时间序列分析与预测

第八讲

黄嘉平

深圳大学｜中国经济特区研究中心
粤海校区汇文楼办公楼 1510
课程网站 https://huangjp.com/TSAF/

1. 时间序列的平稳性和差分

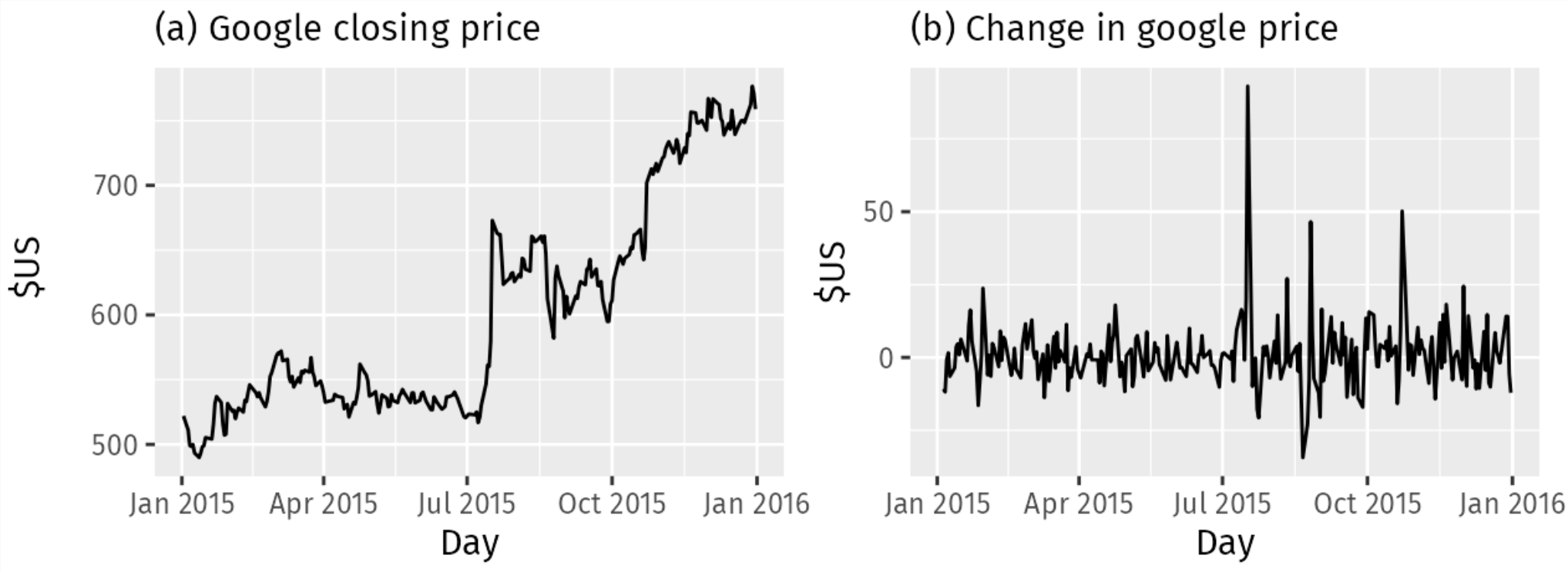
1.1. 平稳时间序列

当一个时间序列的分布特征与观测时间点无关时，我们称其为**平稳的（stationary）**时间序列。这里的分布特征一般表现为均值和方差。

- 不平稳序列的例子
  - 包含趋势和季节性的序列（均值随时间变化）
  - 局部变动幅度不固定的序列（方差随时间变化）
- 平稳序列的例子
  - 白噪声序列（注意平稳序列允许存在自相关，但白噪声序列不行）
  - 包含非季节性周期的序列（因为周期不固定，无法预知波峰和波谷）

通常，平稳时间序列不包含长期不变的可预测特征。

1.1. 平稳时间序列



左图是 2015 年 Google 股票的每日收盘价数据，右图是该价格的日度变化数据。

左图存在明显的趋势，因此是不平稳的。右图则大体上是平稳的。

1.2. 通过差分获得平稳序列

对于非平稳序列，我们可以通过**差分（differencing）**将其变换为平稳序列。

序列 y_t 的**一阶差分（first-order difference）序列**定义为

$$y'_t = y_t - y_{t-1}$$

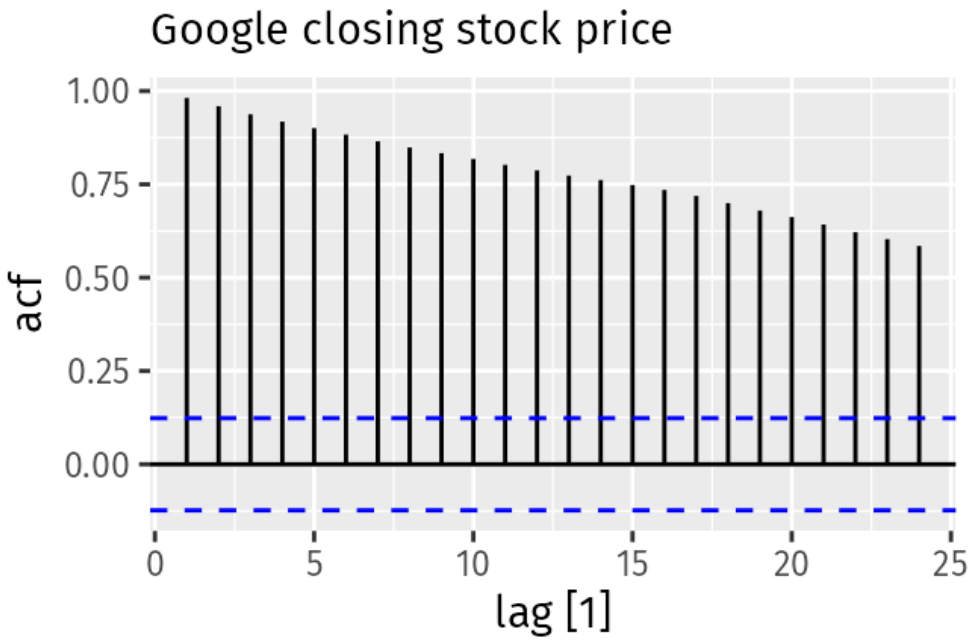
二阶（second-order）差分序列定义为

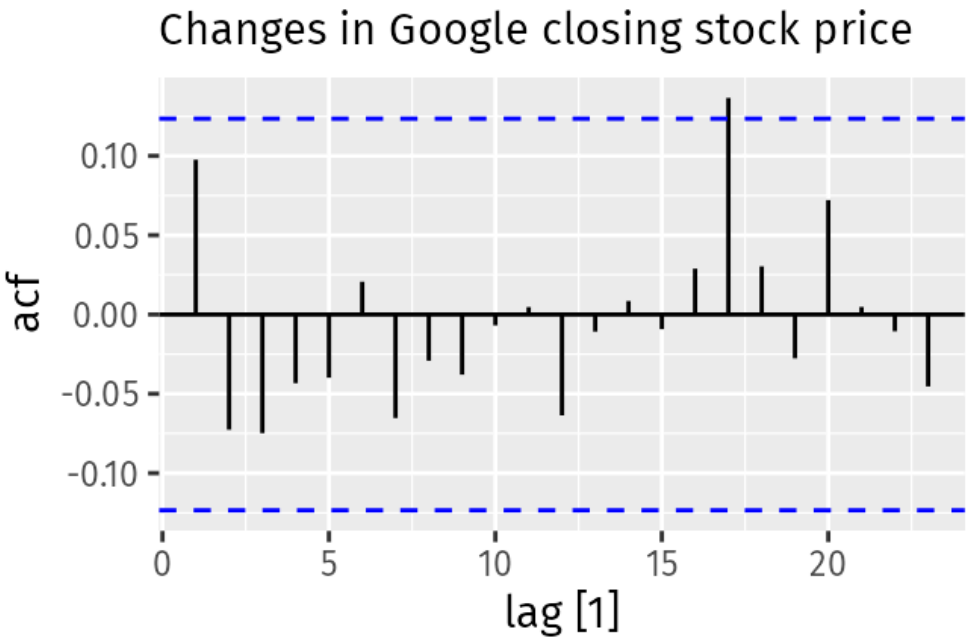
$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

以此类推，我们就可以定义 n **阶（nth-order）差分序列**。

1.2. 通过差分获得平稳序列

Google 的股价变化数据就是收盘价数据的一阶差分。通过观察二者的 ACF 图可以看出，差分消除了原序列中的趋势。





差分可以通过 tsibble 包中的 `difference()` 函数实现。

1.2. 通过差分获得平稳序列

如果一阶差分后的序列是白噪声，即

$$y_t - y_{t-1} = \varepsilon_t, \quad \varepsilon_t \text{ 是白噪声}$$

则有

$$y_t = y_{t-1} + \varepsilon_t$$

此模型称为**随机游走（random walk）**模型。随机游走模型在金融和经济领域拥有广泛的应用空间，例如股票价格的基本模型就是随机游走。

随机游走的特征体现为：(1) 明显的长期上升或下降趋势，(2) 突然的无法预测的趋势变化。

对于随机游走模型，由于其上升和下降的可能性相同，因此最好的预测方法是朴素法。

1.2. 通过差分获得平稳序列

当一阶差分的均值不为零时，

$$y_t - y_{t-1} = c + \varepsilon_t, \quad c \neq 0$$

则有

$$y_t = c + y_{t-1} + \varepsilon_t$$

此模型称为**带漂移项的随机游走（random walk with drift）**。当 $c > 0$ 时，y_t 向上方漂移，反之则向下方漂移。

此时最好的预测方法是漂移法。

1.2. 通过差分获得平稳序列

我们看到差分可以消除趋势，那么如果原序列具有季节性，则需要用**季节性差分（seasonal difference）**去消除。当季节性周期为 m 时，季节性差分定义为

$$y'_t = y_t - y_{t-m}$$

也称为**滞后-m期差分（lag-m difference）**。

如果季节性差分后的序列为白噪声，则原序列服从

$$y_t = y_{t-m} + \varepsilon_t$$

此时最好的预测方法是季节性朴素法。

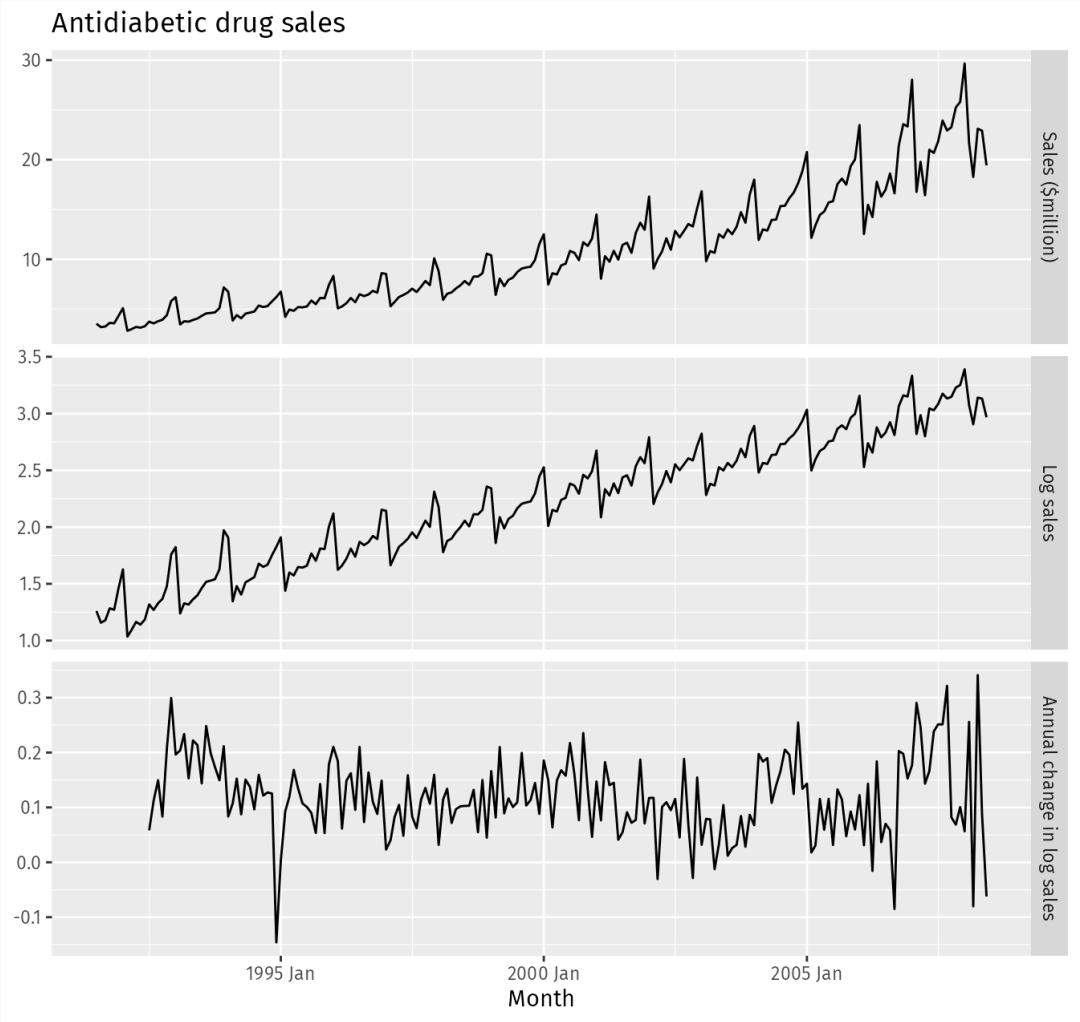
`difference()` 函数可以指定滞后期 `lag` 和差分阶数 `differences`，例如

```
difference(y, lag = 4, differences = 2) # 滞后4期的二阶差分
```

1.2. 通过差分获得平稳序列

观察澳大利亚糖尿病药物的月度销售数据（上图）可知，该序列中既存在趋势和季节性（均值变化），同时季节性变动的幅度也随着时间的推移而增大（方差变化）。

我们首先可以通过对数变换消除方差的变化（中图），然后通过季节性差分（也就是求同比变化值）消除趋势和季节性（下图）。



1.3. 需要差分几次？—— 单位根检验

在前面的学习中，我们介绍了用 Ljung-Box 检验判断序列是否是白噪声。那么，是否有统计检验能够帮助我们判断一个序列是否平稳呢？答案是肯定的，这类检验被称为**单位根检验（unit root tests）**。

单位根检验的原理超出了本门课程的水平，因此不做详细介绍。传统的单位根检验是 Dickey-Fuller 检验，而教科书中推荐的检验是 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) 检验。

KPSS 检验的假设为：

H_0：序列是平稳的
H_1：序列是不平稳的

KPSS 检验可以通过 `features()` 函数结合 `unitroot_kpss()` 函数实现。

1.3. 需要差分几次？—— 单位根检验

对 Google 股票收盘价序列进行 KPSS 检验

```
google_2015 <- gafa_stock |>
  filter(Symbol == "GOOG", year(Date) == 2015)
google_2015 |>
  features(Close, unitroot_kpss) # 收盘价保存在 Close 列中
```

```
# A tibble: 1 × 3
  Symbol kpss_stat kpss_pvalue
  <chr>       <dbl>       <dbl>
1 GOOG         3.56        0.01
```

这里需要注意的是，当 p 值小于 0.01 时，结果中都会显示为 0.01，而当 p 值大于 0.1 时，则都会显示为 0.1。因此，以上结果代表在 1% 显著性水平下拒绝零假设。即收盘价序列是不平稳的。

1.3. 需要差分几次？—— 单位根检验

对于 KPSS 检验为不平稳的序列，我们可以将其差分后，再进行 KPSS 检验。重复此过程直至检验结果提示平稳性，就可以知道使不平稳学列变成平稳序列所需的最小差分阶数。

但是 `unitroot_ndiffs()` 和 `unitroot_nsdiffs()` 函数提供了更加方便的功能，可以直接显示最佳差分阶数和最佳季节性差分阶数。

```
google_2015 |>
  features(Close, unitroot_ndiffs)
```

```
# A tibble: 1 × 2
  Symbol ndiffs
  <chr>   <int>
1 GOOG        1
```

结果显示最佳差分阶数为一阶。

1.4. 利用后移算子表达差分

为方便滞后项的表达，我们可以定义**后移算子（backward shift operator）**B，使 y_t 的一阶滞后项表达为

$$By_t = y_{t-1}$$

可以理解为对 y_t 实施 B 运算，就能获得一阶滞后项 y_{t-1}。

后移算子的方便之处是它虽然是一种函数，但可以像变量一样操作。例如

$$\begin{aligned}
B(By_t) &= B^2 y_t = y_{t-2} \\
y'_t &= y_t - y_{t-1} = y_t - By_t = (1-B)y_t \\
y''_t &= y_t - 2y_{t-1} + y_{t-2} = \left(1 - 2B + B^2\right)y_t = (1-B)^2 y_t
\end{aligned}$$

因此，d 阶差分可以表达为 $(1-B)^d y_t$。

1.4. 利用后移算子表达差分

m 阶季节性差分后再进行一次一阶差分：

$$\begin{aligned}(1-B)(1-B^m)y_t &= \left(1-B-B^m+B^{m+1}\right)y_t \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1} \\ &= (y_t - y_{t-1}) - (y_{t-m} - y_{t-m-1}) \\ &= (1-B)y_t - B^m(1-B)y_t \\ &= (1-B^m)(1-B)y_t\end{aligned}$$

最后的表达式代表一阶差分后再进行 m 阶季节性差分。因此普通差分和季节性差分的顺序并不影响结果。

2. 自回归模型和移动平均模型

2.1. 自回归模型

在回归模型的学习中，我们用预测变量 x_t 预测 y_t。类似的，我们也可以用 y_t 的滞后项（也就是历史观测值）作为预测变量。这种回归模型被称为**自回归（autoregression, autoregressive model）**。

p 阶自回归模型，即 AR(p) 模型，可以写成

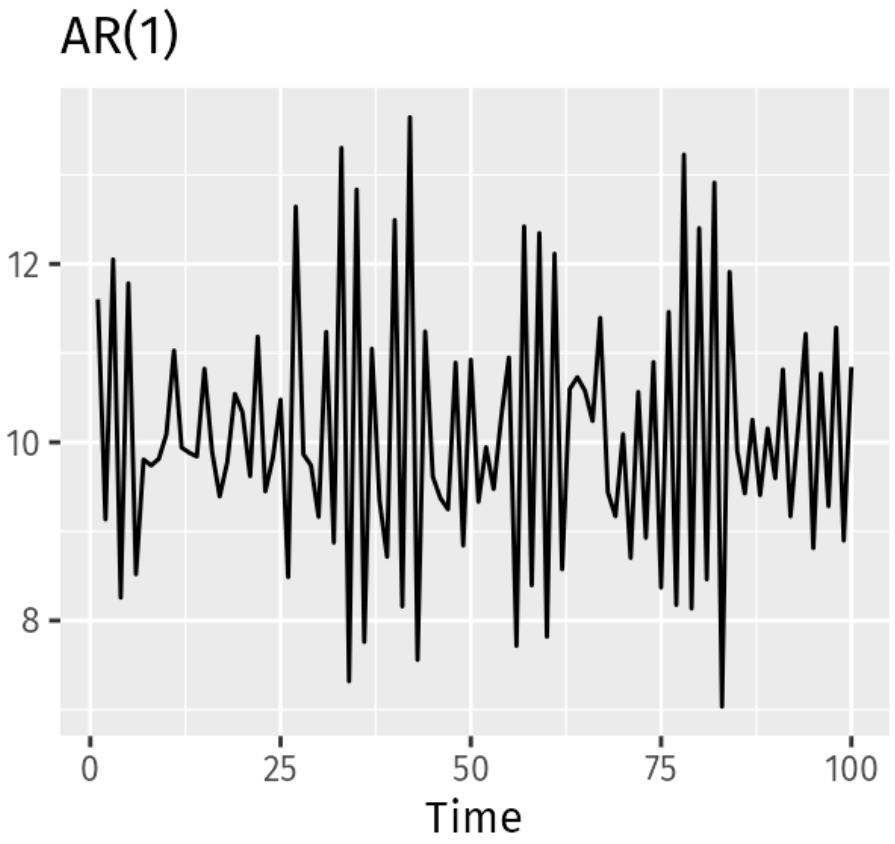
$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

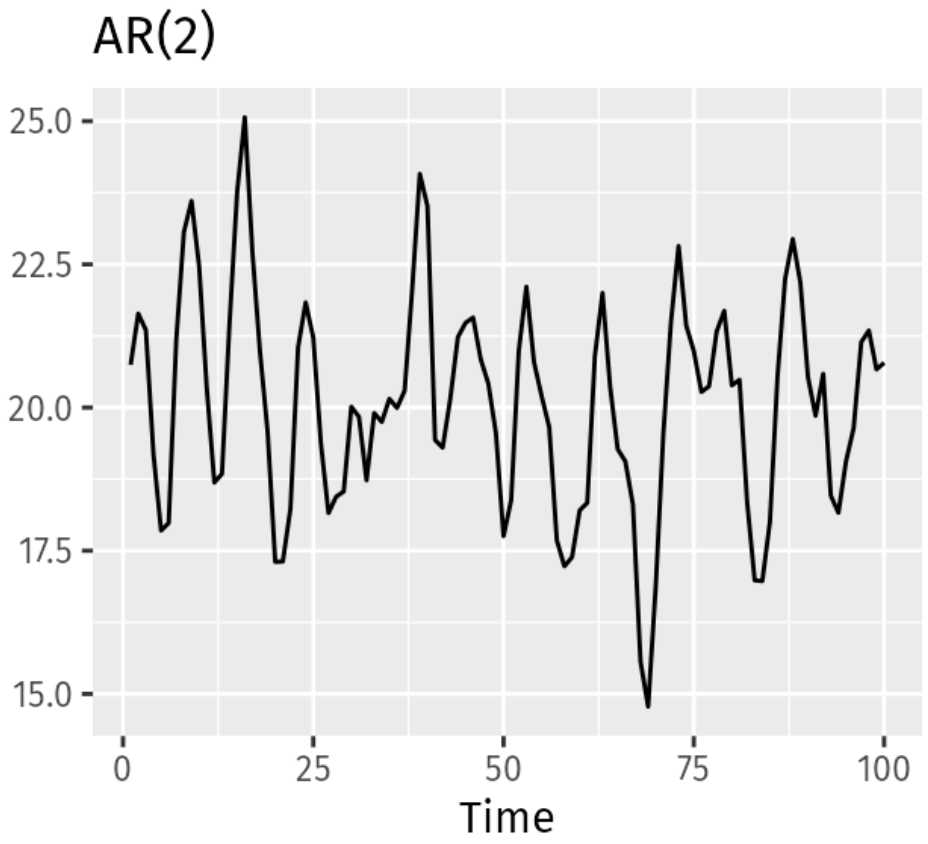
其中 ε_t 是白噪声。

不同的系数组合 $\phi_t, \ldots, \phi_p$ 可以产生不同的时间序列特征，因此自回归模型有广泛的应用场景。但通常我们将其限定在平稳时间序列上，这等于给系数设定了约束条件。例如

- AR(1) 模型：$-1 < \phi_1 < 1$
- AR(2) 模型：$-1 < \phi_2 < 1, \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$

2.1. 自回归模型





左图：$y_t = 18 - 0.8y_{t-1} + \varepsilon_t$，右图：$y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + \varepsilon_t$，$\varepsilon_t \sim N(0,1)$。

2.2. 移动平均模型

另一类类似回归的模型是利用过去的误差项作为预测变量，即

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

我们称之为 MA(q) 模型，也就是 q **阶移动平均（moving average）模型**。注意这里的移动平均模型和进行成份分解中用到的移动平均法不是一回事。

任意的平稳 AR(p) 模型都可以写成 MA(∞) 的形式，例如 AR(1) 模型

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t \\ &= \phi_1 (\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \ldots \\ &= \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \ldots \end{aligned}$$

2.2. 移动平均模型

如果对 MA(q) 模型的系数加以限制，我们也可以将其写成 AR(∞) 模型的形式，此时的 MA(q) 模型是**可逆的（invertible）**。例如 MA(1) 模型

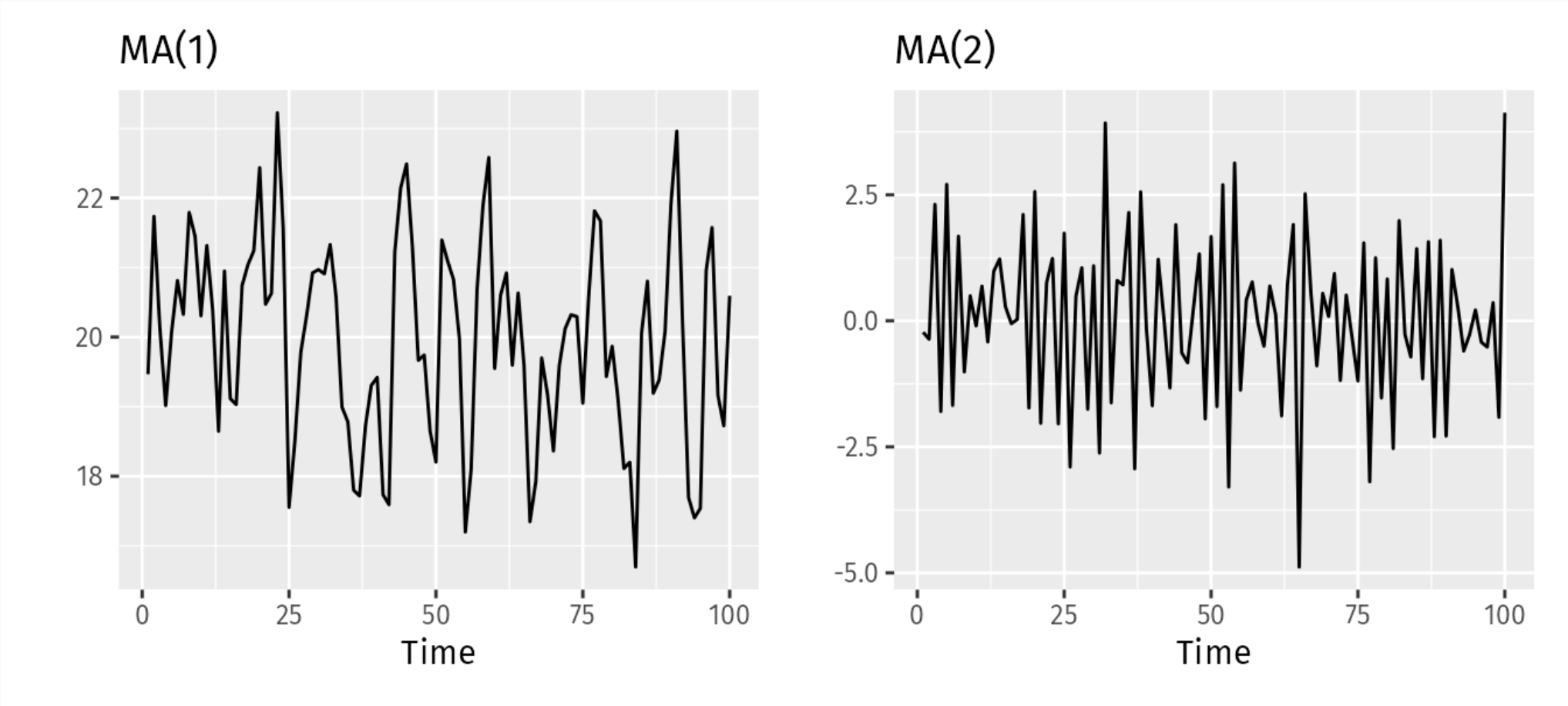
$$
\begin{aligned}
y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad \Leftrightarrow \quad \varepsilon_t &= -\theta_1 \varepsilon_{t-1} + y_t \\
&= -\theta_1(-\theta_1 \varepsilon_{t-2} + y_{t-1}) + y_t \\
&= (-\theta_1)^2 \varepsilon_{t-2} + (-\theta_1) y_{t-1} + y_t \\
&= \dots \\
&= y_t + (-\theta_1) y_{t-1} + (-\theta_1)^2 y_{t-2} + \dots
\end{aligned}
$$

当且仅当 $-1 < \theta_1 < 1$ 时最后一行的表达才有意义。

MA 模型的可逆性条件和 AR 模型的平稳性条件类似

- MA(1) 模型：$-1 < \theta_1 < 1$
- MA(2) 模型：$-1 < \theta_2 < 1, \theta_2 + \theta_1 > -1, \theta_1 - \theta_2 < 1$

2.2. 移动平均模型



左图：$y_t = 20 + \varepsilon_t + 0.8\varepsilon_{t-1}$，右图：$y_t = \varepsilon_t - \varepsilon_{t-1} + 0.8\varepsilon_{t-2}$，$\varepsilon_t \sim N(0,1)$。

3. ARIMA 模型

3. ARIMA 模型

如果我们把差分序列、 AR 模型、以及 MA 模型结合起来，就可以得到（非季节性）**ARIMA 模型**。ARIMA 是 AutoRegressive Integrated Moving Average 的缩写，这里的 integrated 指的是差分的逆向操作（也就是加总）。

ARIMA(p, d, q) 模型指 d 阶差分后的序列包含 p 阶自回归项和 q 阶移动平均项，即

$$y_t^{(d)} = c + \phi_1 y_{t-1}^{(d)} + \ldots + \phi_p y_{t-p}^{(d)} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

很多已知模型都可以看作 ARIMA 模型的特殊形式：

白噪声	ARIMA(0,0,0), $c = 0$
随机游走	ARIMA(0,1,0), $c = 0$
带漂移项的随机游走	ARIMA(0,1,0), $c \neq 0$
自回归	ARIMA(p,0,0)
移动平均	ARIMA(0,0,q)

3. ARIMA 模型

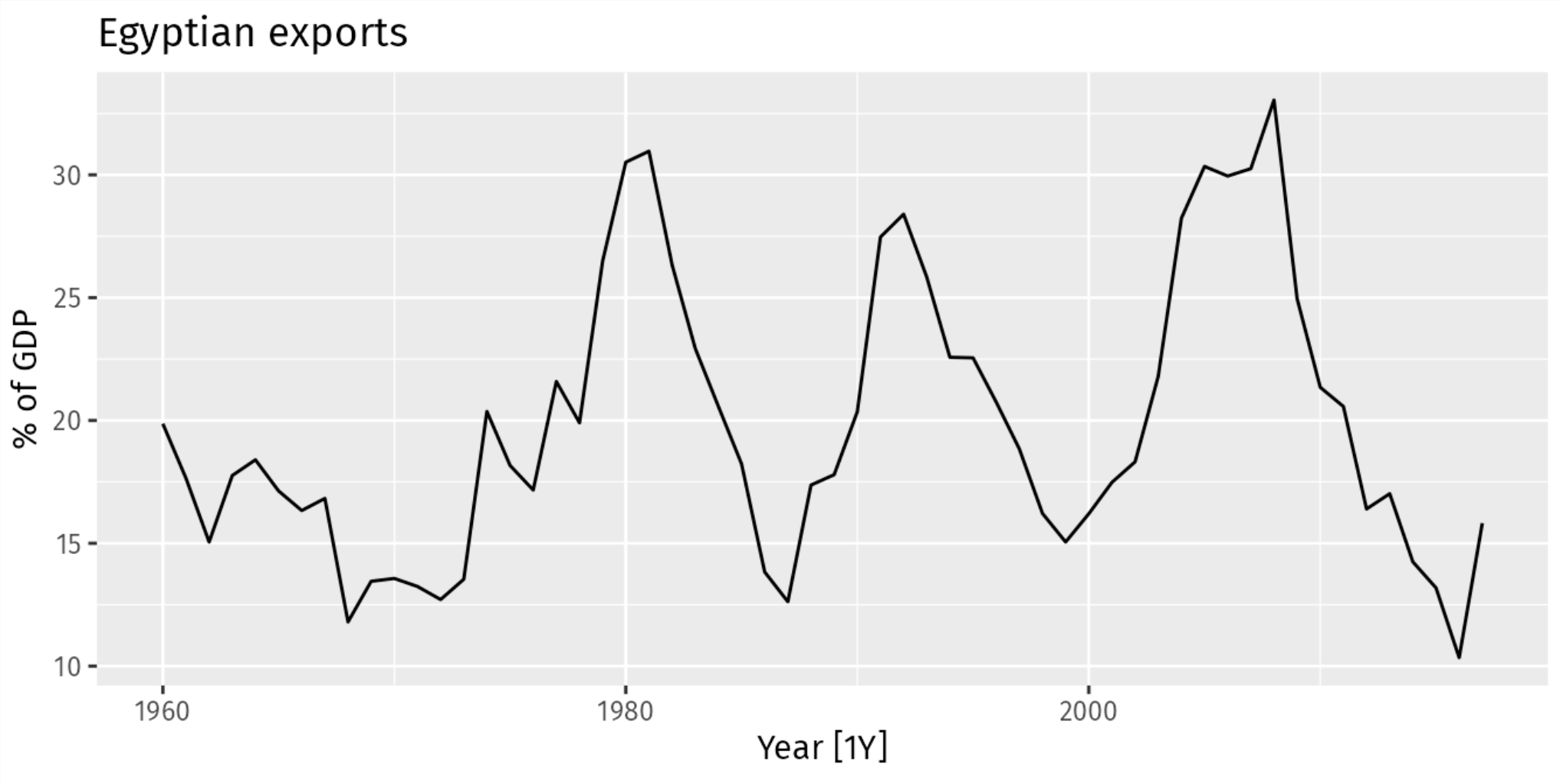
后移算子可以帮助我们更方便地表达 ARIMA(p, d, q) 模型

$$\underbrace{\left(1 - \phi_1 B - \ldots - \phi_p B^p\right)}_{\uparrow} \quad \underbrace{(1 - B)^d y_t}_{\uparrow} = c + \underbrace{\left(1 + \theta_1 B + \ldots + \theta_q B^q\right)}_{\uparrow} \varepsilon_t$$

$$\text{AR}(p) \qquad d \text{ differences} \qquad \text{MA}(q)$$

在实践中，如何正确选择 p, d, q 的取值至关重要。fable 包提供的 **`ARIMA()`** 函数在拟合模型的同时，也可以根据 AICc 自动判断合适的阶数。需要注意的是，**`ARIMA()`** 函数在默认状态并不会比较所有可能的阶数组合，因此最好是在了解它的工作原理的基础上，结合其他方法进行综合判断。详细信息参见教科书第 9.7 节。

3.1. 埃及的出口额数据



3.1. 埃及的出口额数据

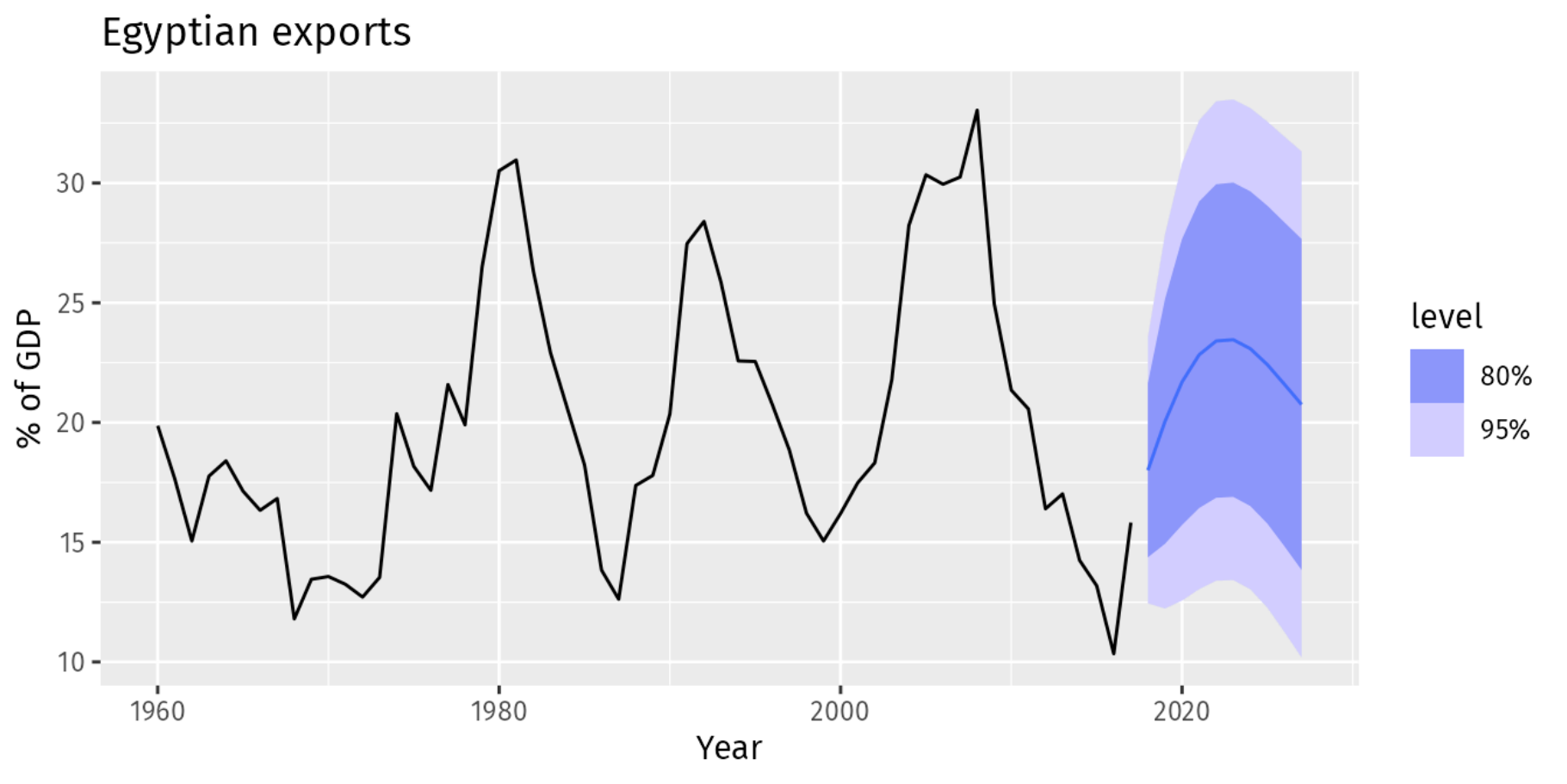
```
fit <- global_economy |>
  filter(Code == "EGY") |>
  model(ARIMA(Exports))
report(fit)
```

```
Series: Exports
Model: ARIMA(2,0,1) w/ mean

Coefficients:
         ar1      ar2      ma1  constant
      1.6764  -0.8034  -0.6896    2.5623
s.e.  0.1111   0.0928   0.1492    0.1161

sigma^2 estimated as 8.046:  log likelihood=-141.57
AIC=293.13   AICc=294.29   BIC=303.43
```

`ARIMA()` 函数自动选择了 ARIMA(2,0,1) 模型（$c \neq 0$）。

3.1. 埃及的出口额数据

对未来 10 期的预测结果：



3.2. 自相关系数和偏自相关系数

另一种判断 AR 和 MA 阶数的方法是利用自相关系数（ACF）和偏自相关系数（PACF）。

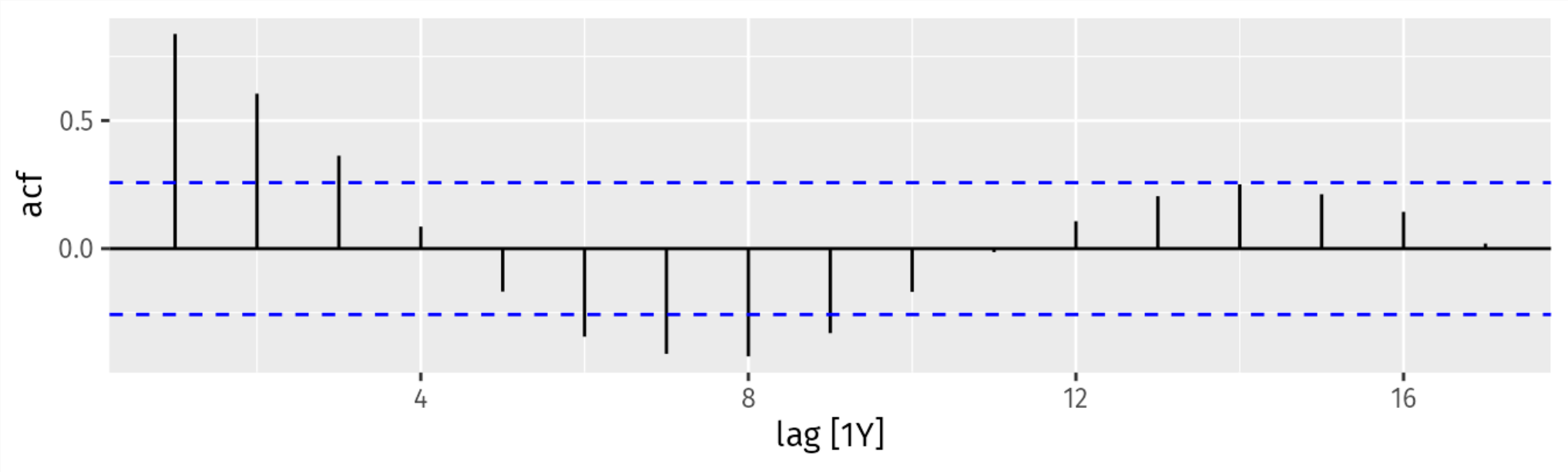
自相关系数可以衡量 y_t 和任意 y_{t-k} 之间的线性相关关系。但是当 y_t 和 y_{t-1} 相关时，y_{t-1} 和 y_{t-2} 也相关，这会导致 y_t 和 y_{t-2} 间存在相关性。因此，单纯依靠二阶自相关系数无法判断二阶滞后项是否真正在预测中起作用。

偏自相关系数（partial autocorrelation coefficient）衡量的是移除了 $y_{t-1}, \ldots, y_{t-k+1}$ 影响的情况下， y_t 和 y_{t-k} 间的线性相关关系。具体的说，k 阶偏自相关系数是 AR(k) 模型中 k 阶滞后项的系数估计值（实际采用更加高效的方法计算）。

偏自相关系数可以通过 `PACF()` 函数求得。

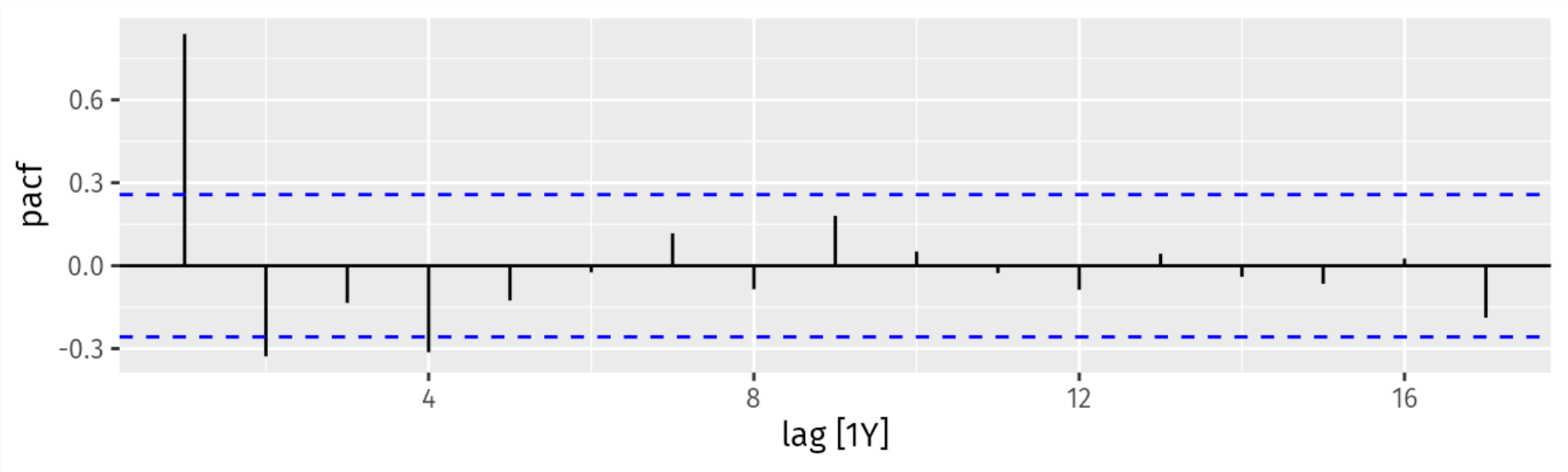
3.2. 自相关系数和偏自相关系数

```
global_economy |>
  filter(Code == "EGY") |>
  ACF(Exports) |> # 求埃及出口额序列的自相关系数
  autoplot()
```



3.2. 自相关系数和偏自相关系数

```
global_economy |>
  filter(Code == "EGY") |>
  PACF(Exports) |> # 求埃及出口额序列的偏自相关系数
  autoplot()
```

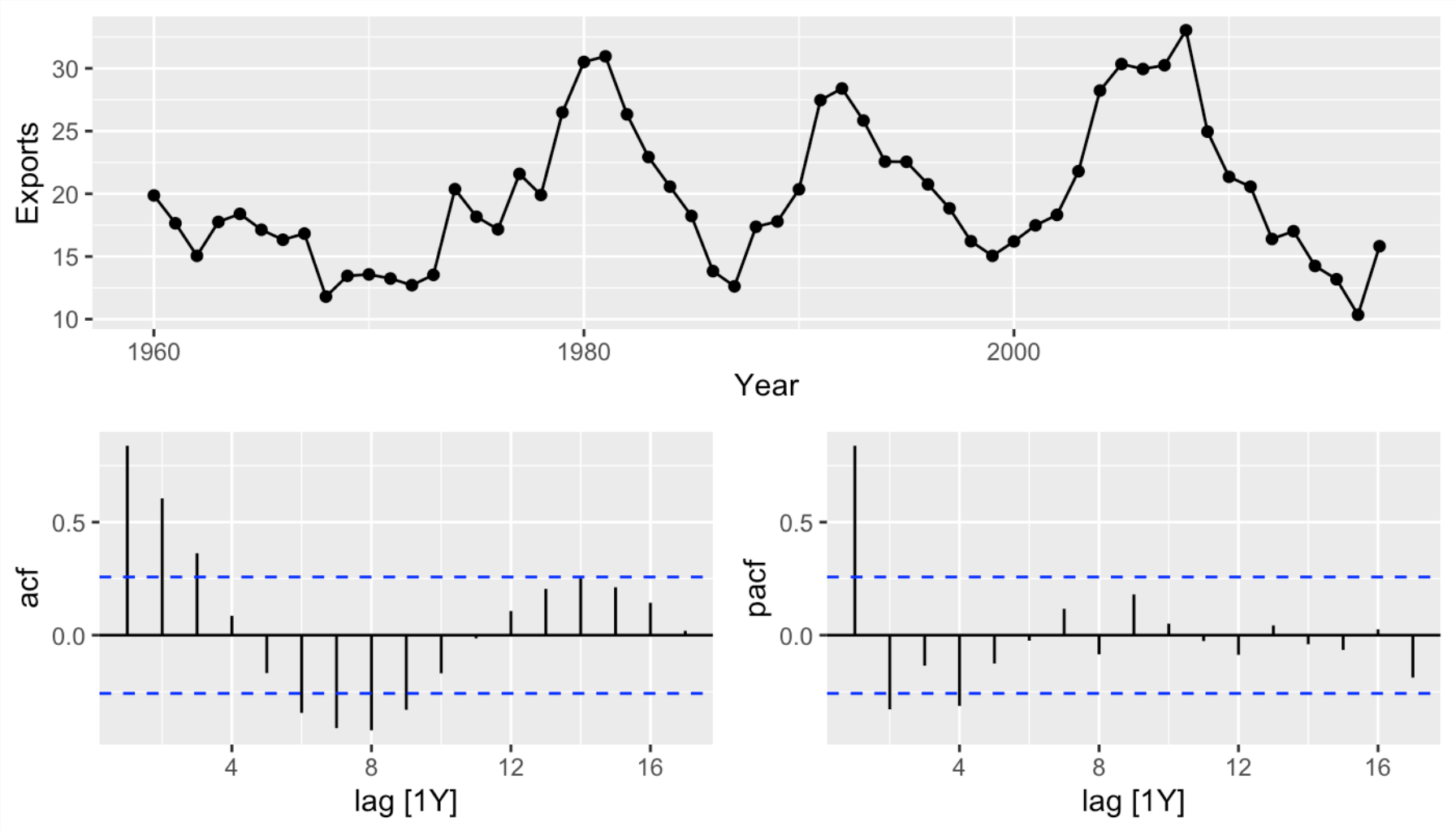


3.2. 自相关系数和偏自相关系数

利用 `gg_tsdisplay()` 的 `plot_type = "partial"` 设定可以将时序图、ACF 图和 PACF 图同时绘制。例如

```
global_economy |>
  filter(Code == "EGY") |>
  gg_tsdisplay(Exports, plot_type = "partial")
```

3.2. 自相关系数和偏自相关系数



如何利用 ACF 和 PACF 判断模型阶数

1. 首先确定差分阶数 d 并取d 次差分。
2. 差分后的数据如果呈现下列特征
   - ACF 呈指数衰减或正弦曲线波动
   - PACF 在 p 阶处有明显凸起，但在 p 阶以上处则没有

   则原数据可能服从 ARIMA(p, d, 0) 模型。
3. 差分后的数据如果呈现下列特征
   - PACF 呈指数衰减或正弦曲线波动
   - ACF 在 q 阶处有明显凸起，但在 q 阶以上处则没有

   则原数据可能服从 ARIMA(0, d, q) 模型。

3.2. 自相关系数和偏自相关系数

```
fit2 <- global_economy |>
  filter(Code == "EGY") |>
  model(ARIMA(Exports ~ pdq(4,0,0))) # 用 pdq() 函数指定阶数
report(fit2)
```

```
Series: Exports
Model: ARIMA(4,0,0) w/ mean

Coefficients:
         ar1      ar2     ar3      ar4  constant
      0.9861  -0.1715  0.1807  -0.3283    6.6922
s.e.  0.1247   0.1865  0.1865   0.1273    0.3562

sigma^2 estimated as 7.885:  log likelihood=-140.53
AIC=293.05   AICc=294.7   BIC=305.41
```

ARIMA(4,0,0) 模型的 AICc 值为 294.7，而 ARIMA(2,0,1) 模型的 AICc 值为 294.29。

3.3. ARIMA 模型的分析流程

利用 ARIMA 模型分析非季节性时间序列数据时，可以参考下面的流程进行

1. **绘制时序图**，观察数据并检查有无异常值。
2. 如有必要，通过数学**变换**（如 Box-Cox）消除方差的变化。
3. 如果数据不平稳，通过**差分**使其平稳。
4. 检查 ACF 和 PACF，并判断合适的阶数（**模型选择**）。
5. 尝试**拟合**你选择的模型，然后通过 AICc 寻找更好的模型
6. 通过残差诊断图或 Ljung-Box 检验**判断残差序列是否为白噪声**。如果不是，那么换一个模型继续尝试。
7. 当残差符合白噪声特征时，即可进行**预测**。

更多信息可参考教科书 9.7 节。

3.4. 适应季节性数据的 ARIMA 模型

虽然 ARIMA(p,d,q) 模型无法应对季节性，但是通过对 ARIMA 模型进行修改，即可获得适应季节性数据的 ARIMA 模型，通常称为 Seasonal ARIMA 或 SARIMA。

SARIMA 就是在 ARIMA 的基础上通过季节性差分等方式消除季节性的影响，它由非季节性成份和季节性成份组成，可以写成

$$\underset{\text{非季节性成份}}{\text{ARIMA}(p,d,q)}\underset{\text{季节性成份}}{(P,D,Q)_m}$$

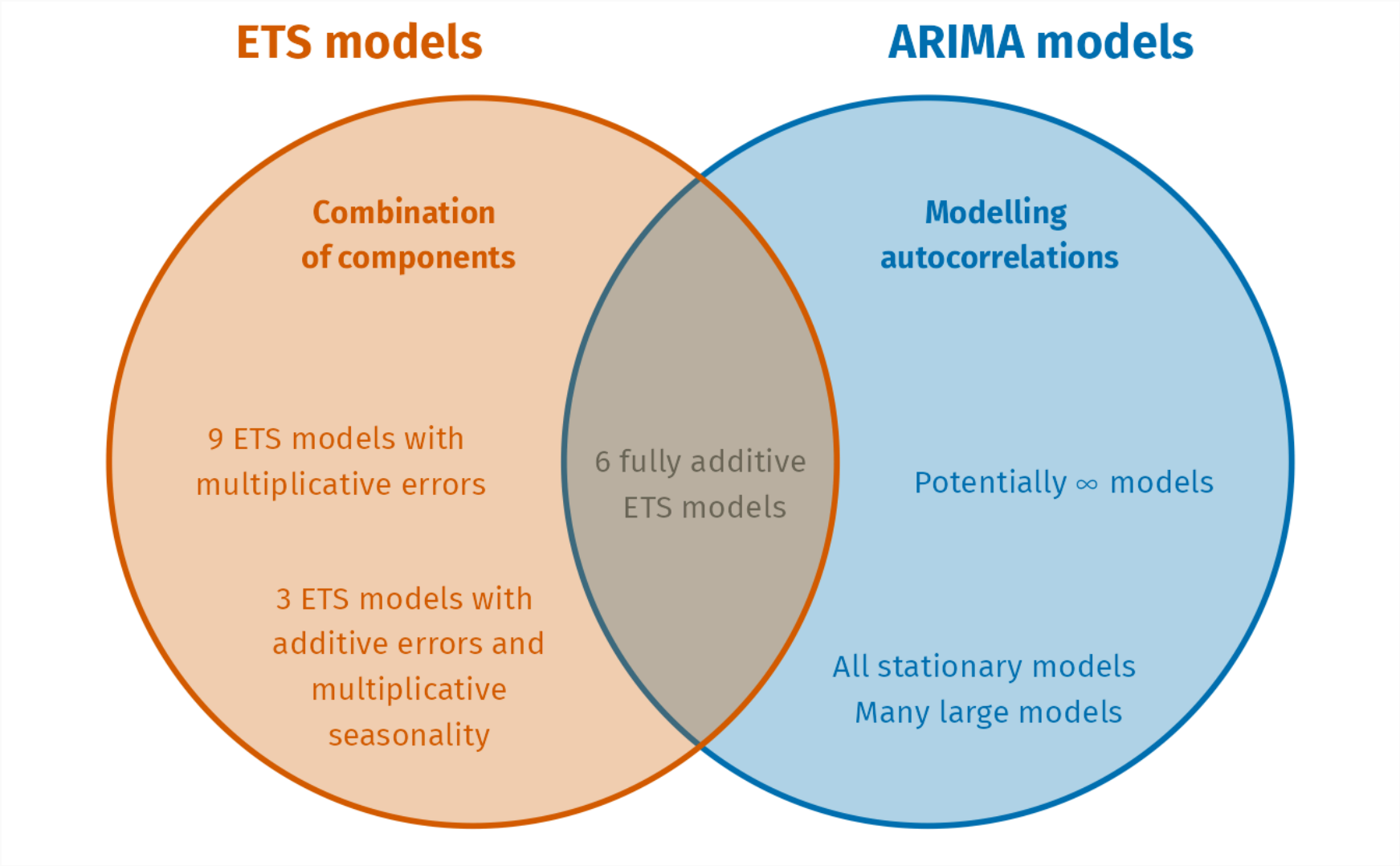
例如，ARIMA(1,1,1)(1,1,1)$_4$ 的表达式为

$$(1-\phi_1 B)\left(1-\Phi_1 B^4\right)(1-B)\left(1-B^4\right)y_t = (1+\theta_1 B)\left(1+\Theta_1 B^4\right)\varepsilon_t$$

更多信息可参考教科书 9.9 节。

4. ARIMA 和 ETS

4. ARIMA 和 ETS


ETS models
ARIMA models
Combination of components
Modelling autocorrelations
9 ETS models with multiplicative errors
6 fully additive ETS models
Potentially ∞ models
3 ETS models with additive errors and multiplicative seasonality
All stationary models
Many large models

ETS 模型和 ARIMA 模型间的等价关系

ETS 模型	ARIMA 模型	参数
ETS(A,N,N)	ARIMA(0,1,1)	$\theta_1 = \alpha - 1$
ETS(A,A,N)	ARIMA(0,2,2)	$\theta_1 = \alpha + \beta - 2,$ $\theta_2 = 1 - \alpha$
ETS(A,Ad,N)	ARIMA(1,1,2)	$\phi_1 = \phi,$ $\theta_1 = \alpha + \phi\beta - 1 - \phi,$ $\theta_2 = (1 - \alpha)\phi$
ETS(A,N,A)	ARIMA$(0,1,m)(0,1,0)_m$	
ETS(A,A,A)	ARIMA$(0,1,m+1)(0,1,0)_m$	
ETS(A,Ad,A)	ARIMA$(1,0,m)(0,1,0)_m$	

4. ARIMA 和 ETS

那么如何在 ETS 和 ARIMA 模型间做出选择呢?

AICc 只能用于同一类模型间的比较，当比较 ETS 和 ARIMA 模型时，我们需要利用 RMSE 或 MAE 等误差测度。此时可以有两种操作方式：

1. 当样本量足够大时，可将数据分成训练集和测试集，并利用测试集计算误差测度值。
2. 当样本量比较小时，可利用交叉验证法（cross-validation，详见教科书第 5.10 节）。

更多信息可参考教科书 9.10 节。

5. 课后练习

5. 课后练习

- 学习教科书第 9 章（ARIMA Models）中的内容，并尝试在自己的电脑上复现书中的结果。

- tsibbledata 包中的 `pelt` 数据集包含了加拿大 Hudson Bay Company 在 1845–1935 年间交易的雪靴兔（snowshoe hare，保存在 `Hare` 列中）和加拿大猞猁（Canadian lynx，保存在 `Lynx` 列中）的毛皮数量。利用雪靴兔的交易量数据回答下列问题：

  1. 绘制时序图。此序列是稳定的吗?
  2. 利用 KPSS 检验、ACF、PACF 说明为什么 ARIMA(4,0,0) 模型是合适的。
  3. 拟合模型，并预测未来 5 年的交易量。