卒業研究報告書

クラスタ数推定に用いる最適な情報量基準の探求

指導教員

藤田 一寿

津山工業高等専門学校

情報工学科

萩原 涼介

平成 29 年 2 月 13 日

Abstract

Clustering is to divide data into several groups called “cluster”. Clustering is used in various fields such as data analysis, data mining, image processing, and pattern recognition. The k-means is one of the most famous clustering methods. Almost clustering methods require that the number of clusters is known in advance. Clustering is performed by specifying the number of clusters. In spite of it, in general, there are few data whose number of cluster is known in advance. Even when the number of clusters is unknown, it is important to estimate the number of clusters appropriately. X-means is one way to estimate the number of clusters. It is to suppose to the data generated from mixed isotropic Gaussian distribution and estimates the number of clusters by estimating the parameters of the probability distribution. It uses an estimator called information criterion. Many of information criterions are proposed by researchers like AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). In to estimate the number of clusters, however, it is not known which information criterion is valid for what kind of data. In this study, the optimal information criterion used for estimating the number of clusters is clarified through quantitative experiments.

目次

1. 緒言 1
2. 先行研究 1
   2.1 クラスタリング 1
   2.2 *k*-means 1
   2.3 最尤推定 2
   2.4 Mean shift 5
3. 実験手法 5
   3.1 X-means 5
   3.2 情報量規準 7
4. クラスタリング実験 8
   4.1 実験環境 8
   4.2 精度の評価 8
   4.3 2 次元データに対するクラスタ数推定 9
   4.4 3 次元データに対するクラスタ数推定 9
   4.5 手書き数字データのクラスタリング 11
5. 結論 12

参考文献 12

1. 緒言

クラスタリングとはデータを教師なし学習により任意の数のクラスタ（データのグループ）に分ける手法である．クラスタリングはデータ解析，データマイニング，画像処理，パターン認識など様々な分野で用いられる．k-means を始めとする多くのクラスタリング手法では，予めクラスタ数がわかっているものとして，クラスタ数を指定しクラスタリングを行う．しかし，データに対し最適なクラスタ数を指定しなければ，最適なクラスタリング結果を得ることはできない．それにも関わらず，一般にクラスタ数が事前にわかっているデータは少ない．その為，クラスタ数が未知である場合に対しても，適切にクラスタ数を推定しクラスタリングを行うことは重要な課題となっている．

クラスタ数推定を行う手法の一つに X-means がある．X-means は，データが混合等方 Gauss 分布から生成されたと想定して，その確率分布のパラメータを推定することにより，クラスタ数推定を行う．X-means は情報量規準とよばれる，確率分布とデータの分布の当てはまり具合（モデルの良さ）を表す指標を用い，クラスタ数推定を行う．その情報量基準は多くの研究者により様々なものが提案されている．代表的なものに，1973 年に赤池が提案した AIC (Akaike Information Criterion) や，Bayes の定理によって算出される事後確率を用いる BIC (Bayesian Information Criterion)，AIC の課題を解決した cAIC (Conditional Akaike Information Criterion) などがある．一般に X-means は，クラスタ数を決定する上で BIC を用いている．しかし，クラスタ数推定においてどの情報量規準がどのようなデータに対し有効かは分かっていない．そこで本研究では，クラスタ数推定に用いる情報量規準として最適なものを数値実験を通し明らかにする．

本研究では，AIC, cAIC, BIC と呼ばれる情報量規準をそれぞれ用いた X-means により混合等方 Gauss 分布から生成されたデータと実データのクラスタ数推定およびクラスタリングを行った．そして，どの情報量基準がクラスタ数推定に有効かを調査した．

本報告書の構成は次のとおりである．第 2 章では，既存のクラスタリング手法である k-means および Mean shift のアルゴリズムの紹介を行う．第 3 章では，本研究で利用する X-means の理論の説明を行う．まず，モデルと真の確率分布との近さを計る指標である情報量基準の例として Kullback-Leibler 情報量について述べ，それと最尤推定との情報量規準の関係性について詳しく述べる．その後，X-means の手法について述べる．第 4 章では，本研究により得られた実験結果について述べる．第 5 章では，本研究を通してのまとめおよび今後の課題について述べる．

2. 先行研究

2.1 クラスタリング

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である．クラスタとは，似ているデータの集まりである．クラスタリング手法において，データの類似の度合いはユークリッド距離やコサイン距離などの距離尺度を用い計算する．クラスタリングはデータ解析，データマイニング，画像処理，パターン認識など様々な分野で用いられる．

2.2 k-means

様々なクラスタリング手法の中で最も有名な手法が k-means である．k-means[1] は，d 次元空間上のデータについて，ユークリッド距離を用い各データ点が属するクラスタを決定する手法である．

d 次元空間上の確率変数 $\boldsymbol{x}$ の N 個のデータ点で構成されるデータ $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ があるとする．このデータを K 個のクラスタに分割することを考える．ここで，セントロイド $\boldsymbol{\mu}_k$ を導入する．セントロイド $\boldsymbol{\mu}_k$ とは

クラスタ k の重心を表す．k-means はクラスタに所属するデータ点とそのクラスタのセントロイド間のユークリッド距離の総和を最小にすることで，クラスタリングを行う．

ここで，各データ点 $\boldsymbol{x}_n$ に対し，対応する 2 値指示変数 $r_{nk} \in \{0,1\}$ $(k = 1, \cdots, K)$ を定める．これは，そのデータ点 $\boldsymbol{x}_n$ がクラスタ k に割り当てられるかを表す変数である．すなわち，データ点 $\boldsymbol{x}_n$ がクラスタ k に割り当てられる場合は $r_{nk} = 1$ とし，そうでない場合は $r_{nk} = 0$ とする．これは，1-of-K 符号化法として知られている．

次に，k-means における目的関数 J を定義する．

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \tag{1}$$

これは，各データ点からそれらが割り当てられたクラスタのセントロイド $\boldsymbol{\mu}_k$ までのユークリッド距離の 2 乗の総和を表している．k-means によるクラスタリングは，J を最小にする $\{r_{nk}\}$ と $\{\boldsymbol{\mu}_k\}$ の値を求めることに換言できる．

まず r_{nk} の決定を考える．(1) 式における J は r_{nk} についての線形関数なので，最適化は代数的に解くことができる．異なる n を含む項は互いに独立である．よって，各 n について別々に $r_{nk} = 1$ としたときに，$||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2$ が最小になるような k の値に対して r_{nk} を選んで 1 とおけばよい（(2) 式）．

$$r_{nk} = \begin{cases} 1 & k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{\mu}_j\| \text{のとき} \\ 0 & \text{それ以外} \end{cases} \tag{2}$$

つまり，単純に n 番目のデータ点がそれに最も近いセントロイドを持つクラスタに割り当てるのである．

次に，r_{nk} を固定したもとで $\boldsymbol{\mu}_k$ の最適化を考える．対象関数 J は $\boldsymbol{\mu}_k$ の二次関数であり，次のように $\boldsymbol{\mu}_k$ に関する偏微分を 0 とおく事で最小化できる（(3) 式）．

$$2 \sum_{n=1}^{n} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{3}$$

これを $\boldsymbol{\mu}_k$ についてとくと，(4) 式を得る．

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}} \tag{4}$$

この式の分母は k 番目のクラスタに割り当てられたデータの数に等しい．それゆえ $\boldsymbol{\mu}_k$ は k 番目のクラスタに割り当てられた全てのデータ点 $\boldsymbol{x}_n$ の平均値と単純に解釈することができる．

図 1 に k-means によるクラスタリングの具体例を示す．k-means では，r_{nk} と $\boldsymbol{\mu}_k$ をそれぞれ最適化する 2 つのステップを交互に繰り返す手続きでクラスタリングを実現する．最初に，$\boldsymbol{\mu}_k$ の初期値を選ぶ (図 1a)．次に，最初のフェーズで $\boldsymbol{\mu}_k$ を固定しつつ，r_{nk} について J を最小化する (図 1b)．第二フェーズでは，r_{nk} を固定しつつ，$\boldsymbol{\mu}_k$ について J を最小化する (図 1c)．そして，このような二段階最適化を収束するように繰り返す．

データ点のクラスタへの再割り当てと，クラスタ平均の再計算という 2 つのフェーズは，再割当てが起こらなくなるまで（もしくはあらかじめ定めた最大繰り返し数を超えるまで）繰り返される．各フェーズは，対象関数 J の値を減少させるので，このアルゴリズムの収束は保証されている．しかしながら，大域的最小点ではなく極小点に収束する可能性はある．

なお，k-means によるクラスタリングは，事前にクラスタ数を指定することによりクラスタリングを行うためクラスタ数が未知の場合，k-means を用いることはできない．

2.3 最尤推定

別の見方をすると k-means は確率分布の平均を求めているといえる．その考え方を発展させると，k-means はデータを生成する真の確率分布を推定する手法であるといえる．では k-means の結果が正しいかを判断す

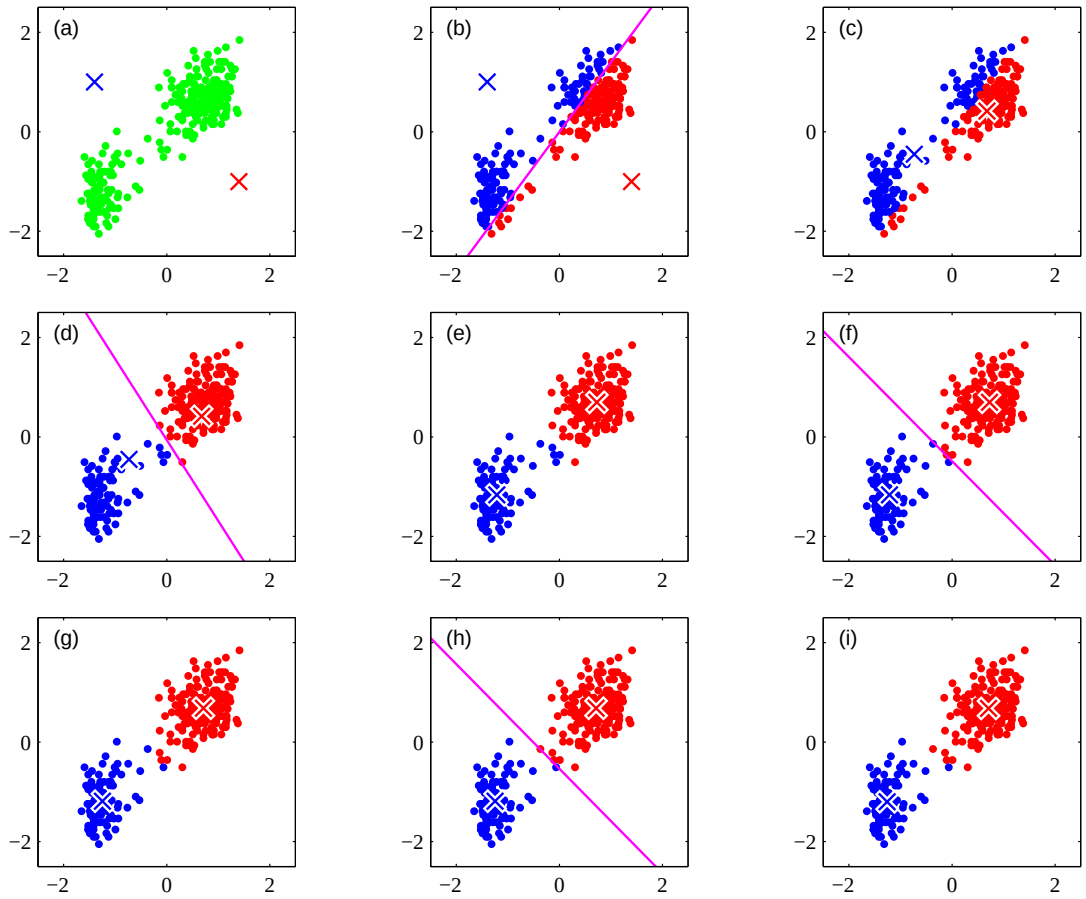

図 1　k-means の動作

るためには，k-means によって推定された確率分布（以後「モデル」）とデータを生成する真の分布がどれほど近いかを判断する必要がある．

真の分布（データ）と指定された確率分布の近さを測る客観的な規準として Kullback-Leibler 情報量（以後「K-L 情報量」）がある．連続型の確率分布のとき，$g(x)$ を真の確率密度関数，$f(x)$ をモデルが定める確率密度関数とすると，モデルに関する真の分布の K-L 情報量は $\log\{g(X)/f(X)\}$ の期待値を取り (5) 式で表される．

$$I(g \mid f) = \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx \tag{5}$$

ただし，log は常用対数ではなく自然対数で，注記がない限り一貫してこの意味で用いる．

このように，真の分布がわかっている場合には K-L 情報量によってモデルの良し悪しを比較できた．しかし，通常は真の分布が未知で，真の分布から得られたデータだけが与えられていることが多い．したがって，データから K-L 情報量を推定する必要がある．(5) 式を展開すると

$$\begin{aligned} I(g \mid f) &= \int_{-\infty}^{\infty} \left\{\log \frac{g(x)}{f(x)}\right\} g(x)dx \\ &= -\int_{-\infty}^{\infty} \{\log f(x)\} g(x)dx - \left(-\int_{-\infty}^{\infty} \{\log g(x)\} g(x)dx\right) \end{aligned} \tag{6}$$

となるが，右辺の第 2 項は定数であり，右辺第 1 項が大きいほど K-L 情報量 $I(g \mid f)$ は小さくなることがわかる．すなわち，K-L 情報量の大小比較のためには，本質的には $\int_{-\infty}^{\infty} \{\log f(x)\} g(x)dx$ だけを推定すれば良いことがわかる．右辺第 1 項の $\int_{-\infty}^{\infty} \{\log f(x)\} g(x)dx$ は，確率密度関数 $\log f(x)$ の期待値であり，平均対数尤度

と呼ばれている．ここで，

$$\sum_{i=1}^{n} \log f(x_i) \tag{7}$$

を対数尤度と呼ぶことにすると，n 個の独立な観測値 $\{x_1, x_2, \cdots, x_i\}$ が得られると，この平均対数尤度は，(8) 式のように対数尤度の n 分の 1

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i) \tag{8}$$

で近似される．したがって，符号に注意すると，対数尤度が大きいほど，そのモデルは真の分布に近いと考えられる．このようにして，対数尤度を K-L 情報量の推定値と考えることにすると異なったタイプのモデルの良し悪しも比較できるのである．

ところで，確率変数 $(X_1, X_2, \cdots, X_n)$ の同時密度関数が $f(x_1, x_2, \cdots, x_n \mid \theta)$ で与えられているものとする．θ は確率密度関数を規定するパラメータである．この時，観測値 $(x_1, x_2, \cdots, x_n)$ は与えられたものとして固定し，f を θ の関数と考える時，この関数を尤度と呼び，$L(\theta)$ で表す．すなわち，

$$L(\theta) = f(x_1, x_2, \cdots, x_n \mid \theta) \tag{9}$$

である．特に，確率変数が独立な場合には $(X_1, X_2, \cdots, X_n)$ の確率密度関数は，各 $X_i (i = 1, \cdots, n)$ の確率密度関数の積に等しいことから，

$$\begin{aligned} L(\theta) &= f(x_1 \mid \theta) f(x_2 \mid \theta) \cdots f(x_n \mid \theta) \\ &= \prod_{i=1}^{n} f(x_i \mid \theta) \end{aligned} \tag{10}$$

となる．この両辺の対数をとると，すでに求められた対数尤度関数

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{11}$$

が導かれる．

ここでは，平均対数尤度の推定量から対数尤度を直接導入した．しかし，モデルが確率分布の形で与えられている場合には，まず観測値の同時分布から尤度を定義し，その対数として対数尤度を求めるほうが都合が良い．$(X_1, X_2, \cdots, X_n)$ が独立でない場合にも，尤度の対数として対数尤度

$$l(\theta) = \log f(x_1, \cdots, x_n \mid \theta) \tag{12}$$

が定義できる．

ここまで，データに基づいて K-L 情報量の大小を比較するためには対数尤度を比較すれば良いことを示した．あらかじめ与えられたいくつかのモデルがある場合には，対数尤度が最大となるモデルを選択することによって，近似的には真の分布にいちばん近いモデルが得られることになる．したがって，モデルがいくつかの調整できるパラメータを保つ場合には，対数尤度を最大とするようにパラメータの値を選ぶことによって良いモデルが得られることがわかる．この推定を最大尤度法，略して最尤法と呼ばれている．また，最尤法で導かれた推定量は最尤推定量と呼ばれ，この最尤推定量によって定められるモデルが最尤モデルである．最尤モデルの対数尤度を最大対数尤度という．

2.4 Mean shift

本実験では，クラスタリング結果の比較のため，Mean shift[2] によるクラスタリングも行った．Mean shift は，データ点 $\boldsymbol{x}$ を標本点として得られるような確率密度関数 $f(\boldsymbol{x})$ を想定し，その標本点から確率密度関数 $f(\boldsymbol{x})$ の極大点を探索する手法である．図 2 に Mean shift によるクラスタリングの概念図を示す．ある任意の観測点 $\boldsymbol{y}_j$ から半径 h の超球（2 次元の場合は円）を考え，その範囲にある点群 $\boldsymbol{x}_i$ の平均 $\boldsymbol{x}_c$ を計算し，その位置に観測点 $\boldsymbol{y}_{j+1}$ を移動する．同様の操作を繰り返すと観測点は最大勾配の方向に移動し，極大点に収束する．

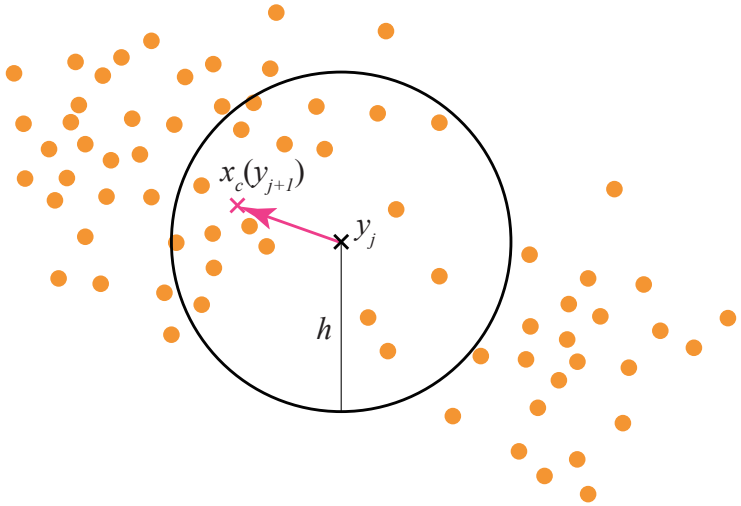

図 2　Mean shift によるクラスタリング

Mean shift を用いたクラスタリングは以下のように行う．

(1) 各点 $\boldsymbol{x}_i$ に Mean shift を適用し，その収束位置 $\boldsymbol{x}_i^c$ を計算する．
(2) 任意の 2 個の点 $\boldsymbol{x}_i, \boldsymbol{x}_j$ について，$\|\boldsymbol{x}_i^c - \boldsymbol{x}_j^c\| < \text{thereshold}$ ならこの 2 点を同じ極大点として同じクラスタとして扱う．

Mean shift は，k-means と異なり，クラスタ数をあらかじめ指定する必要がない．

3. 実験手法

本研究では，クラスタ数推定に X-means を用いる．そして，X-means で用いる情報量基準を変更することで，どの情報量基準がクラスタ数推定に有効か検証する．

3.1 X-means

X-means[3] は，データ分布が混合等方 Gauss 分布から生成されたと想定してクラスタ数推定及びクラスタリングを行う手法である．k-means の逐次繰り返しと，BIC による分割停止規準を用いることで，クラスタ数を推定しクラスタリングを実行する．

具体的には以下の手順で行われる．

(1) クラスタ数 k を初期化する (通常は $k = 2$) .
(2) k-means を実行する．

(3) 次の処理を $j = 1$ から $j = k$ まで繰り返す．

(a) クラスタ j の BIC_j を計算する．

(b) クラスタ j に所属するデータに対し，クラスタ数 2 として k-means を行う．

(c) クラスタ数 2 としてクラスタリングした結果に対し $\mathrm{BIC'}_j$ を計算する．

(d) BIC_j と $\mathrm{BIC'}_j$ を比較し，$\mathrm{BIC'}_j$ が大きければクラスタ数 k に 1 を足す．

(4) 前の処理で k が増加した場合は処理 (2) へ戻る．そうでない場合は終了する．

図 3 に X-means の具体的な動作例を示す．図 3a は 5 つの等方 Gauss 分布から生成されたデータである．クラスタ数の初期値 $k = 2$ で k-means を実行後，親と子の BIC を計算した結果が図 3b である．$k = 2$ で k-means によりクラスタリングした結果，上の 4 つのデータの塊と右下の 1 つの塊にクラスタリングされた．図中の丸印が親クラスタのセントロイド，ばつ印が子クラスタのセントロイドを表す．子クラスとは，クラスタ j に対しクラスタ数 2 でクラスタリングした結果得られたクラスタのことである．また，BIC ($k = 1$) が親クラスタの BIC，BIC ($k = 2$) が子クラスタの BIC を表す．図 3b の場合，赤色のクラスタでは子の BIC が親クラスタのものに比べ大きく，紫色のクラスタでは子の BIC が親クラスタのものに比べ小さい．よって親クラスタと子クラスタの BIC の大小関係から，赤色のクラスタのセントロイドは 2 つに分割し，紫色のクラスタのセントロイドは分割しない．その結果，クラスタ数が 1 つ増え，クラスタ数は 3 となる．次に，全データに対し，クラスタ数 3 で k-means を実行し，BIC を計算した結果が図 3c である．その結果，親クラスタと子クラスタの BIC の大小関係からクラスタを 1 つ増やす．同様に，図 3d では，クラスタ数 4 で k-means を実行し BIC 大小結果を比較し，クラスタを 1 つ増やしている．そして，クラスタ数が収束するまで行った結果が図 3e である．この場合，全てのクラスタにおいて親の BIC が大きくなっているので，クラスタ数 5 と推定された．この推定結果は実際のクラスタ数と一致しており，正確に推定できていることがわかる．

X-means で用いる BIC は次のように求められる．d 次元のデータ $\boldsymbol{D} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_d)$ を K 個のクラスタに分割することを考える．モデル M_j の評価に用いる BIC は (19) 式で与えられる．p_j はモデル M_j のパラメータ数であり，R は M_j のデータ数，$\hat{l}_j(D)$ は p 変量 Gauss 分布の対数尤度関数である．

等方 Gauss 分布を考えると分散 σ^2 は (13) 式により表される．

$$\hat{\sigma}^2 = \frac{1}{R-K} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}\right)^2 \tag{13}$$

すると，確率は次で表される．

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} \frac{1}{\sqrt{2\pi}\hat{\sigma}^d} \exp\left(-\frac{1}{2\hat{\sigma}^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2\right) \tag{14}$$

ここで $\boldsymbol{\mu}_i$ は d 次元の平均ベクトルである．したがって対数尤度関数は

$$\begin{aligned} l(D) &= \log \prod_i \hat{P}(x_i) \\ &= \sum_i \left(\log \frac{1}{\sqrt{2\pi}\sigma^d} - \frac{1}{2\sigma^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2 + \log \frac{R_{(i)}}{R}\right) \end{aligned} \tag{15}$$

となる．ここでクラスタ $n (1 \leq n \leq K)$ のデータ D_n に着目する．クラスタ n のデータ数を R_n と表記すると，(15) 式は以下で表される．

$$\begin{aligned} \hat{l}(D_n) = &-\frac{R_n}{2}\log(2\pi) - \frac{R_n \cdot d}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} \\ &+ R_n \log R_n - R_n \log R \end{aligned} \tag{16}$$

一般的には，分割停止規準として BIC を用いるが，本研究においては，BIC 以外の情報量規準を用いたクラスタリングも行い，クラスタ数推定の精度を検証する．

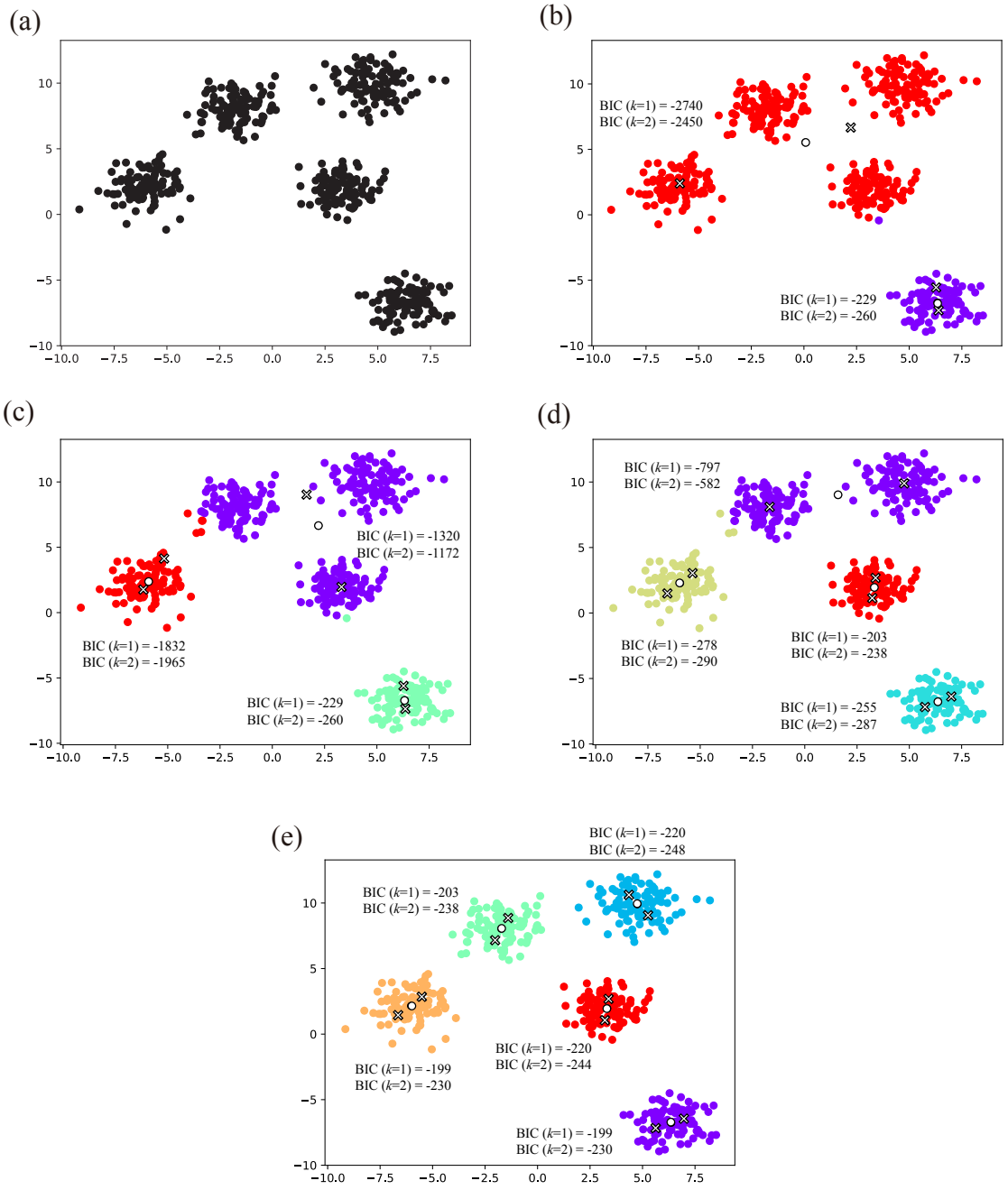

図 3　X-means の動作

3.2　情報量規準

情報量規準とは，最尤推定によって当てはめられたモデルが複数個あるときに，その中の一つを選択する規準である．

モデルの良し悪しは，最尤モデルの平均対数尤度のデータ $\boldsymbol{x}$ に関する期待値（期待平均対数尤度）によって考えることができる．期待平均対数尤度の値が大きいほど，そのモデルは良いといえる．モデルの最大対数尤度を期待平均対数尤度の 1 つの推定量と捉えることができるが，詳しく調べると，最大対数尤度そのものは期待平均対数尤度の不偏推定量にならないことがわかる．一般に最大対数尤度は，期待平均対数尤度の本当の値に比べて大きく出やすいという偏りを持つ．この傾向はモデルのパラメータ数が大きいほど著しい．これ

は最大対数尤度の比較によってモデルを選択すると，パラメータ数の大きいモデルほど選ばれやすいことを示している．

最大対数尤度の期待平均対数尤度に対する偏りの程度とモデルのパラメータ間の関係を調べると，

(モデルの最大対数尤度) – (モデルのパラメータ数)

が近似的に期待平均対数尤度の不偏推定量となることが導かれる．これが AIC と呼ばれる情報量規準である．AIC を最大とするモデルが最適なモデルと考えられる．AIC は最大対数尤度が同程度のモデルがあるときには，その中で実際に推定しなければならないパラメータの数が最も少ないものを選ぶべきであることを示している．

多くの情報量規準は，AIC の形式を踏襲し，

(モデルの最大対数尤度) – (モデルのパラメータ数などの罰則項)

という形をしている．

以下に，いくつか情報量規準の例を示す．なお，モデル M_j の p_j 変量等方 Gauss 分布の対数尤度関数を $\hat{l}_j$ と表し，モデルのデータ数を R と表す．

AIC (Akaike Information Criterion; **赤池情報量規準**)

1973 年に H. Akaike[4] によって提案された情報量規準である．

$$\mathrm{AIC}(M_j) = \hat{l}_j(D) - p_j \tag{17}$$

cAIC (Conditional Akaike Information Criterion; **条件付き赤池情報量規準**)

AIC は導出に漸近理論を使っているため，標本サイズが無限に大きいことを想定している．したがって，標本サイズが小さい場合はその過程が成り立たず，AIC によるモデル決定はパラメータ数を過大に見積もってしまう．

そこで，N. Sugiura[5] は漸近理論を使わない不偏推定量である cAIC を導出した．

$$\mathrm{cAIC}(M_j) = \hat{l}_j(D) - \frac{p_j \cdot R}{R - p_j - 1} \tag{18}$$

BIC (Beyesian Information Criterion; **ベイズ情報量規準**)

BIC は 1978 年に Schwartz[6] によって提案された．AIC とは異なり，罰則項に事後確率を利用する．

$$\mathrm{BIC}(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \ln R \tag{19}$$

4. クラスタリング実験

4.1 実験環境

実験には Python3.5 を用い，機械学習のライブラリとして TensorFlow を用いてアルゴリズムを実装した．プログラムは macOS 10.12 上で実行した．

4.2 精度の評価

クラスタリング精度の評価は以下の 3 つの指標により行った．

ARI; Adjusted Rand Index, **調整ランド指数**

クラスタの正解ラベルに対してクラスタリング結果の一致度を評価する指標である．1 に近づくほどよいクラスタリング結果と言える．

NMI; Normalized Mutual Information, **正規化相互情報量**

相互情報量を正規化した指標である．1 に近づくほどよいクラスタリング結果と言える．

Purity

生成されたクラスタがどれだけ多数派で占められているかを表す指標である．1 に近づくほどよいクラスタリング結果と言える．

これらの指標は，Python のライブラリである scikit-learn で用意されている関数により求めた．

4.3 2 次元データに対するクラスタ数推定

まず，2 次元データに対するクラスタ推定の結果を比較する．実験では，2 次元空間に図 4 のような分散 $\sigma^2 = 1$ の混合等方 Gauss 分布を用いた．この混合等方 Gauss 分布は 5 つの等方 Gauss 分布で構成される．そして，各クラスタは 500 個のデータ点を持つ．このデータに対し，対数尤度関数，BIC, AIC, cAIC を分割停止規準として採用して X-means によるクラスタ推定およびクラスタリングを行った．

表 1 は X-means により推定されたクラスタ数，X-means により得られたクラスタリング結果の ARI, NMI, Purity を示す．各データは，100 回ランダムに生成されたデータに対してクラスタ数推定およびクラスタリングをそれぞれ行ったあと得られた結果の平均であり，括弧内の数値は分散を表す．BIC，cAIC，AIC を情報量基準を用いた X-means により推定されたクラスタ数はほぼ真のクラスタ数と一致していた．さらに，推定したクラスタ数の分散もあまり大きな値とはなっておらず，安定して真のクラスタ数を推定している．しかし，AIC を分割停止規準として採用した場合に着目すると，推定したクラスタ数の平均は真のクラスタ数と近い値となっているが，その分散は大きくなった．これは，3.3 節で述べたように，AIC はクラスタ数を過大に見積もってしまうことに起因すると思われる．実際に，推定されたクラスタ数を見ると，クラスタ数を 20 や 22 と推定しているものが多く存在した．しかし，対数尤度関数を分割停止規準として採用した場合のクラスタリング結果と比較すると，比較的安定したクラスタ数推定を行っていることが見て取れる．一方で，どの情報量基準を使った場合もクラスタリングの精度に大きな差がないことが分かった．

表 1　2 次元空間におけるクラスタリング結果

分割停止規準	クラスタ数 (分散)	ARI	NMI	Purity
BIC	4.58 (0.9836)	0.84458792	0.88281495	0.84458792
cAIC	4.55 (0.6475)	0.85329139	0.89992544	0.85329139
AIC	4.69 (3.8739)	0.83642236	0.88147442	0.83642236
対数尤度関数	5.32 (10.236)	0.85699618	0.91572100	0.85699618

4.4 3 次元データに対するクラスタ数推定

次に，3 次元データに対するクラスタ推定の結果を比較する．3 次元空間に図 5 のようにデータを分散 $\sigma^2 = 1$ の混合等方 Gauss 分布より生成した．実験で使用するデータは図のように 5 つのクラスタで構成されており，各クラスタは 500 個のデータ点を持つ．このデータに対し，対数尤度関数，BIC, AIC, cAIC を分割停止規準として採用してクラスタリングを行った．

表 2 に都度ランダムに生成されたデータに対して 100 回クラスタリングを行ったときの，推定されたクラスタ数，ARI, NMI, Purity の平均値および推定されたクラスタ数の分散を示す．2 次元データの場合とは異なり，3 次元データのクラスタ推定およびクラスタリングにおいては BIC を分割停止規準として採用した場合の精度が最も良くなっている．推定されたクラスタ数の分散も非常に小さいため，安定して精度の高いクラス

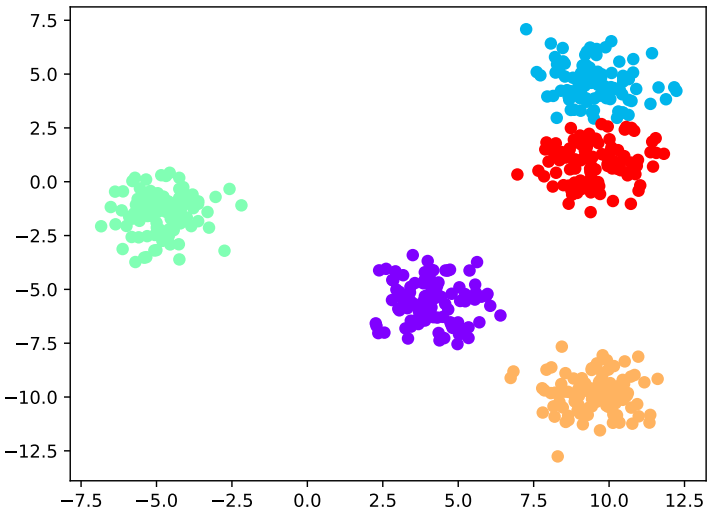

図 4　2 次元空間のクラスタリング例

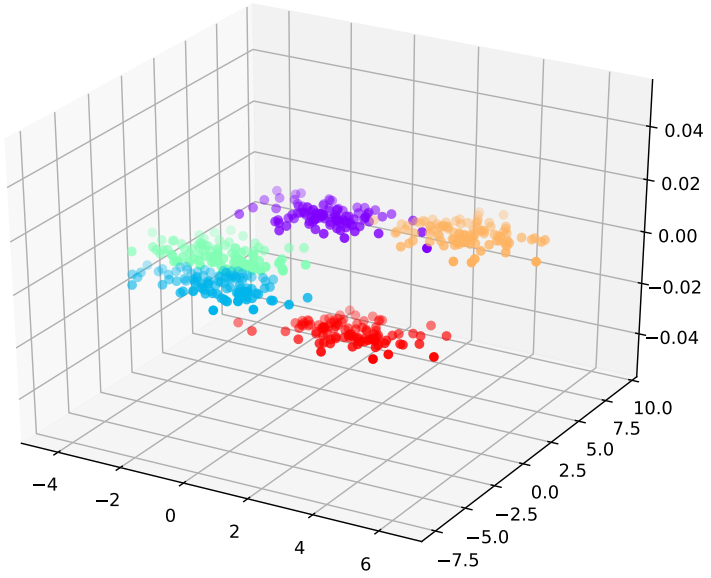

図 5　3 次元空間のクラスタリング例

タ数推定を行っていることが伺える．cAIC と AIC の場合を比較した場合，cAIC のほうがクラスタリングの精度が高いことがわかる．また，AIC を用いた場合では，2 次元データのときほど推定されたクラスタ数の分散が大きくないことがわかる．この原因は，2 次元空間ではクラスタ数を過大に見積もってしまう問題があったが，3 次元空間においてその問題は発生していなかったからだと考えられる．

全体的に 3 次元空間に生成したデータのクラスタリング結果が 2 次元のものと比べ向上している．これは，2 次元空間と 3 次元空間とでは，空間の体積が異なるためだと考えられる．同一条件でデータを生成したため，2 次元空間においてはクラスタ同士が重なり合っているものが多かったが，3 次元空間ではほぼ重なり合ったクラスタは存在していなかった．したがって，3 次元空間のほうがクラスタリング精度が良くなっているようにみえるデータを取得したものと考えられる．

表 2　3 次元空間におけるクラスタリング結果

分割停止規準	クラスタ数 (分散)	ARI	NMI	Purity
BIC	4.95 (0.0669)	0.97179074	0.97913818	0.97179074
cAIC	4.92 (0.2313)	0.96312702	0.97023920	0.96312702
AIC	4.88 (0.1443)	0.95216819	0.96855698	0.95216819
対数尤度関数	5.12 (4.1443)	0.95731637	0.96541468	0.95731637

4.5　手書き数字データのクラスタリング

次に，実データの例として MNIST (Modified National Institute of Standards and Technology database) と呼ばれる手書き数字データのクラスタリングの実験を行った．

MNIST は 70,000 枚の手書きのアラビア数字 (0-9) の 28×28 画素の画像データセットである．図 6 にデータセットの一部を示す．

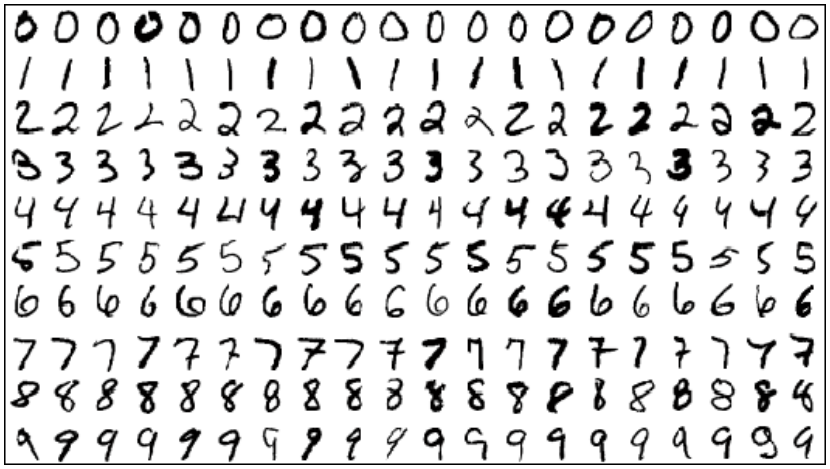

図 6　MNIST の一部

この MNIST のデータを X-means および Mean shift によりクラスタリングした．X-means の分割停止規準には，AIC, cAIC, BIC, 対数尤度関数を用いた．また，Mean shift の際に用いる超球の大きさは Python のライブラリである scikit learn の関数 (cluster.estimate_bandwidth) により決定した．

結果を表 3 に示す．

表 3　MNIST のクラスタリング結果

クラスタリング手法	クラスタ数	ARI	NMI	Purity
X-means (AIC)	32	0.282377190261	0.562954278053	0.282377190261
X-means (cAIC)	32	0.28312802996	0.562243771104	0.28312802996
X-means (BIC)	32	0.27811599645	0.56101889037	0.27811599645
X-means (対数尤度関数)	32	0.28736538966	0.569692653863	0.28736538966
Mean shift	1	0.0	$-6.93889390391 \times 10^{-6}$	0.0

X-means では，どの分割停止規準を用いた場合も同様のクラスタリング結果となり，適切なクラスタリングが行われていないことがわかる．本実験では X-means のモデルとして等方混合 Gauss 分布を用いたが，実

データのクラスタリングを行うときにはより適したモデルを選択する必要があると考えられる．また，Mean shift によるクラスタリングの結果，確率密度の極大点を見つけることができず，1 クラスタに収束した．

5. 結論

本研究では，いくつかの情報量規準を分割停止規準として採用して X-means でクラスタ数推定を行った．

まず，混合等方 Gauss 分布から生成したデータのクラスタ数推定を行った．2 次元空間における混合等方 Gauss 分布から生成したデータセットのクラスタ数推定においては BIC や cAIC が，3 次元空間における混合等方 Gauss 分布から生成したデータセットのクラスタ数推定においては BIC が適していることがわかった．2 次元空間においては AIC を採用した場合，クラスタ数を過大に見積もってしまう問題が見られた．AIC は導出に漸近理論を用いているため，パラメータ数を過大に見積もるという問題点があり，それに起因してクラスタ数を課題に見積もったと考えられる．cAIC は AIC のこの問題を解決するために，漸近理論を用いずに導出された情報量規準である．本研究でも cAIC を用いることで，AIC を用いた場合よりも非常に高い精度でクラスタ数を推定することが確認できた．したがって，データ数が比較的少ない場合は cAIC を情報量規準として採用することで，AIC よりもよい精度でクラスタ数推定ができるといえる．しかし，cAIC は 3 次元の等方 Gauss 分布から生成されるデータのクラスタ数推定では BIC の結果より大きく劣っている．したがって，混合等方 Gauss 分布により生成されたデータのクラスタリングを行う際は等方 Gauss 分布をモデルとする X-means の分割停止規準として BIC を採用することで適切なクラスタ数推定を行うことができるといえる．また，本研究では 2 次元空間と 3 次元空間で同一条件でデータを生成しクラスタリングを行ったため，後者のほうがデータが分布する空間が広い．そのため，クラスタ同士の重なり合いが少なくなり，クラスタ数推定の精度がいずれも向上している．

実データ (MNIST) のクラスタ数推定を行った場合，いずれの情報量規準を分割停止規準を用いた場合も，ほぼ同一の結果となり，等方 Gauss 分布をモデルとする X-means では適切なクラスタ数推定を行うことができないことがわかった．クラスタ数推定に失敗した原因は 3 つ考えられる．1 つは，データ自体が 10 のクラスタに分かれていないかもしれない点である．手書き文字の場合，同一の数字でも様々な書き方が存在しているため，同一の数字の画像が 1 つのクラスタを形成しているとは限らない．2 つは，クラスタ推定をしやすくするための前処理をデータに対し施さなかった点である．手書き画像は，線の太さも濃さも様々である．そのような画像データの場合，同じ形の文字であっても，線の太さや濃さの違いでデータとしては全く別のものになってしまう．3 つは，MNIST においてクラスタ数推定，MNIST の分布が等方 Gauss 分布に従っていないことが要因と考えられる．適切なクラスタ数推定を行うためには，他のモデルを用いるなどの工夫をする必要があると考えられる．また，確率ベースではないクラスタ数推定を行う手法として Mean shift によるクラスタリングも同様に行ったが，複数の確率密度関数の極大点を見つけることができず，適切なクラスタリング結果を得ることはできなかった．

以上の段落より，混合等方 Gauss 分布から生成される人工データのクラスタリングには BIC が最も適していることがわかった．また，AIC は少量のデータのクラスタ数推定には向かず，cAIC を利用することで適切なクラスタ数推定を行うことができる．実データのクラスタ数推定を行う際には，そのデータにあったモデルやクラスタ数推定の手法を採用する必要があり，今後実データのクラスタ数を適切に推定するための手法を検討していく必要がある．

参考文献

1) James MacQueen et al.: Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281–297 (1967).
2) 奥富 正敏他: ディジタル画像処理 [改訂新版], pp.205-208, 公益財団法人 画像情報教育振興協会 (2015).
3) Dan Pelleg, Andrew W Moore, et al.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters., ICML, Vol. 1, pp. 727–734 (2000).
4) Akaike, H.: Information theory and an extension of the maximum likelihood principle, Proceedings of the 2nd International Symposium on Information Theory, pp. 267-281 (1973).
5) Sugiura, N.: Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's, Communications in Statistics-Theory and Methods, 7(1), pp. 13-26 (1978).
6) Gideon Schwarz et al.: Estimating the dimension of a model., The annals of statistics Vol. 6, No. 2, pp. 461-464 (1978).