VAR 模型金融数据应用实例

美国标普指数与我国沪深指数联动性

```r
# 加载 R 软件包
library(showtext)
library(WindR)
library(TSA)
library(MTS)
library(vars)
library(dplyr)

# 启动 Wind 量化平台
w.start()
```

VAR(vector autoregressive) 模型在探索多元时间序列之间相关性的问题中有广泛的应用。VAR(p) 模型的基本设定如下

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}_1\mathbf{X}_{t-1} + \cdots + \mathbf{A}_p\mathbf{X}_{t-p} + \epsilon_t$$

这里 $\mathbf{X}_t$，$\mathbf{c}$ 以及 ϵ_t 均为 $d\times 1$ 向量，而 $\mathbf{A}_1, ..., \mathbf{A}_p$ 均为 $d\times d$ 矩阵。p 决定了多元时间序列模型中作为解释变量的滞后项的阶数，而下标 $t, t-1, ..., t-p$ 则可理解为时间。假设我们可以观测到时间序列 $\{\mathbf{X}_1, ..., \mathbf{X}_T\}$，而无法观测到白噪声扰动项 $\{\epsilon_1, ..., \epsilon_T\}$。我们希望估计模型中的参数 $(\mathbf{c}, \mathbf{A}_1, ..., \mathbf{A}_p)$。

以下面 $d=2$ 的 VAR(2) 模型为例，如果将矩阵表达式展开，我们可以得到

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} \end{pmatrix} \begin{pmatrix} X_{t-2,1} \\ X_{t-2,2} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix}$$

这里 $a_{ij}^{(k)}$ 表示系数矩阵 $\mathbf{A}_k$ 的第 (i,j) 元素。显然，如果我们能够将系数矩阵 $(\mathbf{A}_1, \mathbf{A}_2)$ 中的所有 $a_{ij}^{(k)}$ 元素均估计出来，则我们就可以推断任意 $X_{t,i}, i=1,2$ 如何受到 $X_{t-1,1}, X_{t-1,2}, X_{t-2,1}, X_{t-2,2}$ 影响，并作出相应预测。相较于一元时间序列模型，多元模型使得研究多元时间序列之间的〝交叉〞影响成为可能。

VAR(p) 模型的估计非常简单，主要的估计方法包括最小二乘法估计与极大似然法估计，两种方法均有相应的 R 软件包可以让使用者通过简单地输入数据与命令完成估计，并获得估计结果。

- 最小二乘法估计：使用者需要先安装并调用 vars 软件包，然后使用其 VAR() 命令。
- 极大似然法估计：使用者需要先安装并调用 MTS 软件包，然后使用其 VARMA() 命令。

上述两种估计方法的 R 命令的语法和输出结果非常类似（而且在满足一定条件下，估计的结果也在理论上是一样的），这里我们只演示如何使用最小二乘法来估计美国 S&P500（SP500）与我国沪（SH）深（SZ）指数之间的联动关系。具体而言，我们的模型中的 $\mathbf{X}_t$ 设定为 3×1 向量 $(SP500_t, SH_t, SZ_t)$，而模型的阶数 p 可以通过 AIC，BIC 或 HQIC 等信息准则自动选取（在 VAR() 命令中设定）。

首先，我们通过 Wind 量化平台（Wind Data Feed Services，WDFS）下载 S&P500 指数与我国沪深指数自 2011.1.1 至 2014.12.31 之间的每日收盘值，并计算其各自的日〝收益〞率。

```r
# 使用 WDFS 下载原始数据并产生“收益”数据
data <- w.wsd("000001.SH,399106.SZ,SP500.SPI","close",
              "2011-01-01","2014-12-31","TradingCalendar=NYSE")$Data %>%
  rename(SH = `000001.SH`, SZ = `399106.SZ`, SP500 = `SP500.SPI`,
         date = `DATETIME`) %>%
  mutate(R.SH = (SH-lag(SH, n = 1L))/lag(SH, n = 1L),
         R.SZ = (SZ-lag(SZ, n = 1L))/lag(SZ, n = 1L),
```

```
        R.SP500 = (SP500-lag(SP500, n = 1L))/lag(SP500, n = 1L)) %>%
  select(R.SP500, R.SH, R.SZ) %>%
  filter(!is.na(R.SP500*R.SH))
```

使用已经准备好的多元时间序列数据，我们可以用 VAR() 命令来估计相应的 VAR(p) 模型。在下面的命令中，ic = ”SC” 表示采用 BIC（即 Schwarz criterion）来选取最恰当的 p，lag.max = 3 表示限定 p 最大值为 3，两者同时使用即表示我们首先比较 VAR(1)，VAR(2) 和 VAR(3) 三个模型的 BIC，然后选取 BIC 最小的模型来作为最终选定的模型，并输出其结果。最小二乘法估计的结果将以 List 格式存储于 var.bic 变量中，而后我们使用 summary() 命令来输出（更完整的）估计结果。

```
# 使用 VAR() 命令估计 VAR(p) 模型
var.bic <- VAR(data, lag.max = 3, ic = "SC")
summary(var.bic)
```

```
##
## VAR Estimation Results:
## =========================
## Endogenous variables: R.SP500, R.SH, R.SZ
## Deterministic variables: const
## Sample size: 1004
## Log Likelihood: 9894.493
## Roots of the characteristic polynomial:
## 0.1581 0.09534 0.002063
## Call:
## VAR(y = data, lag.max = 3, ic = "SC")
##
##
## Estimation results for equation R.SP500:
## ========================================
## R.SP500 = R.SP500.l1 + R.SH.l1 + R.SZ.l1 + const
##
##              Estimate Std. Error t value Pr(>|t|)
## R.SP500.l1 -0.0717810  0.0316558  -2.268   0.0236 *
## R.SH.l1     0.0542360  0.0530986   1.021   0.3073
## R.SZ.l1    -0.0455520  0.0432490  -1.053   0.2925
## const       0.0005654  0.0003074   1.839   0.0662 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.009724 on 1000 degrees of freedom
## Multiple R-Squared: 0.00602, Adjusted R-squared: 0.003038
## F-statistic: 2.019 on 3 and 1000 DF,  p-value: 0.1096
##
##
## Estimation results for equation R.SH:
## =====================================
## R.SH = R.SP500.l1 + R.SH.l1 + R.SZ.l1 + const
##
##              Estimate Std. Error t value Pr(>|t|)
## R.SP500.l1  1.814e-01  3.576e-02   5.072 4.69e-07 ***
## R.SH.l1    -5.301e-03  5.997e-02  -0.088    0.930
## R.SZ.l1    -2.276e-02  4.885e-02  -0.466    0.641
## const       9.421e-05  3.472e-04   0.271    0.786
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.01098 on 1000 degrees of freedom
## Multiple R-Squared: 0.02585, Adjusted R-squared: 0.02293
## F-statistic: 8.845 on 3 and 1000 DF,  p-value: 8.652e-06
##
##
## Estimation results for equation R.SZ:
## =====================================
## R.SZ = R.SP500.l1 + R.SH.l1 + R.SZ.l1 + const
##
##               Estimate Std. Error t value Pr(>|t|)
## R.SP500.l1  0.1883520  0.0438446   4.296 1.91e-05 ***
## R.SH.l1    -0.1628486  0.0735437  -2.214   0.0270 *
## R.SZ.l1     0.1377281  0.0599016   2.299   0.0217 *
## const       0.0000731  0.0004258   0.172   0.8637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.01347 on 1000 degrees of freedom
## Multiple R-Squared: 0.02243, Adjusted R-squared: 0.0195
## F-statistic: 7.648 on 3 and 1000 DF,  p-value: 4.674e-05
##
##
##
## Covariance matrix of residuals:
##           R.SP500      R.SH      R.SZ
## R.SP500 9.456e-05 8.568e-06 6.313e-06
## R.SH    8.568e-06 1.206e-04 1.262e-04
## R.SZ    6.313e-06 1.262e-04 1.814e-04
##
## Correlation matrix of residuals:
##         R.SP500    R.SH   R.SZ
## R.SP500 1.00000 0.08022 0.0482
## R.SH    0.08022 1.00000 0.8529
## R.SZ    0.04820 0.85293 1.0000
```

可以看出，系数矩阵中很多 $a_{ij}^{(k)}$ 在统计上并不显著，我们可以使用 restrict() 命令重新对模型进行估计，在新的估计中，之前不显著的系数被设定为为 0，而仅估计之前统计上显著的系数。

```
# 使用 restrict() 命令估计受限模型
restrict(var.bic)
```

```
##
## VAR Estimation Results:
## =======================
##
## Estimated coefficients for equation R.SP500:
## ============================================
## Call:
## R.SP500 = R.SP500.l1
##
```

```
##  R.SP500.l1
## -0.06656503
##
##
## Estimated coefficients for equation R.SH:
## ==========================================
## Call:
## R.SH = R.SP500.l1
##
## R.SP500.l1
##   0.180275
##
##
## Estimated coefficients for equation R.SZ:
## ==========================================
## Call:
## R.SZ = R.SP500.l1 + R.SH.l1 + R.SZ.l1
##
## R.SP500.l1    R.SH.l1    R.SZ.l1
##  0.1887594 -0.1627676  0.1377290
```

下面，我们对模型估计的结果加以总结（注意我们这里使用的均为 "收益" 数据）。可以看出我国沪深指数明显受到美国 S&P500 指数的影响，而 S&P500 指数几乎与沪深指数无关，因此可以将 S&P500 指数视为我国沪深指数的先行指标。此外，我国沪指似乎并不受深指影响，而沪指却可被视为深指的领先指标。

$$\begin{aligned}
SP500_t &= -0.067 \times SP500_{t-1} \\
SH_t &= 0.180 \times SP500_{t-1} \\
SZ_t &= 0.189 \times SP500_{t-1} - 0.163 \times SH_{t-1} + 0.138 \times SZ_{t-1}
\end{aligned}$$

当然，需要注意的是，模型的拟合度 R^2 很低，这说明该模型并不应被用于预测走势，而仅提供了不同指数之间相关性的信息。要获得更高的模型拟合，使用者应考虑加入更多解释变量（如制造业采购经理人指数 PMI，利率，货币供应等宏观经济数据）。

要了解更详细的有关 VAR 模型的理论及应用，请参考我的金融计量课程 Lecture 8-10（⇒ 课程主页）。