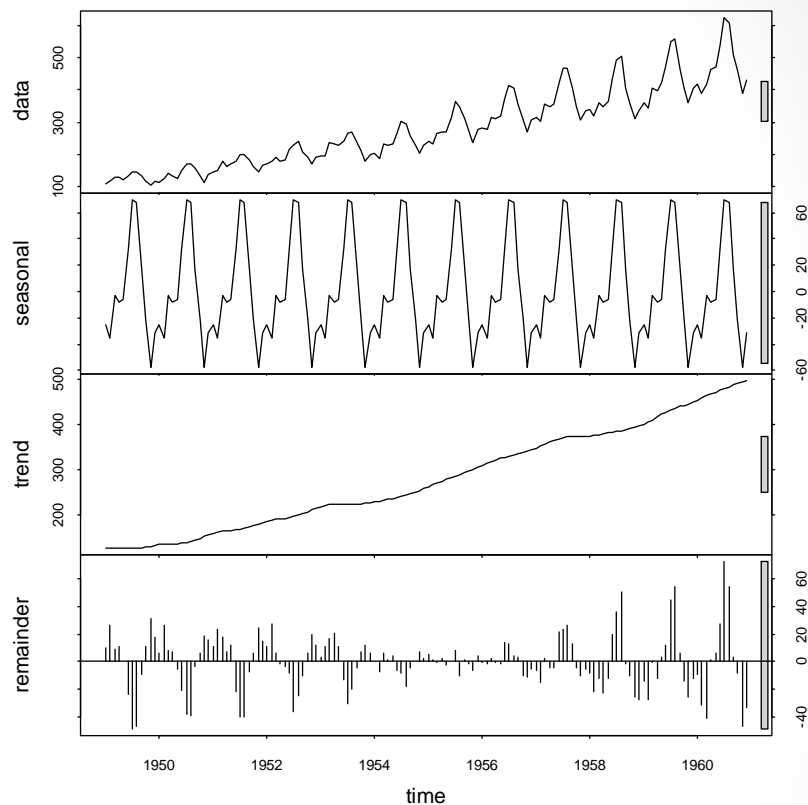
```
stl.y$time.series[, "seasonal"]
```

```
stl.y$time.series[, "trend"]
```

```
stl.y$time.series[, "remainder"]
```

При этом в наших

обозначениях

$$stl.y$time.series[, "trend"] = c_t + tr_t$$


Разделение цикла и тренда

Разделение тренда и цикла

Фильтр Ходрика-Прескотта

$$y_t = tr_t + c_t$$

Оптимизационная задача:

$$\sum_{t=1}^T (y_t - tr_t)^2 + \lambda \sum_{t=3}^T ((tr_t - tr_{t-1}) - (tr_{t-1} - tr_{t-2}))^2$$
$$\rightarrow \min_{\{tr_t\}_{t=-1}^T} .$$

λ — сглаживающий параметр, не допускающий сильной вариации тренда; чем больше λ , тем более гладок ряд $\{tr_t\}$

$\lambda_A = 6.25$ — для годовых данных

$\lambda_Q = 1600$ — для квартальных данных

$\lambda_M = 129600$ — для месячных данных

Разделение тренда и цикла в R

Для выделения тренда воспользуемся фильтром Ходрика-Прескотта

```
library(mFilter)
hp.trend <- hpfilter(stl.y$time.series[, "trend"],
type = "lambda", freq = [...], drift = FALSE)
```

Вместо [...] следует подставить либо NULL для автоматического подбора параметра λ , либо число, задающее его в явном виде

Разделение тренда и цикла в R

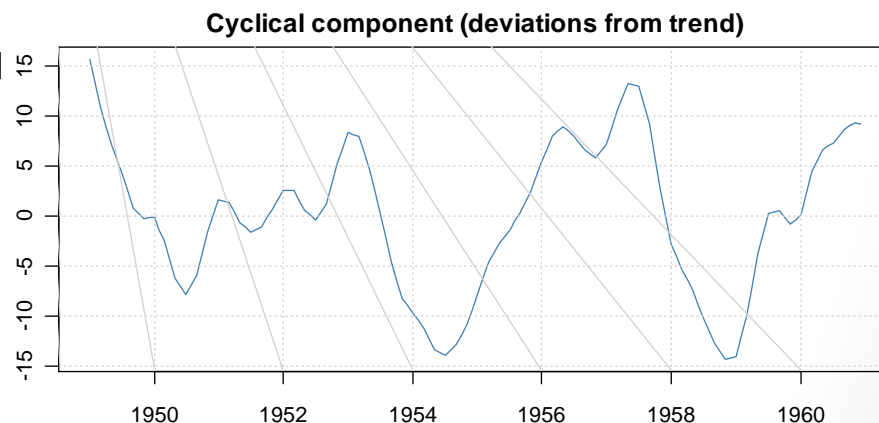
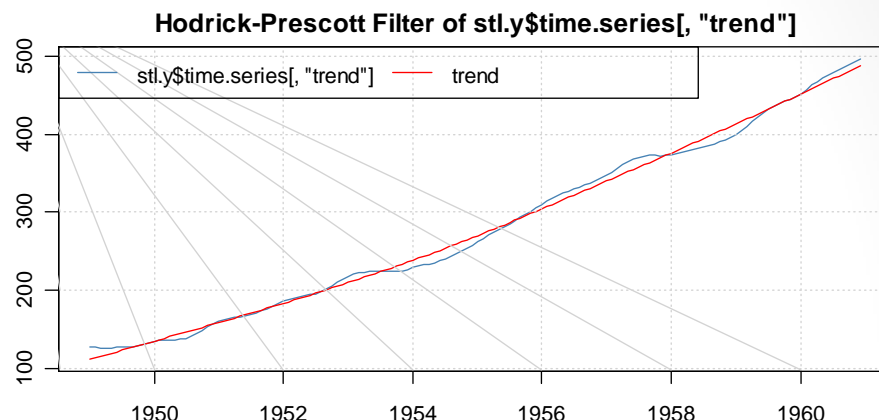
```
plot (hp.trend)
```

Процедура **hpfilter** разделяет исследуемый ряд на две составляющие:

`hp.trend$trend`

`hp.trend$cycle`

Для получения более гладкой циклической компоненты можно повторить процедуру для малых значений λ



```
cycle <- hpfilter(hp.trend$cycle, type="lambda", freq=1600)$trend
```

Домашнее задание

В файле «prod_indx_train.csv» находятся данные о динамике индекса промышленного производства (ИПП) одной из капиталистических стран за период с августа 1959-го года по сентябрь 1988-го (350 наблюдений)

Вашей задачей является построение прогноза уровня ИПП в периодах с октября 1988-го года по декабрь 1993-го (63 значения)

Ответ состоит из двух файлов:

- csv со столбцом прогнозных значений ИПП
- docx или pdf с описанием решения, комментариями и картинками

Исходные данные

```
tmp <- read.csv("prod_indx_train.csv", header = TRUE)[,1]  
y <- ts(log(tmp), start = c(1959, 8), freq = 12)
```

