估计联合概率分布的简单方法

分析 S&P500 指数与上证综指收益率之间的联动性

1. 估计方法简介

金融市场上随处可见的随机性（如价格波动等）都可以用随机变量来加以描述。很多情况下，我们希望了解这些随机变量之间的相关性，以便于进行建模和预测。比较简单的研究相关性的方法是依据观测到的样本计算相关系数或进行回归分析。但是，这些方法通常只能用来研究随机变量之间的线性相关关系，而现实中很多的相关关系实际上是（高度）非线性的，很难用简单的线性关系来准确地近似。如果我们能够比较准确地估计一组随机变量（看作一个随机向量）的（联合概率）分布函数，那么我们理论上就可以对它们之间真实的相关关系进行准确的描述。

估计随机向量的分布函数可以采用参数方法，即预先假设随机向量的联合分布类型。这样一来，对其分布函数的估计就简化到估计有限个参数（比如，在多元正态分布假设下的均值向量和协方差矩阵），从而可以使用极大似然方法来进行估计。但是，在很多情况下，我们无法预先判断所要研究的随机向量的分布的类型。相对地，使用非参数方法估计随机向量的分布函数不需要对其分布函数形式作出假设，可以在相对较弱的条件下进行估计，其结果也相应地更加“稳健”（由于依赖更弱的假设，而更不容易犯错）。但是，相应地，非参数估计在操作上是比较麻烦的。我们在这里介绍一种相对简单易操作的估计随机向量分布函数的非参数方法[1]。

假设我们可以观测到一组来自分布 $f_X(x)$ 的随机样本 $x_1, ...x_N$。我们可以选定一个参考（reference）分布 $g_X(x)$[2]，然后使用 Monte Carlo 方法，抽取 $g_X(x)$ 的一组（M 个）随机样本 $x_{N+1}, ..., x_{N+M}$[3]。将两组样本混合在一起，并定义一个随机变量 Y，如果样本来自分布 $f_X(x)$，则 $Y = 1$，反之，如果样本来自分布 $g_X(x)$，则 $Y = 0$。

那么，依据概率理论，我们知道

$$\mu(x) = E(Y|X = x) = \frac{f_X(x)}{f_X(x) + g_X(x)} = \frac{f_X(x)/g_X(x)}{1 + f_X(x)/g_X(x)} = \frac{\exp(\phi(x))}{1 + \exp(\phi(x))} \tag{1}$$

其中 $\phi(x) = \log(f_X(x)/g_X(x))$[4]。注意到，由于 $g_X(x)$ 是我们选定的，所以其函数形式对于我们是已知的。如果我们能够得到一个关于函数 $\mu(x)$ 的比较准确的估计 $\widehat{\mu}(x)$，则我们就可以通过下面的公式得到关于 $f_X(x)$ 的估计 $\widehat{f_X}(x)$

$$\widehat{f}_X(x) = g_X(x)\left(\frac{\widehat{\mu}(x)}{1 - \widehat{\mu}(x)}\right) \tag{2}$$

注意到，(1) 式实际上可以看作一个包含了未知函数 $\phi(x)$ 的 Logistic 回归（即 Logit 模型），而由近似理论可知，当函数 $\phi(x)$ 足够平滑（“smooth”）的时候，我们总会是可以找到一个多项式很好地近似 $\phi(x)$[5]，即

$$\phi(x) \approx b^k(x)'\theta$$

那么，我们就有

$$\mu(x) = E(Y|X = x) = P(Y = 1|X = x) \approx \frac{\exp(b^k(x)'\theta)}{1 + \exp(b^k(x)'\theta)} \tag{3}$$

这样一来，(1) 中估计函数 $\phi(x)$ 的问题就被简化到估计一组参数（常数向量）θ，而这在 Logit 模型的框架下是很容易实现的（semiparametric logistic regression）。给定一个关于 θ 的估计 $\widehat{\theta}$，我们就可以得到

$$\widehat{\mu}(x) = \frac{\exp(b^k(x)'\widehat{\theta})}{1 + \exp(b^k(x)'\widehat{\theta})}$$

$$\widehat{f}_X(x) = g_X(x)\left(\frac{\widehat{\mu}(x)}{1 - \widehat{\mu}(x)}\right) = g_X(x)\exp(b^k(x)'\widehat{\theta}) \tag{4}$$

[1] 下面介绍的方法实际上是一个将无监督（unsupervised）统计学习问题转化为有监督（supervised）统计学习问题来处理的例子。

[2] $g_X(x)$ 的选择比较灵活，具体的选择应该考虑具体问题的性质和研究目的。常用的选择为 Uniform 和 Gaussian 分布。

[3] 一般应选择 $M \geq N$。

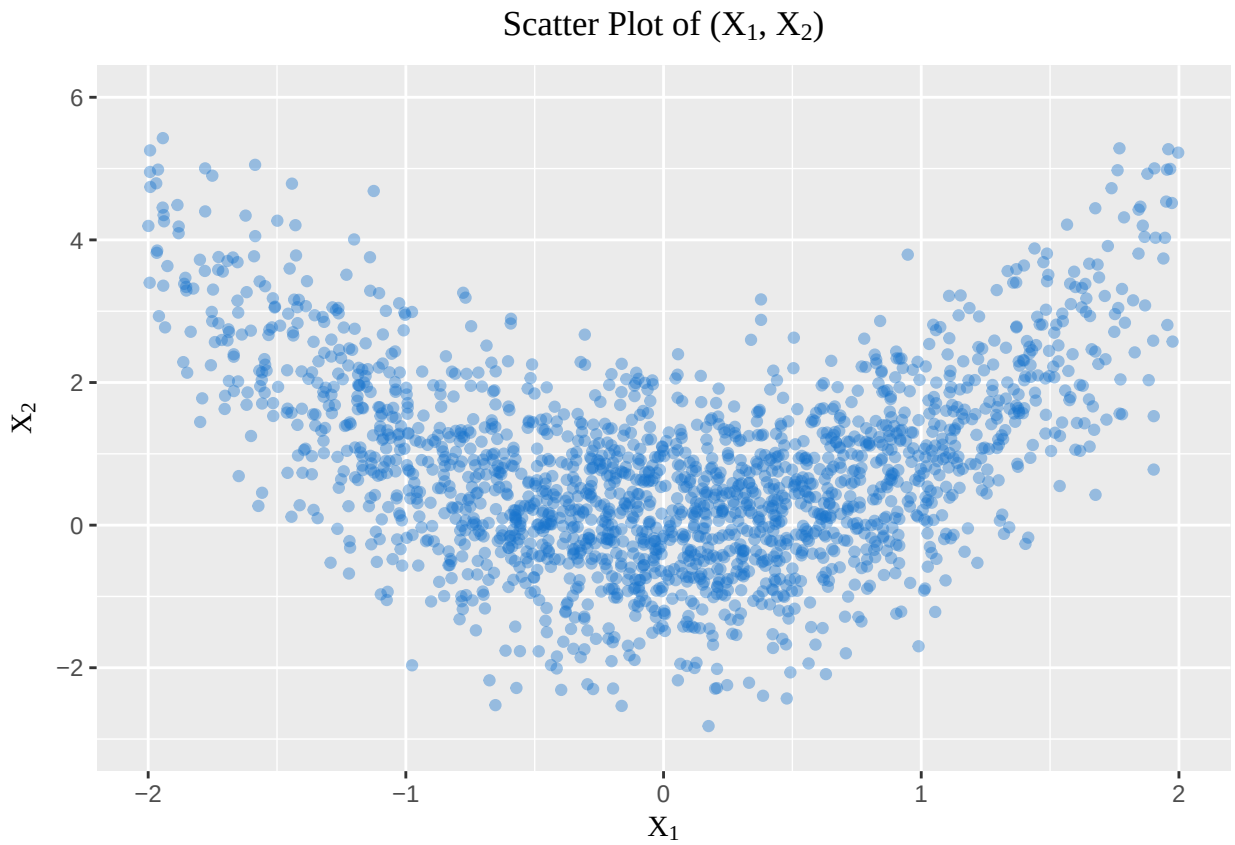
[4] 从这里可以看出，$g_X(x)$ 应该比 $f_X(x)$ 拥有更大的定义域（support）。

[5] 注意到下面公式里多项式的项数 k 应根据样本总量的大小来选取，样本总量越大，多项式应包含的项数越多，即 k 越大。

2. Monte Carlo 实验

这里，我们用 Monte Carlo 方法来验证前面所介绍的方法的有效性。考虑 $X_1 \sim N(0,1)$，$X_2 = X_1^2 + u$，其中 $u \sim N(0,1)$ 且独立于 X_1。注意到 X_1 与 X_2 的相关系数为 0[6]，但实际上它们之间明确地存在非线性相关关系（如下面的散点图所示）。我们希望用 $N = 2000$ 个 $X = (X_1, X_2)$ 的随机样本来估计 $f_X(x_1, x_2)$。

```
# 加载 R 软件包
library(showtext)
library(ggplot2)
library(latex2exp)

# 抽取随机样本
set.seed(123)
N <- 2000
X1 <- rnorm(N, 0, 1); X2 <- X1^2 + rnorm(N, 0, 1)

# 联合分布散点图
scatter <- ggplot(data.frame(X1, X2), aes(X1, X2))
scatter + geom_point(alpha = .4, size = 1.5, color = "dodgerblue3") +
  labs(title = TeX("Scatter Plot of ($X_1$, $X_2$)"),
       y = TeX("$X_2$"), x = TeX("$X_1$")) +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                  face = "bold")) +
  xlim(c(-2, 2)) + ylim(c(-3, 6))
```



[6] 因为 $Cov(X_1, X_2) = Cov(X_1, X_1^2 + u) = Cov(X_1, X_1^2) = E(X_1^3) - E(X_1)E(X_1^2) = 0$。

下面我们使用上一节介绍的方法来估计 $f_X(x_1,x_2)$，从而确切地了解 X_1 与 X_2 的联合分布特征。这里，我们选取 $g_X(x_1,x_2)=g_{X_1}(x_1)g_{X_2}(x_2)$ 作为参照分布的密度函数并从中抽取 $M=2000$ 个随机样本，其中 $g_{X_1}(x_1)$ 和 $g_{X_2}(x_2)$ 分别为 $U(-2,2)$ 和 $U(-3,6)$ 的概率密度函数（U 代表 Uniform，均匀分布）。我们假设 $\phi(x)$ 可以用下面的多项式来近似[7]

$$\phi(x)=\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1^2+\theta_4x_2^2+\theta_5x_1x_2+\theta_6x_1^3+\theta_7x_2^3+\theta_8x_1^2x_2+\theta_9x_1x_2^2$$

然后，我们使用 Nonlinear Least Squares (NLS) 方法来估计 θ。这里之所以没有使用极大似然方法，主要是为了避免直接取 log 时可能产生的近似 $-\infty$ 的情况。

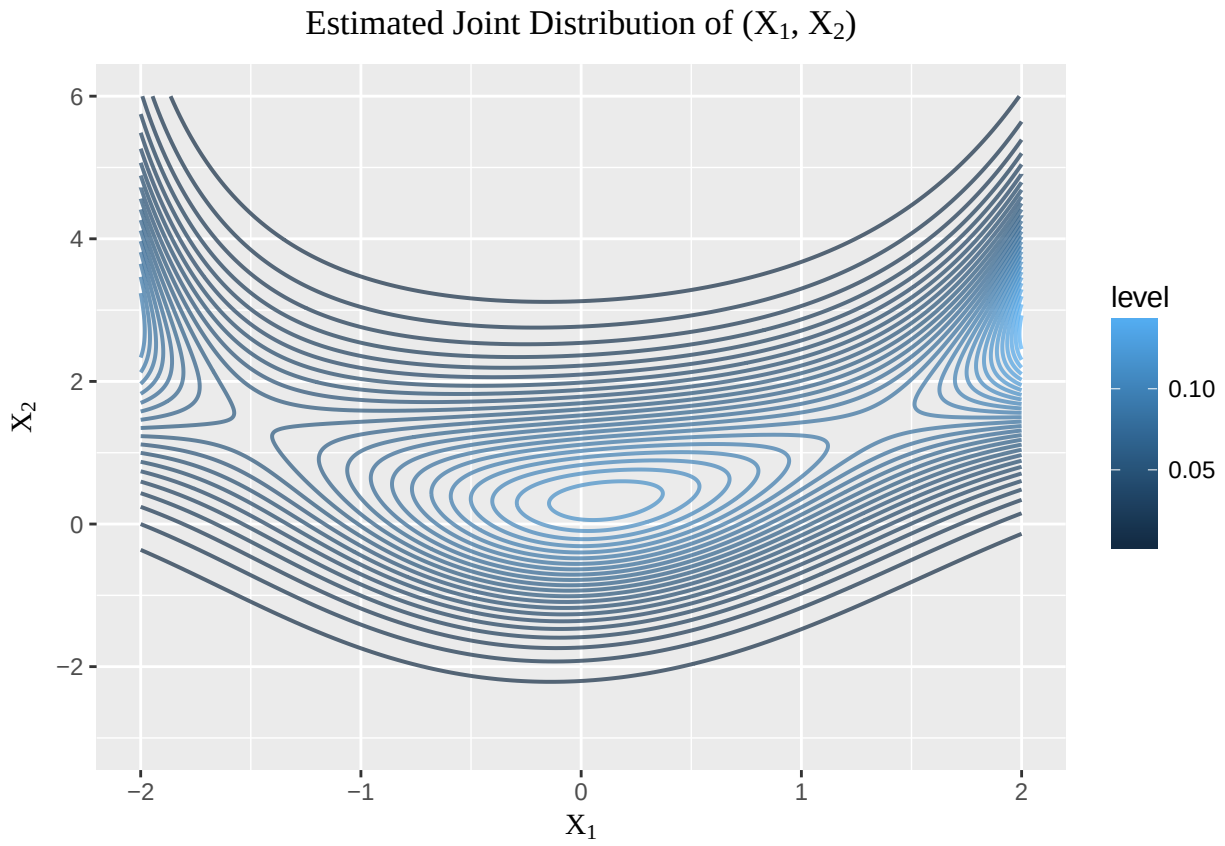
```
# 构造多项式
power.series <- function(x, y) {
  cbind(1, x, y, x^2, y^2, x*y, x^3, y^3, x^2*y, x*y^2)
}

# 定义 NLS 估计目标函数
Logit.nls <- function(theta, data) {
  k <- ncol(data) - 1
  x <- data[, 1:k]; y <- data[, k+1]
  phi <- x %*% theta
  P <- exp(phi)/(1 + exp(phi))
  mean((y - P)^2)
}

# 估计多项式系数
M <- 2000
Y <- rbind(array(1:1, dim = c(N, 1)), array(0:0, dim = c(M, 1)))
Z1 <- runif(M, -2, 2); Z2 <- runif(M, -3, 6)
XZ1 <- c(X1, Z1); XZ2 <- c(X2, Z2)

Data <- cbind(power.series(XZ1, XZ2), Y)
theta0 <- array(0:0, dim = c(ncol(Data)-1, 1))
Logit <- optim(par = theta0, data = Data, fn = Logit.nls)
theta <- Logit$par
```

最后，我们就可以使用估计出的 $\widehat{\theta}$ 来计算 $\widehat{f}_X(x_1,x_2)$。

```
X1grid <- seq(-2, 2, by = 0.01); X2grid <- seq(-3, 6, by = 0.01)
Xgrid <- expand.grid(X1grid, X2grid)

# 估计联合分布密度
fx <- function(x, theta) {
  x1 <- x[, 1]; x2 <- x[, 2]
  phi <- power.series(x1, x2) %*% theta
  dunif(x1, -2, 2)*dunif(x2, -3, 6)*exp(phi)
}
density <- fx(Xgrid, theta)

# 联合分布密度估计的等高线图
fx.contour <- ggplot(data.frame(Xgrid, density),
                     aes(Var1, Var2, z = density))
fx.contour + geom_contour(binwidth = 0.005, alpha = 0.7, size = 0.7,
                          aes(colour = ..level..)) +
  labs(title = TeX("Estimated Joint Distribution of ($X_1$, $X_2$)")) +
```

[7] 一般而言，使用 tensor product of splines 会有更好的近似效果（例如，使用 cpr 程序包中的 btensor 函数）。

```
  labs(y = TeX("$X_2$"), x = TeX("$X_1$")) +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                  face = "bold")) +
  xlim(c(-2, 2)) + ylim(c(-3, 6))
```



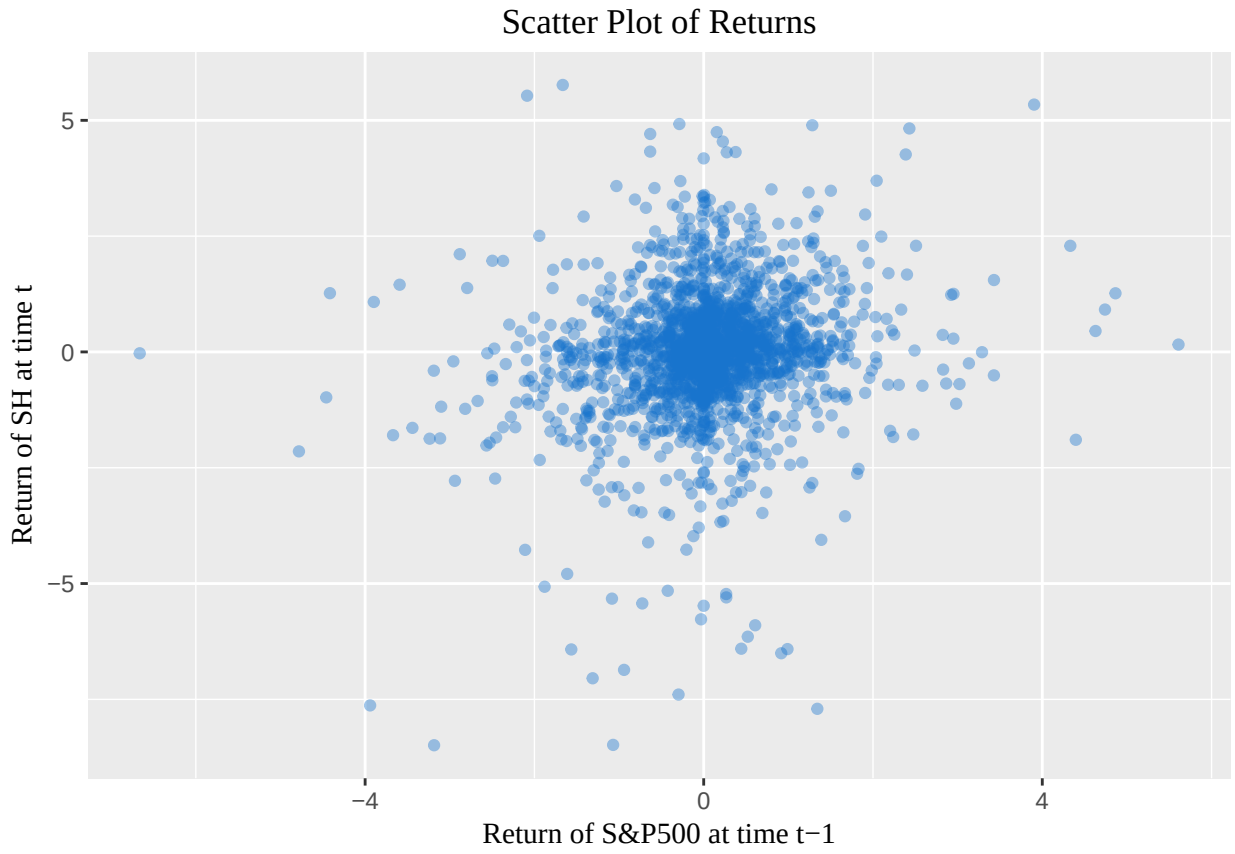
可以看出，上一节中所介绍的方法能够比较好的估计 X_1 与 X_2 的联合分布密度函数。当然，也应注意到，在 X_1 取极端值（如 X_1 位于两个标准差以外）时，由于样本量稀少，所以在这个区域内的估计并不可靠。

3. 应用实例：分析 S&P500 指数与上证综指收益率之间的联动性

下面，我们用第一部分介绍的方法研究 S&P500 指数与上证综指收益率之间的联动性，特别是判断 S&P500 指数是否为上证综指的先行指标。首先，我们通过 Wind 数据终端下载所需数据。这里我们选取 S&P500 指数与上证综指在 2010-1-1 到 2017-12-31 期间的日收盘价，并计算其日收益率（×100）。我们希望研究是否上证综指收益率与滞后一期（交易日）的 S&P500 指数收益率存在某种相关关系。

```
library(WindR)
library(dplyr)
library(readr)

# 下载并整理数据
w.start()
w_wsd_data <- w.wsd("SP500.SPI,399100.SZ,000001.SH","close",
                    "2010-01-01","2017-12-31")
Data <- w_wsd_data$Data
```

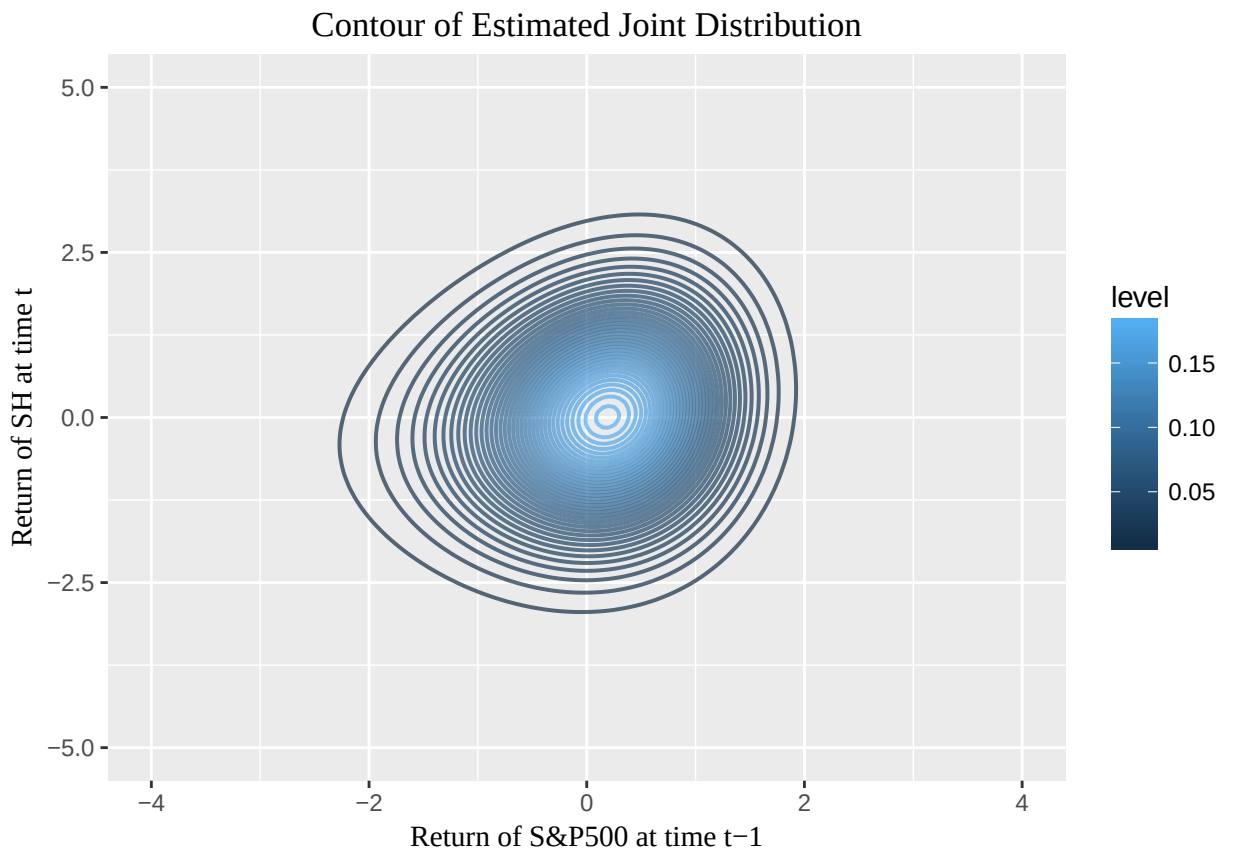
```r
indices <- Data %>%
  rename(sp500 = `SP500.SPI`, sh = `000001.SH`, sz = `399100.SZ`,
         date = `DATETIME`) %>%
  mutate(lsp500 = lag(sp500, n = 1L), Rsp500 = (sp500-lsp500)/lsp500*100,
         lsh = lag(sh, n = 1L), Rsh = (sh-lsh)/lsh*100,
         lsz = lag(sz, n = 1L), Rsz = (sz-lsz)/lsz*100,
         lRsp500 = lag(Rsp500, n = 1L)) %>%
  filter(is.na(lRsp500*Rsh*Rsz*Rsp500) != 1) %>%
  select(lRsp500, Rsz, Rsh)

# 散点图
scatter <- ggplot(indices, aes(lRsp500, Rsh))
scatter + geom_point(alpha = .4, size = 1.5, color = "dodgerblue3") +
  labs(title = TeX("Scatter Plot of Returns"),
       y = TeX("Return of SH at time $t$"),
       x = TeX("Return of S&P500 at time $t-1$")) +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                  face = "bold"))
```



从上面的散点图可以看出，上证综指收益率似乎和滞后一期的 S&P500 指数收益率存在着比较弱的正相关关系。下面我们用第一部分介绍的方法，估计两者的联合分布。

```r
# 估计联合分布密度
N <- nrow(indices); M <- N
Y <- rbind(array(1:1, dim = c(N, 1)), array(0:0, dim = c(M, 1)))
```

```r
set.seed(1)
Z1 <- rnorm(M, 0, 1); Z2 <- rnorm(M, 0, 1)
XZ1 <- c(indices$lRsp500, Z1); XZ2 <- c(indices$Rsh, Z2)

Logit.data <- cbind(power.series(XZ1, XZ2), Y)
theta0 <- array(0:0, dim = c(ncol(Logit.data) - 1, 1))

Logit <- optim(par = theta0, data = Logit.data, fn = Logit.nls)
theta <- Logit$par

X1grid <- seq(-5, 5, by = 0.05); X2grid <- seq(-6, 6, by = 0.05)
Xgrid <- expand.grid(X1grid, X2grid)

fx <- function(x, theta) {
  x1 <- x[, 1]; x2 <- x[, 2]
  phi <- power.series(x1, x2) %*% theta
  dnorm(x1, 0, 1)*dnorm(x2, 0, 1)*exp(phi)
}

# 联合分布密度估计的等高线图
density <- fx(Xgrid, theta)
fx.contour <- ggplot(data.frame(Xgrid, density), aes(Var1, Var2,
                                                     z = density))
fx.contour + geom_contour(binwidth = 0.005, alpha = 0.7, size = 0.7,
                          aes(colour = ..level..)) +
  labs(title = TeX("Contour of Estimated Joint Distribution")) +
  labs(y = TeX("Return of SH at time $t$"),
       x = TeX("Return of S&P500 at time $t-1$")) +
  xlim(c(-4, 4)) + ylim(c(-5, 5)) +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                  face = "bold"))
```



由估计出来的联合概率密度函数的等高线图可以看出，S&P500 指数收益率和上证综指收益率确实存正相关关系。但是这种正相关关系并非完全线性，从图中可以推断，两者在高收益（$>1\%$）与低收益（$<-1\%$）区间的相关系数似乎不同，前者相关度较高，后者相关度较低。