

多項目比率データのクラスター分析

岡山大 養

脇本 和昌

岡山理大

山本 英二

岡山大 養

垂水 共之

Random Distance の分布を用いるクラスター分析法の応用として多項目比率 $(P_1, \dots, P_{k+1})$ が与えられたときも、同様の議論ができる。この場合は、各比率がランダムであるとする $P = (P_1, \dots, P_{k+1})'$ が、 k 次元 Dirichlet 分布 $D(1, \dots, 1; 1)$ に従うので、 D_k^2 の代りに $Q_k^2 = \sum_{i=1}^{k+1} (P_i - P_i')^2$ を使えば Q_k^2 は

$$Q_k^2 = (P_1 - P_2)' \begin{pmatrix} 2 & 1 & \dots & 1 \\ & \ddots & & \\ & & 1 & \\ 1 & \dots & 1 & 2 \end{pmatrix} (P_1 - P_2)$$

の形をしている。ここで P_1, P_2 は互いに独立で、同一分布の k 次元 Dirichlet 分布 $D(1, \dots, 1; 1)$ に従う。

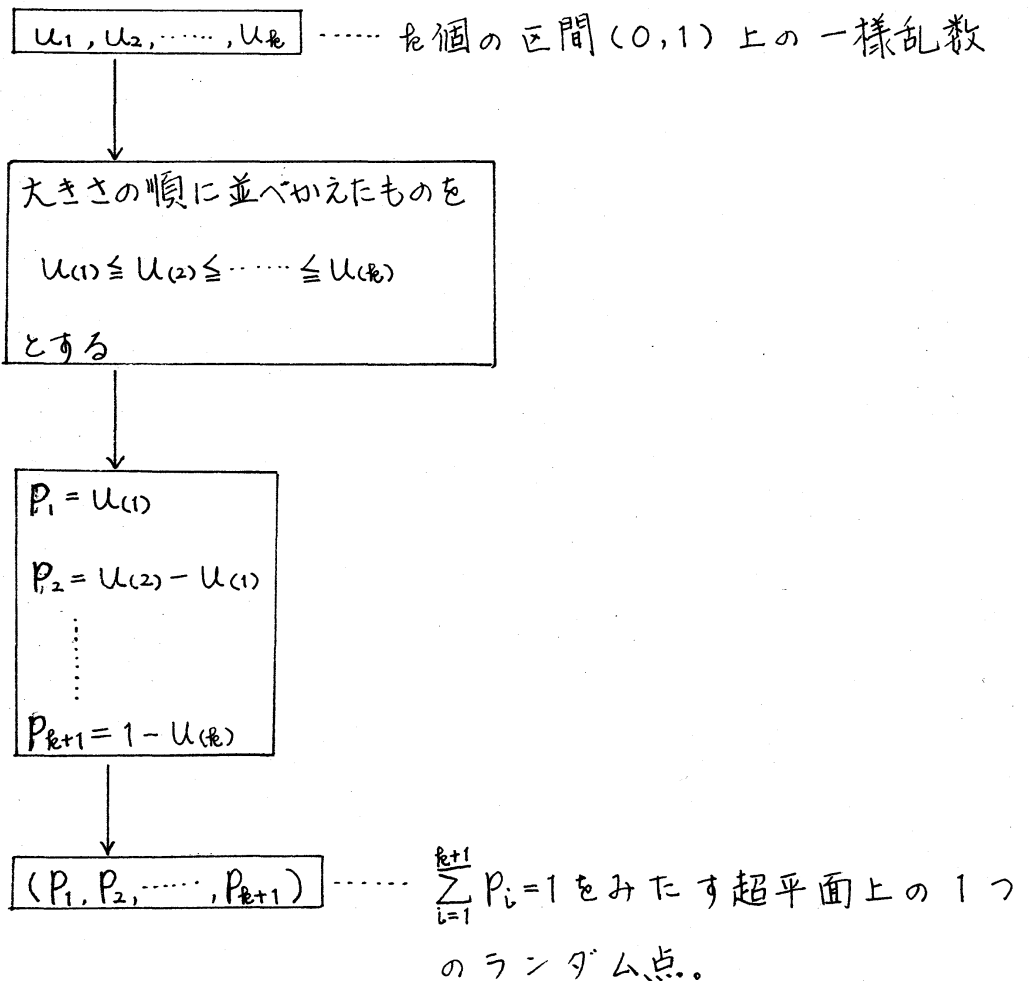
Q_k^2 の平均と分散は

$$E(Q_k^2) = \frac{2k}{(k+2)(k+1)}$$

$$V(Q_k^2) = \frac{4k(4k^3 + 31k^2 + 47k - 24)}{(k+4)(k+3)(k+2)^2(k+1)^2} - (E(Q_k^2))^2$$

となる。 Q_k^2 ($k \geq 2$) の確率分布の計算は大変であり、現在計算中であるが、まだその結果を得てないので、シミュレーションによって近似確率密度関数を求めてみた。その結果を図1に示す。この密度関数を用いれば、前報告と同様にクラスター分析をおこなうことができる。

[注] Q_k^2 の確率分布のシミュレーションによる求め方。



$\sum_{i=1}^{k+1} p_i = 1$ 上の2つのランダム点を作り、2点間の距離を計算して確率密度を求めた。

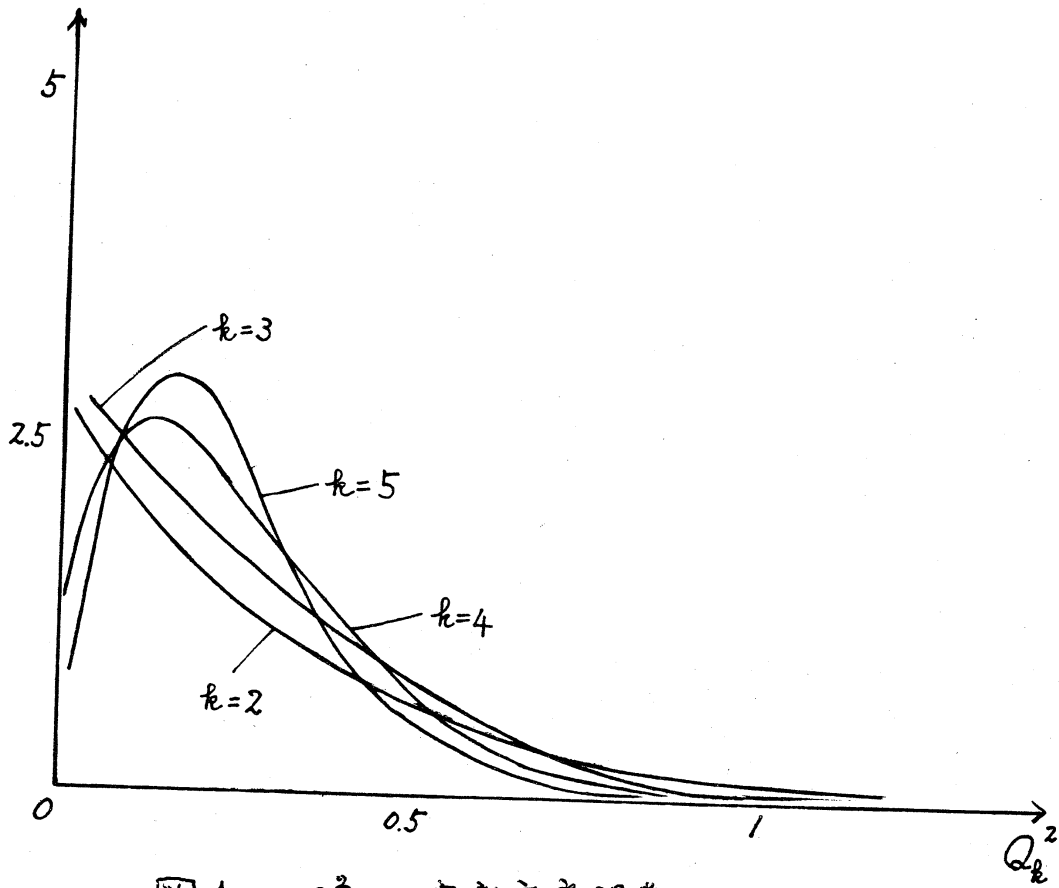


図 1. Q_k^2 の確率密度関数