详细解答

STEP III 2010，第 13 题

概率与统计 · 线性组合的相关性

STEP III 2010，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

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

本题在 *Beyond STEP* 第 13.7 节制造相关性中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/correlation.html

关键思路

这里的一切都是平均值为 0、方差为 1 的独立（independent）变量的线性组合（linear combination），所以每个期望值（expectation）、方差（variance）和协方差（covariance）都可以直接从系数读出：若 $U = \sum u_k Z_k$ 及 $V = \sum v_k Z_k$，则 $\operatorname{Cov}(U, V) = \sum u_k v_k$。这道题其实是在逐个变量地构造三个具有任意指定相关系数（correlation）的标准化（standardised）变量，本质上就是乔装了的格拉姆–施密特（Gram–Schmidt）正交化过程。每个新条件都确定一个新的系数。

解答

全程我们使用协方差的双线性（bilinearity）、$\operatorname{Var}(Z_k) = 1$，以及当 $j \neq k$ 时 $\operatorname{Cov}(Z_j, Z_k) = 0$（这是因为各个 Z_k 互相独立）。特别地，对常数 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 与 Z_2 的相关系数

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

变量 Y_2

记 $s=\sqrt{1-\rho_{12}^2}$，由于 $-1<\rho_{12}<1$，它是正实数。于是 $Y_2=\rho_{12}Z_1+sZ_2$，由期望值的线性（linearity of expectation）得

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

由上述协方差法则，

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

由于 $\mathrm{Var}(Y_1)=\mathrm{Var}(Y_2)=1$，定义中的分母为 1，因此

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

变量 Y_3

条件 $\mathrm{E}(Y_3)=0$ 对任何 a,b,c 都自动成立，所以它没有提供任何信息。其余三个条件各给出一条方程。由于 Y_1,Y_2,Y_3 的方差都是 1，每个相关系数都等于对应的协方差。

与 Y_1 的相关系数。 $\mathrm{Cov}(Z_1,\,aZ_1+bZ_2+cZ_3)=a$，所以

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

与 Y_2 的相关系数。 $\mathrm{Cov}(\rho_{12}Z_1+sZ_2,\,aZ_1+bZ_2+cZ_3)=\rho_{12}a+sb$，所以 $\rho_{12}\rho_{13}+sb=\rho_{23}$，又因 $s\neq 0$，

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

单位方差。 $a^2+b^2+c^2=1$，由于 $c\geq 0$，我们取非负的平方根：

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

其中最后一步里 $\rho_{12}^2\rho_{13}^2$ 项互相抵消。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负数，会怎样?

那么不存在实数 c，而且根本不存在具有这三个相关系数的随机变量：$\rho_{12},\rho_{13},\rho_{23}$ 这三个数彼此并不相容。例如，$\rho_{12}=\rho_{13}=0.9$ 而 $\rho_{23}=-0.9$ 是不可能的，因为如果 Y_2 和 Y_3 都与 Y_1 十分接近，它们之间就不可能强烈地反向相关。题目默认所给的相关系数是可以实现的，而这正正就是该分子非负的条件。

最后部分： X_i

对 $i=1,2,3$，取

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

其中 $\sigma_i>0$（我们可以取 σ_i 为正，因为题目只指定了 σ_i^2）。于是 $\mathrm{E}(X_i)=\mu_i+\sigma_i\cdot 0=\mu_i$，且 $\mathrm{Var}(X_i)=\sigma_i^2\mathrm{Var}(Y_i)=\sigma_i^2$。加上常数不会改变协方差，所以

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

这里用了 $\sigma_i\sigma_j>0$。所以这些 X_i 具备所有要求的性质。

验算

- 若 $\rho_{13}=\rho_{23}=0$，则得 $a=b=0$、$c=1$，所以 $Y_3=Z_3$，与其他变量独立，理应如此。若 $\rho_{12}=0$，则 $Y_2=Z_2$，且 $a=\rho_{13}$、$b=\rho_{23}$、$c=\sqrt{1-\rho_{13}^2-\rho_{23}^2}$，正是显而易见的选择。
- 数值上，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 组三元组进行模拟（所用的 Z_k 是均匀分布而非正态分布），得到 X_i 的样本相关系数为 0.500、-0.302、0.200，平均值和标准差亦正确。

常见错误

- 忘记 $\mathrm{Var}(Y_2)=1$ 意味着相关系数就是协方差，结果除错了东西。
- 写成 $X_i=\mu_i+\sigma_i^2Y_i$：方差按乘数的*平方*缩放。
- 对其中一个变量取 $\sigma_i<0$，这会令它的相关系数变号。

STEP 之外

这三个条件表示系数矩阵 A（各行为 $(1,0,0)$、$(\rho_{12},s,0)$、(a,b,c)）满足 $AA^{\mathsf{T}}=R$，其中 R 是相关矩阵（correlation matrix）。这就是 R 的*乔列斯基分解*（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 $\det R$。统计学家在实际中正是用这个方法模拟相关的变量；若 Z_k 是正态分布的，所得的就是多元正态分布（multivariate normal distribution）。Blitzstein 与 Hwang 的 *Introduction to Probability*（《概率导论》）涵盖协方差、相关系数和多元正态分布；Strang 的 *Introduction to Linear Algebra*（《线性代数导论》）讲解正定矩阵（positive definite matrix）和乔列斯基分解。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。