詳細解答

STEP III 2010，第 13 題

概率與統計 • 線性組合的相關性

STEP III 2010，第 13 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

step.maths.org/questions/10-s3-q13

本題在 *Beyond STEP* 第 13.7 節製造相關性中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch13/correlation.html

關鍵思路

這裡的一切都是平均值為 0、方差為 1 的獨立（independent）變量的線性組合（linear combination），所以每個期望值（expectation）、方差（variance）和協方差（covariance）都可以直接從係數讀出：若 $U = \sum u_k Z_k$ 及 $V = \sum v_k Z_k$，則 $\operatorname{Cov}(U, V) = \sum u_k v_k$。這道題其實是在逐個變量地構造三個具有任意指定相關係數（correlation）的標準化（standardised）變量，本質上就是喬裝了的格拉姆–施密特（Gram–Schmidt）正交化過程。每個新條件都確定一個新的係數。

解答

全程我們使用協方差的雙線性（bilinearity）、$\operatorname{Var}(Z_k) = 1$，以及當 $j \neq k$ 時 $\operatorname{Cov}(Z_j, Z_k) = 0$（這是因為各個 Z_k 互相獨立）。特別地，對常數 u_k、v_k，

$$\operatorname{Cov}\Big(\sum_k u_k Z_k,\ \sum_k v_k Z_k\Big) = \sum_{j,k} u_j v_k \operatorname{Cov}(Z_j, Z_k) = \sum_k u_k v_k.$$

Z_1 與 Z_2 的相關係數

由獨立性得 $\operatorname{Cov}(Z_1, Z_2) = \mathrm{E}(Z_1 Z_2) - \mathrm{E}(Z_1)\mathrm{E}(Z_2) = 0$，所以 $\operatorname{Corr}(Z_1, Z_2) = 0$。

變量 Y_2

記 $s = \sqrt{1-\rho_{12}^2}$，由於 $-1 < \rho_{12} < 1$，它是正實數。於是 $Y_2 = \rho_{12}Z_1 + sZ_2$，由期望值的線性（linearity of expectation）得

$$\mathrm{E}(Y_2) = \rho_{12}\mathrm{E}(Z_1) + s\,\mathrm{E}(Z_2) = 0.$$

由上述協方差法則，

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + s^2 = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + sZ_2) = \rho_{12}.$$

由於 $\mathrm{Var}(Y_1) = \mathrm{Var}(Y_2) = 1$，定義中的分母為 1，因此

$$\mathrm{Corr}(Y_1, Y_2) = \rho_{12}.$$

變量 Y_3

條件 $\mathrm{E}(Y_3) = 0$ 對任何 a, b, c 都自動成立，所以它沒有提供任何資訊。其餘三個條件各給出一條方程。由於 Y_1, Y_2, Y_3 的方差都是 1，每個相關係數都等於對應的協方差。

與 Y_1 的相關係數。 $\mathrm{Cov}(Z_1,\, aZ_1 + bZ_2 + cZ_3) = a$，所以

$$a = \rho_{13}.$$

與 Y_2 的相關係數。 $\mathrm{Cov}(\rho_{12}Z_1 + sZ_2,\, aZ_1 + bZ_2 + cZ_3) = \rho_{12}a + sb$，所以 $\rho_{12}\rho_{13} + sb = \rho_{23}$，又因 $s \neq 0$，

$$b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}.$$

單位方差。 $a^2 + b^2 + c^2 = 1$，由於 $c \geq 0$，我們取非負的平方根：

$$\begin{aligned} c^2 &= 1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2} \\ &= \frac{(1-\rho_{12}^2)(1-\rho_{13}^2) - (\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2} \\ &= \frac{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}, \end{aligned}$$

其中最後一步裡 $\rho_{12}^2\rho_{13}^2$ 項互相抵消。因此

$$c = \sqrt{\frac{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}}.$$

想一想 如果平方根下的式子是負數，會怎樣？

那麼不存在實數 c，而且根本不存在具有這三個相關係數的隨機變量：$\rho_{12}, \rho_{13}, \rho_{23}$ 這三個數彼此並不相容。例如，$\rho_{12} = \rho_{13} = 0.9$ 而 $\rho_{23} = -0.9$ 是不可能的，因為如果 Y_2 和 Y_3 都與 Y_1 十分接近，它們之間就不可能強烈地反向相關。題目默認所給的相關係數是可以實現的，而這正正就是該分子非負的條件。

最後部分：X_i

對 $i = 1, 2, 3$，取

$$X_i = \mu_i + \sigma_i Y_i,$$

其中 $\sigma_i > 0$（我們可以取 σ_i 為正，因為題目只指定了 σ_i^2）。於是 $\mathrm{E}(X_i) = \mu_i + \sigma_i \cdot 0 = \mu_i$，且 $\mathrm{Var}(X_i) = \sigma_i^2 \mathrm{Var}(Y_i) = \sigma_i^2$。加上常數不會改變協方差，所以

$$\mathrm{Corr}(X_i, X_j) = \frac{\sigma_i \sigma_j \,\mathrm{Cov}(Y_i, Y_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} = \frac{\sigma_i \sigma_j}{|\sigma_i \sigma_j|}\,\rho_{ij} = \rho_{ij},$$

這裡用了 $\sigma_i \sigma_j > 0$。所以這些 X_i 具備所有要求的性質。

驗算

- 若 $\rho_{13} = \rho_{23} = 0$，則得 $a = b = 0$、$c = 1$，所以 $Y_3 = Z_3$，與其他變量獨立，理應如此。若 $\rho_{12} = 0$，則 $Y_2 = Z_2$，且 $a = \rho_{13}$、$b = \rho_{23}$、$c = \sqrt{1 - \rho_{13}^2 - \rho_{23}^2}$，正是顯而易見的選擇。
- 數值上，取 $\rho_{12} = 0.5$、$\rho_{13} = -0.3$、$\rho_{23} = 0.2$，得 $a = -0.3$、$b \approx 0.4041$、$c \approx 0.8641$，且 $a^2 + b^2 + c^2 = 1$。以 2×10^5 組三元組進行模擬（所用的 Z_k 是均勻分佈而非正態分佈），得到 X_i 的樣本相關係數為 0.500、-0.302、0.200，平均值和標準差亦正確。

常見錯誤

- 忘記 $\mathrm{Var}(Y_2) = 1$ 意味着相關係數就是協方差，結果除錯了東西。
- 寫成 $X_i = \mu_i + \sigma_i^2 Y_i$：方差按乘數的平方縮放。
- 對其中一個變量取 $\sigma_i < 0$，這會令它的相關係數變號。

STEP 之外

這三個條件表示係數矩陣 A（各行為 $(1,0,0)$、$(\rho_{12},s,0)$、(a,b,c)）滿足 $AA^{\mathsf{T}}=R$，其中 R 是相關矩陣（correlation matrix）。這就是 R 的喬列斯基分解（Cholesky decomposition），它存在若且唯若 R 是半正定（positive semi-definite）的；而 c^2 的分子正是 $\det R$。統計學家在實際中正是用這個方法模擬相關的變量；若 Z_k 是正態分佈的，所得的就是多元正態分佈（multivariate normal distribution）。Blitzstein 與 Hwang 的 *Introduction to Probability*（《概率導論》）涵蓋協方差、相關係數和多元正態分佈；Strang 的 *Introduction to Linear Algebra*（《線性代數導論》）講解正定矩陣（positive definite matrix）和喬列斯基分解。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。