Estatística Computacional

Patrícia de Siqueira Ramos

PPGEAB
UNIFAL-MG

13 de Junho de 2018

Variáveis aleatórias multidimensionais

- A maioria dos fenômenos são multivariados
- As várias variáveis são, geralmente, correlacionadas entre si
  - mudanças em uma ou algumas delas afetam as demais
- Processos para gerar v.a.s multivariadas são considerados difíceis por alguns
- Boa parte deles pode ser obtida com apenas uma linha de comando no `R`
- Veremos detalhes de alguns processos para gerar dados dos principais modelos probabilísticos multivariados (normal multivariada, Wishart, Wishart invertida, t)

Normal multivariada

- Generalização da normal univariada quando há duas ou mais v.a.s simultaneamente
- Para um vetor aleatório com p variáveis, $\mathbf{X^T} = [X_1 \dots X_p]$, a densidade de $\mathbf{X}$ é dada por:

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\},$$

em que $-\infty < X_i < \infty$, $i = 1, \dots, p$
e ainda

Normal multivariada

$\boldsymbol{\mu} \in \mathbb{R}^p = [\mu_1 \dots \mu_p]$ (vetor de médias das variáveis)

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

Normal multivariada

$\mathbf{\Sigma}$: matriz de covariâncias simétrica positiva definida $p \times p$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}.$$

- $\mathbf{\Sigma}^{-1}$ é a inversa de $\mathbf{\Sigma}$ e $|\mathbf{\Sigma}|$ é seu determinante
- Notação: $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Teorema sobre a normal multivariada

Considere o vetor aleatório normal multivariado $\mathbf{X} = [X_1, \ldots, X_p]^T$ *com média* $\boldsymbol{\mu}$ *e covariância* $\boldsymbol{\Sigma}$ *e considere uma matriz* $\mathbf{C}$ *de dimensão* $p \times p$ *e posto* p. *Então a combinação linear* $\mathbf{Y} = \mathbf{CX}$ *de dimensão* $p \times 1$ *tem distribuição normal multivariada com média* $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{C}\boldsymbol{\mu}$ *e covariância* $\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T$.

Então, combinações lineares de variáveis normais multivariadas são normais multivariadas.

Normal multivariada

- O teorema nos fornece o principal resultado para gerar dados de uma normal multivariada
- Para aplicar o método, devemos obter a matriz $\mathbf{\Sigma}^{1/2}$ (matriz raiz quadrada de $\mathbf{\Sigma}$)
  - Para isso, considerar a decomposição espectral da matriz dada por $\mathbf{\Sigma} = \mathbf{P\Lambda P}^T$.
  - Assim, podemos definir a matriz $\mathbf{\Sigma}^{1/2} = \mathbf{P\Lambda^{1/2}P}^T$,
  em que $\mathbf{\Lambda}$ é a matriz diagonal dos autovalores, $\mathbf{\Lambda}^{1/2}$ é a matriz diagonal contendo a raiz quadrada desses elementos e $\mathbf{P}$ é a matriz de autovetores associados aos autovalores
  - Verifica-se que $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}$

Normal multivariada

O método para gerar uma realização p-variada de uma normal multivariada será:

1. Gerar um vetor aleatório $\mathbf{Z} = [Z_1, \ldots, Z_p]^T$ de variáveis normais padrão independentes (usando, por exemplo, o algoritmo de Box-Müller)
   - o que significa que $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbf{0}$ e $Cov(\mathbf{Z}) = \mathbf{I}$
2. Tal vetor deve sofrer a seguinte transformação linear:

$$\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$$

3. $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Normal multivariada

- Para obter a matriz raiz quadrada no `R`, podemos usar a decomposição espectral (em autovalores e autovetores), a decomposição do valor singular `svd` ou o comando `chol`, que retorna o fator de Cholesky de uma matriz positiva definida, o que é um tipo de raiz quadrada
- Para gerar as variáveis normais padrão usaremos a função `rnorm`
- Como exemplo, vamos gerar variáveis bivariadas ($p = 2$) com vetor de médias $\boldsymbol{\mu} = [10, 50]^T$ e matriz de covariâncias

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}.$$

Função rnormmv1

```
rnormmv1 = function(n, mu, Sigma){
  p = nrow(Sigma)
  ev = eigen(Sigma, symmetric = TRUE)
  lambda = ev$values
  P = ev$vectors
  Sigmaroot = P %*% diag(sqrt(lambda)) %*% t(P)
  Z = rnorm(p,0,1)
  X = t(Sigmaroot %*% Z + mu)
  if(n > 1){
    for(ii in 2:n){
      Z = rnorm(p,0,1)
      Y = Sigmaroot %*% Z + mu
      X = rbind(X, t(Y))
    }   # for
  }  # if
X  # matriz nxp dos dados
}  # rnormmv1
# exemplo de utilização
Sigma = matrix(c(4,1,1,1), 2, 2)
mu = c(10, 50)
n = 2500
X = rnormmv1(n, mu, Sigma)
X

Xb = apply(X, 2, mean)  # aplica a função mean às colunas de X
Xb
var(X)
plot(X, xlab = 'X1', ylab = 'X2', pch = 20)
```

Função rnormmv2 - versão otimizada (sem *loops*)

```
rnormmv2 = function(n,mu,Sigma){
  p = nrow(Sigma)
  ev = eigen(Sigma, symmetric = TRUE)
  lambda = ev$values
  P = ev$vectors
  Sigmaroot = P %*% diag(sqrt(lambda)) %*% t(P)
  X = (matrix(rnorm(n * p), n, p) %*% Sigmaroot) +
        matrix(rep(mu, each = n), n, p)
  X # matriz n x p dos dados
} # rnormmv2

# exemplo de uso
(Sigma = matrix(c(4, 1.9, 1.9, 1), 2, 2))
(mu = c(10, 50))
n = 2500
(X = rnormmv2(n, mu, Sigma))
(Xb = apply(X, 2, mean))
(S = var(X))
plot(X, xlab = 'X1', ylab = 'X2', pch = 20)
```

Comparação dos tempos: `rnormmv` × `rnormmultv`

```
# comparação dos tempos
n = 15000
# função com loops
tg1 = system.time(rnormmv1(n, mu, Sigma))
tg1
# função sem loops
tg2 = system.time(rnormmv2(n, mu, Sigma))
tg2
# razão entre tempos gastos pelas duas funções
tg1 / tg2
```

Usando funções prontas do R

- Função `mvrnorm(n, mu, Sigma)` do pacote `MASS`
- Função `rmvnorm(n, mean, sigma)` do pacote `mvtnorm`
- Pacote `MASS`: funções disponibilizadas em *Modern Applied Statistics with S*

```
Sigma = matrix(c(4, 1.9, 1.9, 1), 2, 2)
mu = c(10, 50); n = 30
library(MASS)
X = mvrnorm(n, mu, Sigma)
Xb = apply(X, 2, mean)
Xb # média de X
var(X) # covariância de X
library(mvtnorm)
Y = rmvnorm(n, mu, Sigma)
Yb = apply(Y, 2, mean)
Yb # média de Y
var(Y) # covariância de Y
```

Comparação dos tempos das funções vistas

```
library(MASS)
library(mvtnorm)

n = 30000   # tamanho amostral
p = 5       # variáveis
(mu = numeric(p))   # vetor de médias nulo
(Sigma = cov(matrix(rnorm(n*p), n, p)))

# tempo rnormmv1
set.seed(100)
system.time(rnormmv1(n, mu, Sigma))
# tempo rnormmv2
set.seed(100)
system.time(rnormmv2(n, mu, Sigma))
# tempo mvrnorm - pacote MASS
set.seed(100)
system.time(mvrnorm(n, mu, Sigma))
# tempo rmvnorm - pacote mvtnorm
set.seed(100)
system.time(rmvnorm(n, mu, Sigma))
```

Distribuição Wishart e Wishart invertida

- Distribuições relacionadas às matrizes de somas de quadrados e produtos $\mathbf{W}$ obtidas de amostras de tamanho ν da normal multivariada
- São extensões da distribuição χ^2
- A ideia é obter a matriz

$$\mathbf{W} = (n-1)\mathbf{\Sigma}$$

Distribuição Wishart

- Considere $\mathbf{X}_j = [X_1, \ldots, X_p]^T$ o j-ésimo vetor ($j = 1, \ldots, n$) de uma amostra aleatória de tamanho n de uma $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então a matriz aleatória

$$\mathbf{W} = \sum_{j=1}^{n} (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^T \sim \text{Wishart}(\nu = n - 1, \boldsymbol{\Sigma})$$

Distribuição Wishart

- Se considerarmos uma amostra aleatória de tamanho ν de uma $N_p(\mathbf{0}, \mathbf{\Sigma})$, então a matriz aleatória

$$\mathbf{W} = \sum_{j=1}^{\nu} \mathbf{X}_j \mathbf{X}_j^T \sim \text{Wishart}(\nu, \mathbf{\Sigma})$$

Geração de v.a.s Wishart

Para gerarmos variáveis Wishart($\nu, \mathbf{\Sigma}$) devemos:

1. Gerar variáveis normais multivariadas
2. Obter matriz de covariâncias amostrais dos dados gerados ($\mathbf{S}$)
3. Obter a matriz de somas de quadrados e produtos amostrais

$$\mathbf{W} = (n-1)\mathbf{S}$$

4. Tal matriz ($p \times p$) será uma realização da v.a. Wishart

Função rWishart

```
# Exemplificação da geração de matrizes de somas de quadrados e produtos
# aleatórias W com distribuição Wishart(nu, Sigma)
# utiliza o pacote mvtnorm para gerar amostras normais
rWishart = function(nu, Sigma){
  p = nrow(Sigma)
  mu = matrix(0,p)
  y = rmvnorm(nu + 1, mu, Sigma)
  w = nu * var(y) # var(y) retorna a matriz de covariâncias
  w
}
# exemplo de uso
library(mvtnorm)
Sigma = matrix(c(4,1,1,1),2,2)
nu = 5
w = rWishart(nu,Sigma)
w
```

Distribuição Wishart invertida

- Considere $\mathbf{W}$ uma matriz aleatória $W_p(\nu, \mathbf{\Sigma})$, então a distribuição de $\mathbf{S} = \mathbf{W}^{-1}$ é a Wishart invertida, representada por $W_p^{-1}(\nu, \mathbf{\Sigma})$
- Para gerar uma matriz aleatória de W_p^{-1}, precisamos apenas realizar a transformação

$$\mathbf{S} = \mathbf{W}^{-1}$$

Função rWishart modificada

```
# Exemplo de geração de matrizes de somas de quadrados e produtos
# aleatórias W com distribuição Wishart(nu, Sigma) e Wishart invertida
# WI(nu, Sigma)
# Utiliza o pacote mvtnorm para gerar amostras normais
rWishart = function(nu, Sigma){
  p = nrow(Sigma)
  mu = matrix(0,p)
  y = rmvnorm(nu + 1, mu, Sigma)
  w = nu*var(y)
  wi = solve(w) # solve retorna a inversa
  list(w = w, wi = wi)
}

# exemplo de uso
library(mvtnorm)
Sigma = matrix(c(4,1,1,1),2,2)
nu = 5
W = rWishart(nu,Sigma)
W$w    # Wishart
W$wi   # Wishart invertida
```

Distribuição t de Student multivariada

- Distribuição útil em inferência multivariada, para comparação de vetores de médias
- Para um vetor aleatório com p variáveis, $\mathbf{X^T} = [X_1 \dots X_p]$, com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, a f.d.p. da t multivariada é

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{X}) &= \frac{g(\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})}{|\boldsymbol{\Sigma}|^{1/2}} \\ &= \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}}\left[1+\frac{1}{\nu}(\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}}, \end{aligned}$$

- em que $g(z) = \dfrac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{p/2}\Gamma(\nu/2)}\left(1+\frac{z}{\nu}\right)^{-\frac{\nu+p}{2}}$
- $\mathbf{X}$ tem média $\boldsymbol{\mu}$ e matriz de covariâncias $\nu\boldsymbol{\Sigma}/(\nu-2)$ para $\nu > 2$.

Geração de v.a.s p-dimensionais da t multivariada

- Seja um vetor aleatório $\mathbf{Z}$ com distribuição $N_p(\mathbf{0}, \mathbf{I})$ e a variável aleatória Q com distribuição χ^2 com ν g.l., então o vetor aleatório $\mathbf{Y}$, dado pela transformação

$$\mathbf{Y} = \sqrt{(\nu)}\frac{\mathbf{Z}}{\sqrt{Q}}$$

possui distribuição t multivariada esférica com ν g.l.

- Já o vetor $\mathbf{X}$ obtido pela transformação linear

$$\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Y} + \boldsymbol{\mu} \tag{1}$$

possui distribuição t multivariada elíptica com ν g.l. e parâmetros $\boldsymbol{\mu}$ e $\mathbf{\Sigma}$.

Geração de v.a.s p-dimensionais da t multivariada

- Assim, devemos aplicar a transformação (1) n vezes a n diferentes vetores aleatórios $\mathbf{Y}$ e variáveis Q
- Ao final, teremos uma amostra n da t multivariada
- A função `rtmult` é apresentada para este fim, usando o fator de Cholesky para obter a matriz $\mathbf{\Sigma}^{1/2}$

Função rtmult

```
# Geração de variáveis aleatórias t multivariadas (n, mu, Sigma, nu).
# Devemos carregar o pacote mvtnorm antes de usar a função
rtmult = function (n, mu=c(0, 0, 0), sigma = diag(3), df = 1){
  library(mvtnorm)
  FT = chol(sigma)
  p = nrow(sigma)
  x = (rmvnorm(n, sigma = diag(p))/sqrt(rchisq(n, df)/df))%*%FT +
    matrix(rep(mu, each = n), n, p)
  x
}
# Exemplo de uso
nu = 3
par(mfrow = c(1,2))
plot(rtmult(2000, c(0,0), diag(2), nu), xlab = 'X1', ylab = 'X2', pch = 20)
Sigma = matrix(c(4, 1.9, 1.9, 1), 2, 2)
mu = c(10, 5)
x = rtmult(3000, mu, Sigma, nu) # t com correlaçao 0,95 entre X1 e X2
var(x) # valor esperado é nu Sigma/(nu - 2)
nu*Sigma/(nu - 2) # valor esperado
plot(x, xlab = 'X1', ylab = 'X2', pch = 20)
```