

## Random Distance の分布を用いるクラスター分析

岡山大 養	脇本 和昌
岡山理大	山本 英二
岡山大 養	垂水 共之

### [1] Random Distance の分布

互いに独立で、 $(0,1)$  の一様分布に従う  $k$  個の要素で構成された  $k$  次元 random vector  $(X_1, X_2, \dots, X_k)$  を考える。二つの random vector  $(X_1, X_2, \dots, X_k), (X'_1, X'_2, \dots, X'_k)$  の距離

$D_k = \sqrt{\sum_{i=1}^k (X_i - X'_i)^2}$  を random distance と呼ぶことにする。

random distance の二乗  $D_k^2$  の  $k=1, 2, 3$  のとき分布関数

$F_k(d^2)$  は以下の通り、

$$F_1(d^2) = \begin{cases} 0 & d^2 \leq 0 \\ -d^2 + 2d & 0 < d^2 \leq 1 \\ 1 & 1 \leq d^2 \end{cases}$$

$$F_2(d^2) = \begin{cases} 0 & d^2 \leq 0 \\ \pi d^2 - \frac{8}{3}d^3 + \frac{1}{2}d^4 & 0 < d^2 \leq 1 \end{cases}$$

$$F_3(d^2) = \begin{cases} \frac{1}{3} + (\pi - 2)d^2 + 4(d^2 - 1)^{\frac{1}{2}} + \frac{8}{3}(d^2 - 1)^{\frac{3}{2}} - \frac{d^2}{2} - 4d^2 \sec^{-1} d & 1 < d^2 \leq 2 \\ 1 & 2 < d^2 \end{cases}$$

$$F_3(d^2) = \begin{cases} 0 & d^2 \leq 0 \\ \frac{4}{3}\pi d^3 - \frac{3}{2}\pi d^4 + \frac{8}{5}d^5 - \frac{1}{6}d^6 & 0 < d^2 \leq 1 \\ \left(\frac{5}{2}\pi + \frac{43}{30}\right) - 6(d^2 - 1)^{\frac{1}{2}} + (3\pi + \frac{7}{2})(d^2 - 1) \\ \quad - \frac{8}{3}\pi d^3 - 10(d^2 - 1)^{\frac{3}{2}} + \frac{5}{2}(d^2 - 1)^2 \\ \quad - \frac{16}{5}(d^2 - 1)^{\frac{5}{2}} + \frac{1}{3}(d^2 - 1)^3 + 6d^4 \sec^{-1} d & 1 < d^2 \leq 2 \\ \left(\frac{23}{2}\pi - \frac{343}{30}\right) + 14(d^2 - 2)^{\frac{1}{2}} + (9\pi - \frac{21}{2})(d^2 - 2) + 10(d^2 - 2)^{\frac{3}{2}} \\ \quad + \left(\frac{3\pi - 5}{2}\right)(d^2 - 2)^2 + \frac{8}{5}(d^2 - 2)^{\frac{5}{2}} - \frac{1}{6}(d^2 - 2)^3 \\ \quad - (6d^4 + 12d^2 - 2) \sec^{-1} \sqrt{d^2 - 1} + 8d^3 \sec^{-1}(d^2 - 1) - \frac{8}{3}\pi d^3 & 2 < d^2 \leq 3 \\ 1 & 3 < d^2 \end{cases}$$

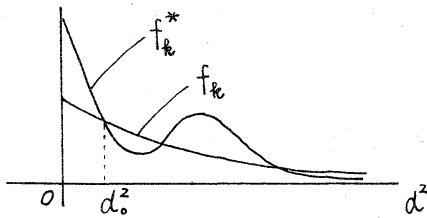
$n$  が大きいときの  $D_n^2$  の漸近分布は、正規分布であるが、原点の回りの  $r$  次のモーメントが  $\frac{1}{(2r+1)(r+1)}$  であることを利用して、Gram-Charlier-Edgeworth 近似によつて、近似の精度を上げることが出来る。

## [2] クラスタ分析のアルゴリズム

1.  $n$ 次元単位立方体上の  $N$ 個 ( $N \gg 1$ ) のデータが作る  $nC_2$

個の線分から復元抽出により大きさ  $n$  のユニタムサンプルを抽出し、それにもとずき距離の経験分布  $F_k^*$  を計算する。

2. [1] で求めた分布  $F_k$  との適合度検定を行い、有意水準  $\alpha\%$  で棄却されたとき、クラスターがあると判断し、さらに下図の様に  $d_0$  を定め、以下、クラスター ( $d_0$ ) を求めていく。



3. 単位立方体の中に一辺  $d'_0 = \frac{1}{\sqrt{k}} \cdot \frac{1}{\sqrt{M}} \cdot d_0$  ( $M = 1 + \frac{\log_{10} N}{\log_{10} 2}$  : ステージエスの式) の立方体を系統的に  $m$  個とり ( $S_1, \dots, S_m$ )、 $S_i$  に含まれる全ての標本点の個数を  $l_i$  とし、 $l_i$  個の点の密度と  $l_i C_2$  個の距離の経験分布を計算する。
4. 各  $S_i$  において、密度と経験分布を用いてクラスターがあるかないか判断し、ある場合は、重心を *seed point* とする。
5. クラスターを有する  $S_i$  の結合を *seed point* 間距離と  $d_0'' = \frac{1}{\sqrt{M}} d_0$  の比較を行う。

[3] この手法の特徴

1. クラスターの決定を自動的に行う。2点間距離の経験分布を使って、クラスターを定義し、それを求めていくことになる。
2. クラスターの判断を密度だけでなく、2点間距離の経験分布を用いて行っていること。これにより、 $S_i$ において、高密度でなくとも、点が偏在しているときは、クラスターと判断可能となる。
3. 大標本への適用可能  
 サンプリングと、立方体の分割により、計算量を大中に減らすことができる。

#### [4] 適用例

図1に示すような5つのクラスターと1つの離れた点からなる2変量のデータを考える。このデータに対してここで提案した方法を用いると、左上5つの *seed point* と右下4つの *seed point* ができ、その *seed point* の結合により2つのクラスターと1つの離れた点とに分けることができた。なお1辺  $d_0$  の正方形の内部に1点しかはいらない場合は、1番近い *seed point* までの距離が  $d_0$  内であれば、その点を *seed point* に属させてクラスターを作った。

图 1

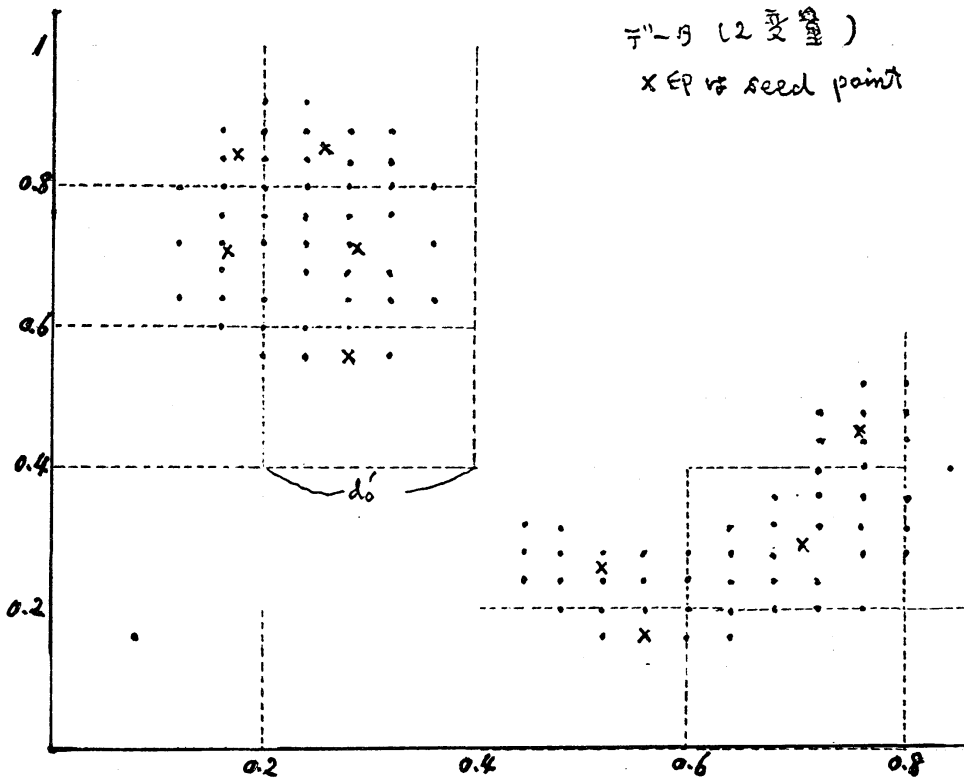
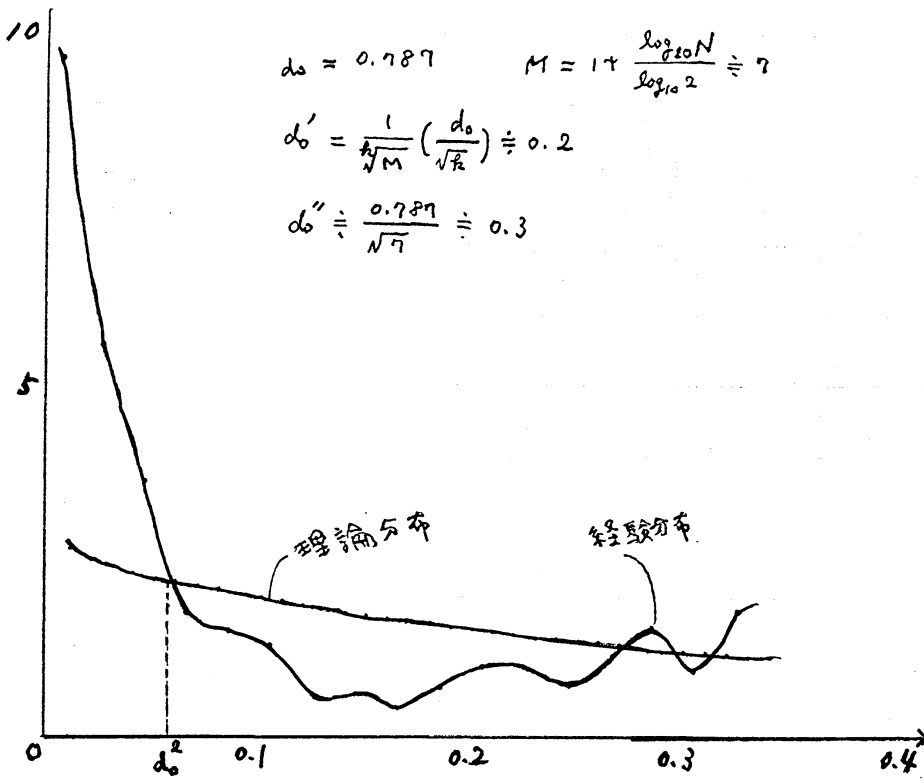


图 2



今後の研究課題としては、他のクラスター分析の手法との比較検討をおこなうこと、また複雑な例題についてその効果を例示することを考えている。

### 参考文献

- [1] F. Gruenberger & A. M. Mark (1951) "The  $d^2$ -test of random digits" Math. Tables other Aids Comp. 5. pp. 109-110.
- [2] A. A. TÖRN (1977) "Cluster analysis using seed points and density-determined hyperspheres as an aid to global optimization" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBNETICS, Vol. SMC-7, NO-8, AUGUST pp. 610-616.