論　文

ディリクレ過程平均法のレートひずみ理論による解釈

小林真佐大[†a)]　　渡辺　一帆[†b)]

A Rate-Distortion Theoretic View of Dirichlet Process Means Clustering

Masahiro KOBAYASHI[†a)] and Kazuho WATANABE[†b)]

あらまし　ディリクレ過程平均法はクラスタリングの代表的手法である K-平均法を拡張した手法であり，クラスタ数をデータから推定することができる．クラスタ数を指定する代わりに，ペナルティパラメータと呼ばれるしきい値を指定する必要があるが，ペナルティパラメータの変化に対するクラスタ数の振る舞いは未だに明らかにされていない．本研究では，ペナルティパラメータとクラスタ数の対数を次元で割った値との組がそれぞれレートひずみ理論における最大ひずみとレートに対応することに着目し，データ数とデータの次元が無限大の極限において，ペナルティパラメータに対応するクラスタ数の曲線がレートひずみ曲線に近づくことを示す．数値実験により，学習データ数が有限であることの影響を受けにくいレートが 0 の近辺において，レートひずみ曲線に近づくことを確認し，ペナルティパラメータと学習データ中の最大ひずみとの対応を示す．

キーワード　クラスタリング，ディリクレ過程，レートひずみ曲線，ひずみ有りデータ圧縮

1. まえがき

クラスタリングはデータ集合を複数の部分集合クラスタに分割するデータ解析の手法であり，画像処理，データマイニングなど様々な分野で利用されている．とりわけよく利用される方法として，K-means 法が挙げられる．K-means 法では事前にクラスタ数の指定をする必要があるが，データ数や次元が膨大な場合，クラスタ数の予想は立てづらく，何らかの発見的手法や複数のクラスタ数での結果を吟味することなどが必要となる．クラスタ数の推定には，ノンパラメトリックベイズ法による混合正規分布の学習法が提案され [1]，その分散 0 の極限において，クラスタ数をデータから推定することができる K-means 法の拡張が得られることが示された．この手法は，ディリクレ過程平均法（以下，DP-means 法）[2] として知られている．

DP-means 法では，事前にクラスタ数を指定せず，アルゴリズム中でクラスタ数の推定を行う．K-means 法と同様にアルゴリズムが簡便であり，データが大規模な場合に適用可能という利点をもつ．また，データが二値や非負整数値などの特殊な型をもつ場合に適切な距離尺度を導入するための指数型分布族を用いた拡張 [3], [4] が与えられている．計算時間削減のためのアプローチとして，楽観的並行性制御を取り入れた並列化による方法 [5] や，データを重み付きサブセットに分割してクラスタリングを行うことで，若干の精度を犠牲にして計算時間を大幅に短縮する方法 [6] がある．更に，外科手術用マニピュレータの位置制御への応用において，オンライン DP-means 法が考案されている [7]．

DP-means 法ではクラスタ数を指定する代わりに，クラスタ数を増やす指標として，ペナルティパラメータと呼ばれるしきい値を指定する必要がある．ペナルティパラメータを自動探索する方法 [8] が試みられているが，ペナルティパラメータの変化に伴うクラスタ数の振る舞いは未だに明らかにされておらず，ペナルティパラメータの設定法は確立していない．

一方で，クラスタリングをひずみ有りデータ圧縮とみなすと，クラスタ数の対数に対応するレートと，データとクラスタ中心間の擬距離で与えられるひずみとの間のトレードオフは，情報理論の一分野であるレートひずみ理論において研究されている [9]．レートひずみ理論では，ひずみの測り方として，一般的に

† 豊橋技術科学大学，豊橋市
Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan
a) E-mail: m143320@edu.tut.ac.jp
b) E-mail: wkazuho@cs.tut.ac.jp

扱われる平均ひずみの他に，最大ひずみも扱われており，対応するレートはデータ数とデータの次元が無限大の極限でレートひずみ曲線に近づくことが示されている [9].

本研究では，DP-means 法のペナルティパラメータはレートひずみ理論における最大ひずみと対応付けられることを指摘し，ペナルティパラメータを変化させたときのクラスタ数の曲線がデータ数及び次元が無限大になる極限においてレートひずみ曲線に近づくことを示す.

また，実際に数値実験を行い，DP-means 法のペナルティパラメータの振る舞いが，最大ひずみとほぼ同様の軌跡を描くことを確認した．特にレートが低い領域において，DP-means 法のペナルティパラメータがレートひずみ曲線の値に収束していく様子を確認した．しかし，レートが一定であるときのペナルティパラメータと最大ひずみの間には多少の差異が存在した．この差異を小さくすることを目的として，DP-means 法のアルゴリズムを改変したアルゴリズムを提案し実験を行う．オリジナルの DP-means 法のアルゴリズムと改変後のアルゴリズムにおいてペナルティパラメータと最大ひずみの差異を比較すると，改変後において特にひずみが大きいときの差異が小さくなることが確認できた.

2. DP-means 法

DP-means 法は，データ $\boldsymbol{x}^n = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ と，ペナルティパラメータ λ を入力として必要とする．なお，データの次元は L 次元とする，すなわち，$\boldsymbol{x}_i = (x_i^{(1)}, ..., x_i^{(L)}) \in \boldsymbol{R}^L$．クラスタは一つから始め，基本的には K-means 法と同様に，クラスタ中心の計算とデータ点のクラスタへの割り当てを収束するまで実行する．$\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ をクラスタ中心とすると，新しいクラスタが追加されるのは，ペナルティパラメータ λ より，データ点 $\boldsymbol{x}_i$ とクラスタ中心 $\boldsymbol{\theta}_{c(i)}$ の擬距離の値が大きい，すなわち (1) を満たすときである.

$$d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right) > \lambda \tag{1}$$

ここで，$c\left(i\right) \equiv \arg\min_k d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_k\right)$ は $\boldsymbol{x}_i$ のクラスタラベルを示す．DP-means 法のアルゴリズムを Algorithm 1 に示す．DP-means 法と K-means 法には二つの違いがある．一つ目は，初期化時のクラスタ数である．DP-means 法では，クラスタ数は一つとして，クラスタ中心をデータ点全体の平均で初期化するが，

Algorithm 1 DP-means

Input: $\boldsymbol{x}^n = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}, \lambda$
Output: $\boldsymbol{l} = \{\boldsymbol{l}_1, ..., \boldsymbol{l}_K\}, K$

```
K = 1
l_1 = x^n
θ_1 = (1/n) Σ_{i=1}^n x_i
c(i) = 1 (i = 1, ..., n)
repeat
  for i = 1 to n do
    d_ik = d_L(x_i, θ_k) (k = 1, ..., K)
    if min_k d_ik > λ then
      K = K + 1
      c(i) = K
      θ_K = x_i
    else
      c(i) = arg min_k d_ik
    end if
  end for
  for j = 1 to K do
    l_j = {x_i | c(i) = j}
    θ_j = (1/|l_j|) Σ_{x∈l_j} x
  end for
until Σ_{i=1}^n d_L(x_i, θ_{c(i)}) + λK converges
```

K-means 法では，事前に指定されたクラスタの数だけクラスタを作り，クラスタ中心は何らかの方法（ランダムなど）で初期化する．二つ目は，クラスタの追加についてである．DP-means 法では，データ点のクラスタへの割り当てを行うループ中において，(1) を満たすかどうかの条件判定を行い，満たすときはクラスタの追加をし，満たさないときはクラスタラベルの更新をする．一方，K-means 法では，データ点のクラスタへの割り当てを行うループ中において，クラスタラベルの更新をするのみである.

なお，本論文では，擬距離には 2 乗距離を一般化したブレグマンダイバージェンスを仮定する．具体的には，凸関数 ϕ から決まるブレグマンダイバージェンス d_ϕ から次元に関して加法的に定義される以下を擬距離とする.

$$d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right) \equiv \frac{1}{L}\sum_{j=1}^{L} d_\phi\left(x_i^{(j)}, \theta_{c(i)}^{(j)}\right)$$

$$d_\phi\left(x, \theta\right) \equiv \phi\left(x\right) - \phi\left(\theta\right) - \left(x - \theta\right)\phi'(\theta)$$

ここで，ϕ は微分可能な凸関数であり，ϕ' はその微分を表す．ブレグマンダイバージェンスは指数型分布族に属する確率分布と一対一に対応している．例えば，データが二値，非負整数値といった特定の型をもつ場合に，通常の実数値に対する 2 乗距離よりも適した距離尺度として，対応するベルヌーイ分布やポアソ

ン分布のブレグマンダイバージェンスが用いられている [3], [4].

3. レートひずみ理論

3.1 平均ひずみと最大ひずみ

レートひずみ理論において，次元はブロック長に対応し，クラスタ数は符号語数に対応するため，レートは次元あたりのクラスタ数の対数として，(2) で定義される．

$$r \equiv \frac{\ln K}{L} \tag{2}$$

また，レートに対応するひずみは平均ひずみ，最大ひずみの 2 種類を考え，それぞれ擬距離より (3)，(4) で定義される．

$$D_a \equiv \frac{1}{n}\sum_{i=1}^{n} d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right) \tag{3}$$

$$D_m \equiv \max_{1\le i\le n} d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right) \tag{4}$$

ここで，データの経験分布に従う L 次元確率ベクトルを $\boldsymbol{X}$ とし，$c^*(\boldsymbol{X}) \equiv \arg\min_k d_L(\boldsymbol{X}, \boldsymbol{\theta}_k)$ とすると，

$$D_m = \inf\left\{\alpha \,\middle|\, \Pr\left\{d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right) > \alpha\right\} = 0\right\}$$

と書き直すことができる．データの次元が無限大の極限において，この値は確率変数列 $\left\{d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)\right\}_{L=1}^{\infty}$ に対して，

$$\inf\left\{\alpha \,\middle|\, \lim_{L\to\infty} \Pr\left\{d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right) > \alpha\right\} = 0\right\}$$

に収束すると考えられる．すなわち，確率的上極限を用いて，

$$p\text{-}\limsup_{L\to\infty} d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)$$

と表される [9]．(2) のレートについて同様の極限をとると，

$$\limsup_{L\to\infty} \frac{\ln K}{L}$$

であり，これらは [9, 定義 5.3] における固定長符号化の最大ひずみ基準とレートの定義にそれぞれ一致し，最大ひずみが一定値 D のときの達成可能なレートの下限値としてレートひずみ曲線が定義される．また，データが情報源の分布に従う場合，その経験分布はデータ数無限大の極限で情報源の分布に近づくことから，上述のレートひずみ曲線は，情報源の分布に対するレートひずみ曲線に近づくと考えられる [10].

情報源 $p(\boldsymbol{x})$ が i.i.d. のとき，すなわち，各次元 L において $p(\boldsymbol{x}) = \prod_{j=1}^{L} p\left(x^{(j)}\right)$ で与えられるとき，分布 $p(x)$ に従う確率変数を X，分布 $\int q(\theta|x)\,p(x)\,dx$ に従う確率変数を Θ とすると，最大ひずみ基準のレートひずみ曲線は，平均ひずみに対して $L\to\infty$，$n\to\infty$ の極限から同様に得られるレートひずみ曲線に一致し，

$$R(D) = \inf_{q(\theta|x):E_{X,\Theta}\left[d_\phi(X,\Theta)\right]\le D} I(X;\Theta)$$

により与えられることが知られている [9, 定理 5.8], [11]．ここで，$I(X;\Theta)$ は X と Θ の相互情報量，$E_{X,\Theta}$ は $q(\theta|x)p(x)$ に関する期待値を表す．なお，レートひずみ曲線の値が 0 となる最小のひずみを $D_{\max}$ と表す．一般のひずみ尺度 d に対し，$D_{\max}$ は $\inf_\theta E_X[d(X,\theta)]$ で与えられ [12]，ブレグマンダイバージェンスに対しては，一般に $\arg\min_\theta E_X[d_\phi(X,\theta)] = E_X[X]$ のため，$D_{\max} = E_X\left[d_\phi(X, E_X[X])\right]$ である [3].

また，データの分布が二つの確率分布 p_1，p_2 の混合によって与えられる混合情報源

$$p(x) = \alpha p_1(\boldsymbol{x}) + (1-\alpha)p_2(\boldsymbol{x}) \tag{5}$$

$(0 < \alpha < 1)$ では，それぞれの確率分布に対応したレートひずみ曲線が $R_1(D)$，$R_2(D)$ としたとき，最大ひずみに対応したレートひずみ曲線は，それらの最大で与えられる [9, 定理 5.10].

$$R(D) = \max\{R_1(D), R_2(D)\} \tag{6}$$

3.2 DP-means 法とレートひずみ曲線の関係

本節では，まず DP-means 法におけるペナルティパラメータと，そのときに与えられる最大ひずみが同様の意味をもつことを示す（定理 1）．そして，データ数と次元数が無限大の極限において，ひずみを一定値 λ としたときに DP-means 法により決まるレートがレートひずみ曲線を達成することについて，情報源が単一の i.i.d. 情報源の場合（定理 2）と混合情報源の場合（定理 3）に分けて示す．

DP-means 法では，推定クラスタ数はデータの並びとクラスタ中心の初期値に依存する．ここでは，議論を単純化するため，仮定 1 をおく．

［仮定 1］ DP-means 法停止時の推定クラスタ数が最小となり，そのときのクラスタ中心 $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ に対する平均ひずみを最小とする最適解が得られるような，データの並びとクラスタ中心の初期値を仮定する．

[定義 1] 仮定 1 のもとで，クラスタ数 K とクラスタ中心 $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ から決まる，レート，平均ひずみ，最大ひずみを (2)〜(4) より，それぞれ $r(K)$，$D_a(K)$，$D_m(K)$ と定義する．

[仮定 2] $D_m(K)$ に対し，単調性

$$D_m(K) > D_m(K+1) \tag{7}$$

が全ての $K \geq 1$ で成り立つとする．

次元数 L が大きいとき，チェビシェフの不等式より最大ひずみは平均ひずみに近づくため（後述 (12)），仮定 2 の単調性は高い確率で成立すると考えられる．これらの仮定のもと，次の定理が成り立つ．

[定理 1] DP-means 法実行時におけるクラスタ数が K になるペナルティパラメータ λ の下限を $\lambda(K)$ とすると，全ての K について (8) が成り立つ．

$$\lambda(K) = D_m(K) \tag{8}$$

[証明 1] データの平均 $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ に対し，

$$\lambda(1) = D_m(1) = \max_{1 \leq i \leq n} d_L\left(\boldsymbol{x}_i, \bar{\boldsymbol{x}}\right)$$

が成り立つ．$k \leq K-1$ $(K \geq 2)$ において $\lambda(k) = D_m(k)$ を仮定すると，$\lambda = D_m(K)$ とした DP-means 法は，全ての i で $d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right) \leq \lambda$ が成り立つことからクラスタ数 K 以下で停止するが，(7) より

$$\lambda = D_m(K) < D_m(K-1) = \lambda(K-1)$$

となることから，クラスタ数は K 未満とはならないため，クラスタ数 K で停止する．また，$\lambda = D_m(K) - \varepsilon(\varepsilon > 0$ は任意の小さい定数$)$ とすると，$i^* = \arg\max_i d_L\left(\boldsymbol{x}_i, \boldsymbol{\theta}_{c(i)}\right)$ に対し，

$$d_L\left(\boldsymbol{x}_{i^*}, \boldsymbol{\theta}_{c(i^*)}\right) = D_m(K) > \lambda$$

が成立し，少なくとも $K+1$ 個のクラスタが必要となる．以上より，$D_m(K) - \varepsilon \leq \lambda(K) \leq D_m(K)$ が全ての $\varepsilon > 0$ に対し成り立ち，(8) の成立が確かめられる．(証明終)

(8) は，ペナルティパラメータが最大ひずみと同様の意味をもつということを意味する．なお，本論文では，DP-means 法のペナルティパラメータを変化させたときのクラスタ数から (2) で決まるレートの曲線のことを「クラスタ数の曲線」と呼んでいる．

以降の議論では，仮定 1，仮定 2 に加え以下の仮定をおく．

[仮定 3] i.i.d. 情報源を表す確率変数 X が離散の場合は，有限集合上に値を取るとし，連続の場合は，その確率密度関数 $p(x)$ が有界な台 S をもつとする(注1)．

[仮定 4] ひずみ尺度に対し，

$$E_X\left[d_\phi(X, E_X[X])^2\right] < \infty$$

とする．

[定理 2] i.i.d. 情報源に対して，$n \to \infty$ かつ $L \to \infty$ の極限において，ペナルティパラメータを λ とした DP-means 法により決まるレートは，ひずみを λ としたレートひずみ曲線の値を達成する．

$$\lim_{L\to\infty} \lim_{n\to\infty} r(K) = R(\lambda) \tag{9}$$

[証明 2] レート r を固定すると (2) より，クラスタ数は

$$K = \lceil \exp(Lr) \rceil$$

で与えられる．このとき定義 1 より平均ひずみ $D_a(K)$ と最大ひずみ $D_m(K)$ が一意に決まる．更に仮定 2 と定理 1 より，ペナルティパラメータ $\lambda(K)$ に対する単調性を考慮すると，クラスタ数が K となるペナルティパラメータ λ の区間は次式で表される．

$$\lambda(K) \leq \lambda < \lambda(K-1) \tag{10}$$

ここで，情報源の分布 $p(\boldsymbol{x})$ 及び，それに従う L 次元確率ベクトル $\boldsymbol{X}$ に対して，クラスタ数が K，クラスタ中心が $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ であるときに与えられる平均ひずみと最大ひずみを

$$D_a^*(K) \equiv E_{\boldsymbol{X}}\left[d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)\right]$$
$$D_m^*(K) \equiv \max_{\boldsymbol{x} \in S^L} d_L\left(\boldsymbol{x}, \boldsymbol{\theta}_{c^*(\boldsymbol{x})}\right)$$

と定義する（S^L は S の L 次の直積）．すると，$n \to \infty$ の極限において，グリベンコ・カンテリの定理 [13, 定理 17.3] より，データの経験分布は情報源の分布に収束するため，

$$D_a^*(K) = \lim_{n\to\infty} D_a(K)$$
$$D_m^*(K) = \lim_{n\to\infty} D_m(K) \tag{11}$$

が成り立つ．更に，チェビシェフの不等式より，任意

(注1)：$\{x : p(x) \neq 0\}$ の閉包を $p(x)$ の台という．

の $\gamma > 0$ に対し，

$$\Pr\left\{ \left| d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right) - D_a^*(K) \right| \geq \gamma \right\} \leq \frac{1}{\gamma^2} V_{\boldsymbol{X}}\left[d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)\right] \leq O\left(\frac{1}{L}\right) \tag{12}$$

が成り立つ．ここで，$V_{\boldsymbol{X}}$ は $\boldsymbol{X}$ に関する分散を表し，各次元の独立性より，

$$V_{\boldsymbol{X}}\left[d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)\right] = \frac{1}{L^2}\sum_{j=1}^{L} V_{X^{(j)}}\left[d_\phi(X^{(j)}, \theta^{(j)}_{c^*(X)})\right]$$

となること，及び仮定 4 から $E_X\left[d_\phi(X^{(j)}, \theta^{(j)}_{c^*(X)})^2\right] \leq E_X\left[d_\phi(X, E_X[X])^2\right] < \infty$ となることを用いた．(12) より，$L \to \infty$ の極限では，$d_L\left(\boldsymbol{X}, \boldsymbol{\theta}_{c^*(\boldsymbol{X})}\right)$ は $D_a^*(K)$ に確率収束するため，最大ひずみは平均ひずみに近づく．

$$\lim_{L\to\infty} D_m^*(K) = \lim_{L\to\infty} D_a^*(K) \tag{13}$$

ここで，K-means 法は，大域的最適解が得られているとすれば，有限次元において一般のブレグマンダイバージェンスに対し平均ひずみを最小化する最適な手法であるため [3]，$n \to \infty$ かつ $L \to \infty$ の極限ではレートひずみ曲線を達成することが知られている [11, Section III]．したがって，レートひずみ曲線 $R(D)$ の逆関数，ひずみレート曲線を $D(R)$ と表したとき，次式が成り立つ．

$$\lim_{L\to\infty} D_a^*(K) = D(r) \tag{14}$$

(11)，(13)，(14) より，

$$\lim_{L\to\infty}\lim_{n\to\infty} m(K) = D(r) \tag{15}$$

となる．ここで，$r = r(K)$ に対し，

$$\begin{aligned} r(K-1) &= \frac{\ln(K-1)}{L} = \frac{\ln\{K(1-\frac{1}{K})\}}{L} \\ &= \frac{\ln K}{L} + \frac{1}{L}\ln(1-\frac{1}{K}) \\ &\simeq r - \frac{1}{LK} \end{aligned}$$

より，$L \to \infty$ の極限では，$r(K-1)$ は r に近づく．このとき，$D_m(K-1) = \lambda(K-1)$ は $D_m(K) = \lambda(K)$ に近づき，(10) の区間の幅は 0 に近づく．よって，任意の $\lambda \in [\lambda(K), \lambda(K-1))$ に対し，(15) より

$$\lim_{L\to\infty}\lim_{n\to\infty} r(K) = R(\lambda)$$

が成り立ち，(9) の成立が確かめられる．(証明終)

［定理 3］ i.i.d. 情報源により構成される混合情報源 (5) に対し，$n \to \infty$ かつ $L \to \infty$ の極限において，ペナルティパラメータを λ とした DP-means 法により決まるレートは，ひずみを λ とした最大ひずみ基準の混合情報源に対するレートひずみ曲線の値を達成する．

$$\lim_{L\to\infty}\lim_{n\to\infty} r(K) = \max\{R_1(\lambda), R_2(\lambda)\} \tag{16}$$

［証明 3］ 学習データが混合情報源から生成されているとき，p_1 と p_2 のどちらから生成されたかにより，データを二つの部分集合に分割する．そのそれぞれに対し，ペナルティパラメータ λ を用いた DP-means 法を適用したときのクラスタ数をそれぞれ K_1，K_2 とする．学習データ全体に対し同じ λ の値での DP-means 法の結果与えられるクラスタ数が K のときのレートの値は，

$$\frac{\ln K}{L} \leq \frac{\ln(K_1+K_2)}{L} \leq \frac{\ln(2\max\{K_1, K_2\})}{L}$$

より，L が十分大きいとき，$\max\{\ln K_1/L, \ln K_2/L\}$ で上から評価することができる．定理 2 より，この値は $n \to \infty$，$L \to \infty$ の極限で，

$$\max\{R_1(\lambda), R_2(\lambda)\},$$

となり，(16) の成立が確かめられる．(証明終)

定理 2，定理 3 より，単一の i.i.d. 情報源と混合情報源どちらの場合でも，$n \to \infty$ かつ $L \to \infty$ の極限においては，クラスタ数の曲線は最大ひずみ基準のレートひずみ曲線を達成する．

ただし，データの並びやクラスタ中心の初期値などの DP-means 法の解に関する仮定は DP-means 法の理想的な状況での振る舞いを調べるためのものであり，実際の状況では一般には成り立たない．これらの仮定の成り立つ条件を明らかにすることや，3.1 や本節での次元及びデータ数に関する極限操作の正当性を厳密に証明することは本論文の範疇を超える今後の課題であるため，次章において数値実験による検証を行う．

4. 数 値 実 験

本研究では，情報源として，N 回の試行を行ったときの二項分布を仮定する．対応するブレグマンダイバージェンスは (17) で与えられる．

$$d_\phi(x, \theta) = x\ln\frac{x}{\theta} + (N-x)\ln\frac{N-x}{N-\theta} \tag{17}$$

なお，凸関数 ϕ は (18) で与えられる．

$$\phi(\theta) = \theta \ln \frac{\theta}{N} + (N-\theta) \ln \frac{N-\theta}{N} \tag{18}$$

本章では，単一の二項分布から生成した乱数をデータとして与える場合と，混合情報源に対応した二つの二項分布の混合より生成した乱数をデータとして与える場合の 2 通りで実験を行った．生成するデータは L 次元として，単一の二項分布の場合は (19) においてパラメータが $\mu = 0.3, N = 100$ の下で，混合二項分布の場合は (19) においてパラメータが $\mu \in \{0.3, 0.7\}, N = 100$ の下で生成した．

$$p(x) = \binom{N}{x} \mu^x (1-\mu)^{N-x} \tag{19}$$

なお，混合二項分布の場合はデータ 1 点に対して，一様乱数から 50% の確率でパラメータ μ が決まるようにした．また，本章では，比較のためにレートひずみ曲線を頻繁に図示する．(6) 及び対称性から混合二項分布の場合のレートひずみ曲線は，単一の二項分布の場合と一致する．(17) の擬距離に対するレートひずみ曲線は，

$$D_{\max} = N \left\{ h(\mu) - \sum_{x=0}^{N} p(x) h\left(\frac{x}{N}\right) \right\}$$

$$R(0) = -\sum_{x=0}^{N} p(x) \ln p(x)$$

による端点をもつ．ここで，h は二値エントロピー関数である．これらの端点の間のレートひずみ曲線の値は，数値的に計算することができる [3], [14]．この方法は，あらかじめ指定した数の混合数をもつ有限混合分布の最適化を行う．混合数を一つずつ増やしながら最適化を行い，目的関数値の変化が十分小さくなった混合数を採用することで，レートひずみ曲線の計算を行った．収束判定条件の影響から多少の誤差が見られたが，本実験での比較には十分な精度といえる．

4.1 ペナルティパラメータに対するクラスタ数の曲線とレートひずみ曲線の関係

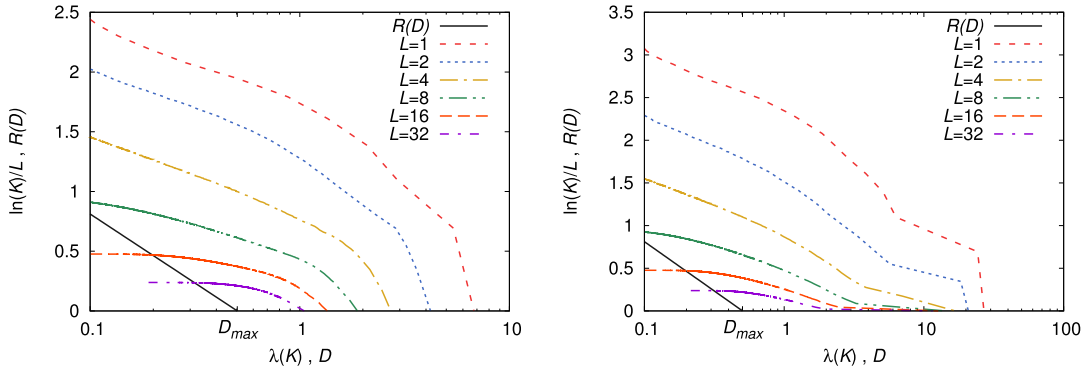
まず，学習データとして，単一の二項分布，混合二項分布のそれぞれより，データ 1 点の次元を $(2^0, ..., 2^5)$ と 2 のべき乗で変化させながら，各次元ごとに 2048 点からなる学習データセットを 100 セット生成した．同様に，単一の二項分布，混合二項分布より，データの次元を変化させながら，次元ごとに 16384 点からなるテストデータを 1 セット生成した．生成した学習データそれぞれに対し，ペナルティパラメータを (20) で変化させ，クラスタ数が 1 になるまで，DP-means 法を実行し，学習データに対応する次元のテストデータでテストを行った．

$$\lambda_i = \begin{cases} 0 & (i=1) \\ 0.01 & (i=2) \\ 1.01\lambda_{i-1} & (i \geq 3) \end{cases} \tag{20}$$

そして，実行ごとに，ペナルティパラメータ，クラスタ数，学習データとテストデータのそれぞれに対する平均ひずみと最大ひずみを記録した．

記録したデータそれぞれに対して，ペナルティパラメータの値を昇順に見たとき，クラスタ数の値が減少しており，なおかつ，同じクラスタ数の場合は，対応するペナルティパラメータの値が最小となるクラスタ数に対応するペナルティパラメータ，学習データとテストデータのそれぞれに対する平均ひずみと最大ひずみの値を次元ごとに対応するクラスタ数で平均をとったものを結果とした．この結果から，ペナルティパラメータとクラスタ数の関係をレートひずみ曲線とともに表したのが図 1，学習データに対する平均ひずみとクラスタ数の関係をレートひずみ曲線とともに表したのが図 2，テストデータに対する平均ひずみとクラスタ数の関係をレートひずみ曲線とともに表したのが図 3 である．なお，図 1〜図 3 では横軸を対数スケールとした．

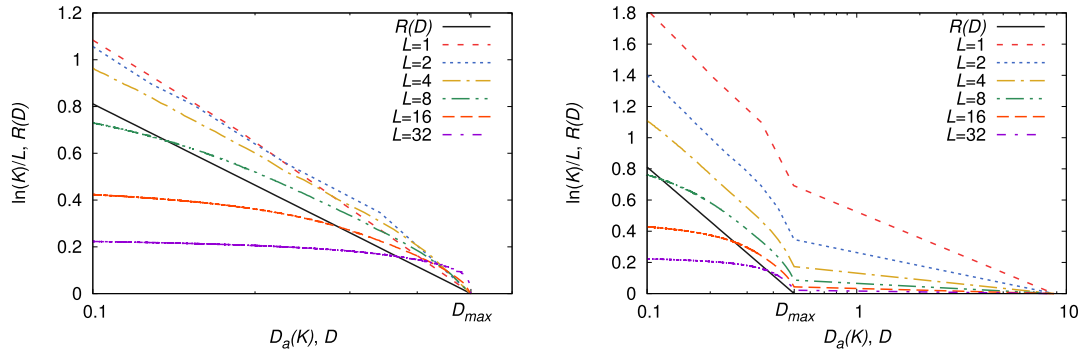
図 1 を見ると (a)，(b) ともに次元が上がると，ペナルティパラメータに対応する曲線は，レートが低い領域においてレートひずみ曲線に近づいていることがわかる．しかし，ひずみが 0 となる近辺においてデータの次元が 4 次元以上（図の範囲では 16 次元以上）では，クラスタ数の曲線がレートひずみ曲線より下にきている．これは本来，L 次元の二項分布のデータ空間は $(N+1)^L$ であるが，学習データ数は有限であり，二項分布のデータ空間と比べるとデータ数が少ないことが原因であると考えられる．実際，**3.1** での議論のように最大ひずみは $n \to \infty$ の極限においてレートひずみ曲線に近づくため，有限のデータ数では，曲線を下回ることがあり得る．図 2 を見ると，図 1 と同様に (a)，(b) ともに次元が上がると，平均ひずみに対応する曲線は下がり，レートひずみ曲線に近づいていることがわかる．4 次元以上（図の範囲では 8 次元以上）の場合でレートひずみ曲線の下にきていることも



(a) Single binomial source　　(b) Mixed binomial source

図 1　ペナルティパラメータに対するレートの曲線 (a) 単一の二項分布 (b) 混合二項分布

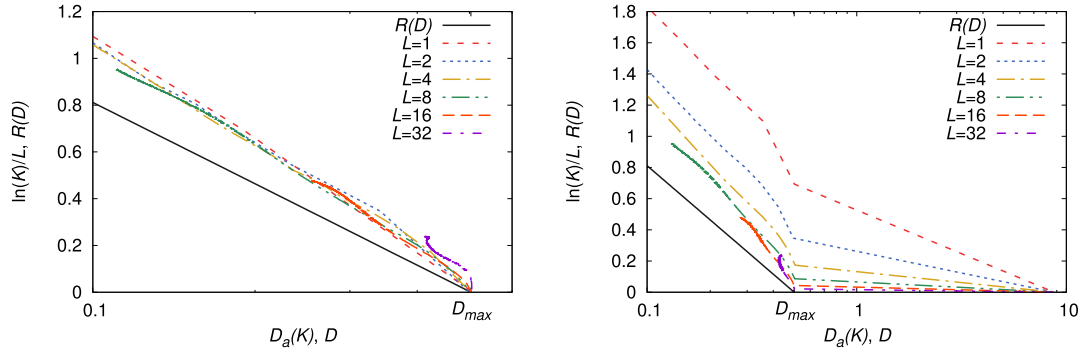
Fig. 1　Rate against the penalty parameters.



(a) Single binomial source　　(b) Mixed binomial source

図 2　学習データの平均ひずみに対するレートの曲線 (a) 単一の二項分布 (b) 混合二項分布

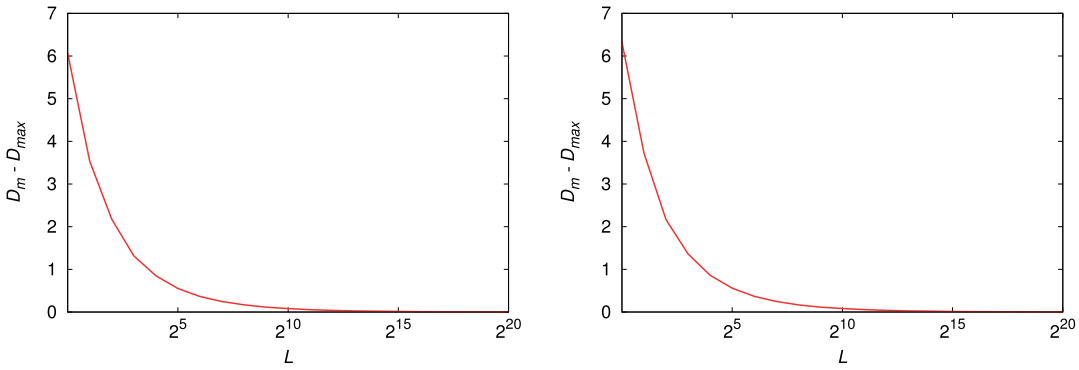
Fig. 2　Rate against the average distortion for training data.



(a) Single binomial source　　(b) Mixed binomial source

図 3　テストデータの平均ひずみに対するレートの曲線 (a) 単一の二項分布 (b) 混合二項分布

Fig. 3　Rate against the average distortion for test data.



(a) Maximum distortion at rate 0, for the single binomial source

(b) Maximum distortion at rate $\ln 2/L$, for the mixed binomial source

図 4　データの次元に対する最大ひずみと $D_{\max}$ との差 (a) 単一の二項分布において，レートの値が 0 となるときの最大ひずみ (b) 混合二項分布において，レートの値が $\ln 2/L$ となるときの最大ひずみ

Fig. 4　The difference between maximum distortion and $D_{\max}$ against the dimensionality of data.

図 1 と同じ理由だと考えられる．また，図 2 と図 3 を比較すると，次元が増大するとともに，学習データに対する平均ひずみとテストデータに対する平均ひずみの差異は増大し，レートひずみ曲線に対して学習データに対する平均ひずみは下に，テストデータに対する平均ひずみは上にきていることがわかる．混合情報源では，レートが $\ln 2/L$ の値で二つの混合成分が検出できるため，どの次元においても，そのレートでほぼ $D_{\max}$ 程度の平均ひずみの値となっている (図 2 (b), 図 3 (b))．

図 1 (a)，図 2 (a) において，レートが 0 になるときのひずみに着目すると，平均ひずみは $D_{\max}$ へ収束しているが，最大ひずみに相当するペナルティパラメータは，32 次元ではまだ収束しておらず，データの次元を更に上げて実験を行う必要がある．しかし，そのためには，データの次元を大きな値にし，なおかつデータ空間に空きが出ないような量のデータで実験を行う必要があり，現実的には困難である．図 1 (b), 図 2 (b) に関しても，レートが $\ln 2/L$ となるときの平均ひずみと最大ひずみに関して同様のことがいえる．

4.2　データの次元増加に伴う最大ひずみの収束

4.1 では，ペナルティパラメータに対応するクラスタ数の曲線が，次元を上げるとレートひずみ曲線に近づくことが示唆された．ここでは，次元が十分大きい場合に両者が限りなく近づくという様子を確認したい．しかし，データ数の影響を受けるため，データの次元を更に上げることは困難である．そこで，データ数の影響が少ないレートひずみ曲線のレートが 0 になる点 $D_{\max}$ に限定して実験を行った．まず，上限を 2^{20} 次元として，**4.1** と同様の条件で，学習データセットを生成する．次に，次元を 2 のべき乗で 2^{20} まで変化させながら，学習データセットごとに，データ点と真のクラスタ中心との最大ひずみを記録した．そして，単一の二項分布と混合二項分布に対しそれぞれ，次元ごとに平均を取った値を結果とした．この結果を表したのが図 4 である．図 4 を見ると，(a), (b) ともに次元が上がると，最大ひずみと $D_{\max}$ の差が 0 に限りなく近づいていくことがわかる．このことから，データの次元を大きくする極限において，ペナルティパラメータに対応するクラスタ数の曲線がレートひずみ曲線に収束することが示唆された．

4.3　最大ひずみの混合二項分布に対するレートひずみ曲線への収束

定理 3 では，学習データが混合情報源から生成されているとき，クラスタ数の曲線はデータ数と次元数が無限大の極限で，(6) のレートひずみ曲線を達成することが示された．すなわち，あるひずみ D に対して，混合情報源を構成する i.i.d. 情報源に対応するレートが最大となる曲線である．ここでは，数値実験でその確認を行う．実験はデータ数の影響を避けるため，**4.2** と同様の条件で行う．混合情報源を構成する i.i.d. 情報源に対するレートひずみ曲線に差を付けるために，単一の二項分布の場合は $\mu = 0.01$ と $\mu = 0.5$, 混合二項分布の場合は $\mu \in \{0.01, 0.5\}$ からなる情報

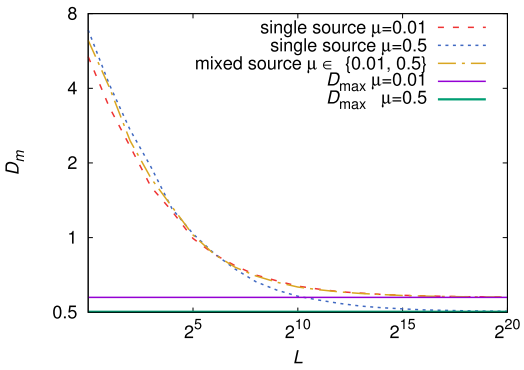


図 5　パラメータ $\mu \in \{0.01, 0.5\}$ からなる混合二項分布におけるデータの次元に対する最大ひずみと $D_{\max}$

Fig. 5　Maximum distortion against the dimensionality of data and $D_{\max}$ for the mixed binomial source with $\mu \in \{0.01, 0.5\}$.

源よりデータの生成を行った．実験の結果を図 5 に示す．図 5 では，単一の二項分布の場合はそれぞれ，生成元の情報源に対応する $D_{\max}$ に近づいていることがわかる．そして，混合二項分布の場合には，レートが $\ln 2/L$ となるときの最大ひずみが二項分布のパラメータが $\mu = 0.01$ であるときの $D_{\max}$ に近づいていることがわかる．つまり，学習データが混合二項分布より生成されているとき，データ数と次元数が無限大の極限で，混合二項分布を構成する二項分布に対応するレートが大きい方の曲線に近づくことが確認できる．

4.4　ペナルティパラメータと最大ひずみの関係

4.1 及び **4.2** では，ペナルティパラメータと最大ひずみは近似的に等しいという前提において，データの次元を十分に大きくした場合，ペナルティパラメータに対するクラスタ数の曲線がレートひずみ曲線に近づくことを示した．しかし，**4.1** でペナルティパラメータに対応するクラスタ数の曲線と学習データに対する最大ひずみに対応するクラスタ数の曲線を比較すると，多少の差異があるとわかった．このことについては，次に挙げる二つの理由が考えられる．一つ目の理由としては，データをクラスタに割り当てるループにおいて新しいクラスタが複数作られることがあるため，余分なクラスタが生成される可能性があることである．二つ目の理由としては，クラスタ中心の計算は全てのデータ点をクラスタに割り当てた後に行うため，クラスタ中心が大雑把にしか動かないことである．そこで，クラスタが増えるのは，データをクラスタに割り当てるループ中につき 1 回，クラスタ中心はデータ 1 点の

Algorithm 2 modified DP-means for maximum distortion

Input: $\boldsymbol{x}^n = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}, \lambda$
Output: $\boldsymbol{l} = \{\boldsymbol{l}_1, ..., \boldsymbol{l}_K\}, K$

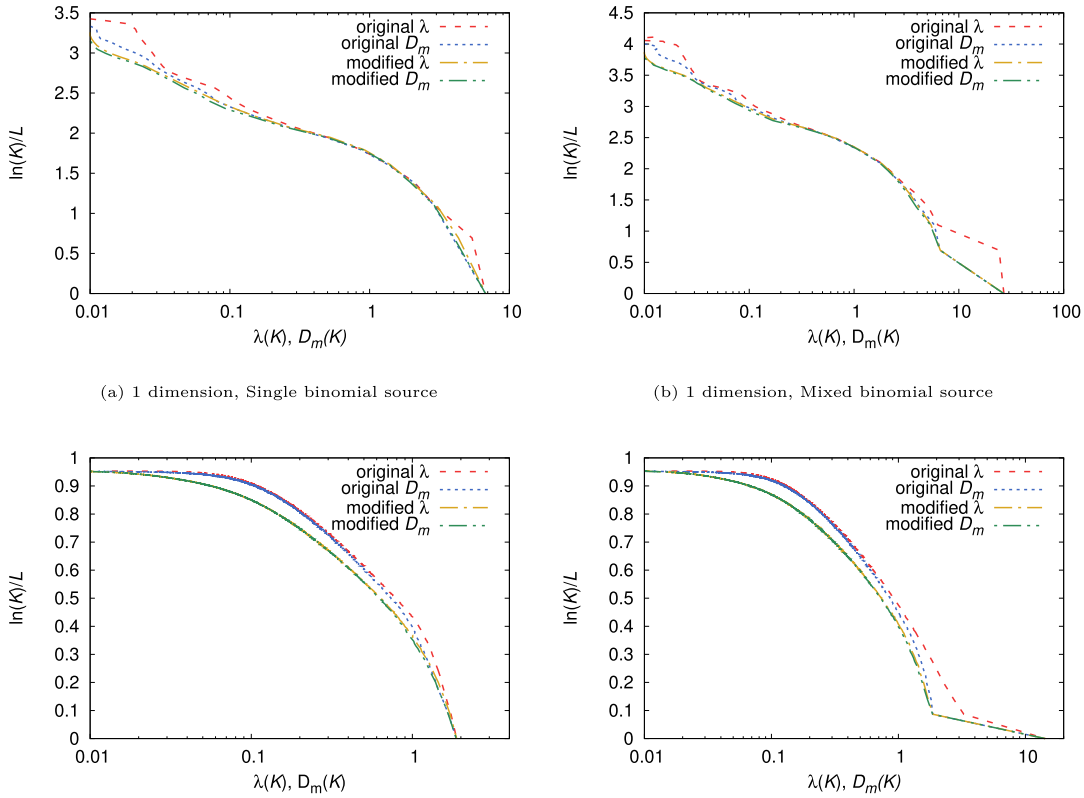
```
K = 1
l_1 = x^n
θ_1 = (1/n) Σ_{i=1}^n x_i
c(i) = 1 (i = 1, ..., n)
repeat
  once = true
  for i = 1 to n do
    d_ik = d_L(x_i, θ_k) (k = 1, ..., K)
    if min_k d_ik > λ && once then
      K = K + 1
      c(i) = K
      θ_K = x_i
      once = false
    else
      c(i) = arg min_k d_ik
    end if
    for j = 1 to K do
      l_j = {x_i | c(i) = j}
      θ_j = (1/|l_j|) Σ_{x∈l_j} x
    end for
  end for
until Σ_{i=1}^n d_L(x_i, θ_{c(i)}) + λK converges
```

割り当てを行うごとに計算するように変更したアルゴリズムが Algorithm 2 である．この Algorithm 2 を用いて **4.1** と同様の実験を行い，ペナルティパラメータと学習データに対する最大ひずみの差異の大きさについて比較を行った．ただし，次元は $(2^0, ..., 2^3)$ の範囲とした．結果から，1 次元と 8 次元を例として，オリジナルの DP-means 法による実行結果と改変した DP-means 法による実行結果から，それぞれ，ペナルティパラメータとクラスタ数の関係と学習データの最大ひずみとクラスタ数の関係を図 6 に示す．図 6 から次元及び二項分布が単一か混合かによらず，改変した DP-means 法による結果では，ペナルティパラメータに対応するクラスタ数の曲線と学習データに対する最大ひずみに対応するクラスタ数の曲線の特にひずみが大きいときの差異が小さくなっていることがわかる．

5.　考　　察

5.1　データの順序に対する依存性

ペナルティパラメータの値を増加させたとき，クラスタ数の値は基本的に減少するが，稀に上昇することがある．これは，DP-means 法がデータの並びに依存して結果が局所解となるためであり，データの並びを変更することで，ペナルティパラメータに対するクラ



(a) 1 dimension, Single binomial source

(b) 1 dimension, Mixed binomial source

(c) 8 dimension, Single binomial source

(d) 8 dimension, Mixed binomial source

図 6　アルゴリズムの改変によるペナルティパラメータと学習データの最大ひずみの差異の比較 (a) 1 次元，単一の二項分布 (b) 1 次元，混合二項分布 (c) 8 次元，単一の二項分布 (d) 8 次元，混合二項分布

Fig. 6　Comparison of the penalty parameter of the modified DP-means to the maximum distortion for training data.

スタ数は単調に減少すると考えられる．また，データの並びを変更して，クラスタ数の値が減少したとき，最適にクラスタリングされたと考えると，最大ひずみの値はペナルティパラメータにより近づくと予想できる．実際に，ペナルティパラメータ 1 点ごとにデータの並びをランダムに 1000 回変更し，その中で，クラスタ数が最小となり最大ひずみが最小となる値をペナルティパラメータごとに記録した．結果としては，ペナルティパラメータの増加に伴いクラスタ数は単調に減少した．一方で，最大ひずみはペナルティパラメータの値によっては，データの並びを変更する前よりも両者の差異は増大した．これは，DP-means 法は平均ひずみ最小化を行っているが，そのときに必ずしも最大ひずみが最小となるわけではないためだと考えられる．また，ペナルティパラメータに対して最大ひずみが最小となるデータの並びを調べようとすると，データ数の階乗通りを調べる必要があり，現実的な計算時間では不可能である．ペナルティパラメータと最大ひずみの間の若干の差異は，**4. 4** の結果から，データの順序に対する依存性よりも，1 ループ中に複数のクラスタが生成されることやクラスタ中心の更新があまり頻繁に行われないことが原因であると考えられる．**3. 2** では，データの並びやクラスタ中心の初期値などの DP-means 法の解に関する理想的な仮定の下，ペナルティパラメータと最大ひずみが一致することを示したが，**4. 4** の数値実験の結果は仮定の成り立たない実際の状況においても頻繁に一致することを示唆している．

5. 2　学習データ数に対する依存性

また，**4.** では，データ数とデータの次元が大きい

場合にペナルティパラメータに対応するクラスタ数の曲線はレートひずみ曲線に近づくことを示した．一方で，有限の次元数において，学習データ数が大きいときの最大ひずみの振る舞いは極値統計論により考察することができる．ペナルティパラメータの最大ひずみとしての解釈から，ペナルティパラメータに対するDP-means 法のクラスタ数変化の学習データ数への依存性を調べることができる．

簡単のため，各次元が独立に平均 0 分散 σ^2 の正規分布である L 次元等方的正規分布に従うデータを考える．2 乗距離をひずみ尺度とすると，クラスタ数が 1 となるペナルティパラメータの下限は，近似的に，

$$\lambda(1) \simeq \max_{1\leq i\leq n} d_L\left(\boldsymbol{x}_i, \boldsymbol{0}\right) = \frac{1}{L}\max_{1\leq i\leq n} \|\boldsymbol{x}_i\|^2$$

で与えられる．極値統計論より，データ数 n が十分大きいとき，

$$\max_{1\leq i\leq n} \|\boldsymbol{x}_i\|^2 \simeq 2\sigma^2 \ln n + O(\ln\ln n)$$

であることが示される [15]．このことから，主要項のみを考えると，

$$\lambda(1) \simeq \frac{2\sigma^2}{L}\ln n$$

すなわち，ペナルティパラメータはデータ数の対数のオーダーで大きくなることがわかる．データ数 n に対し，クラスタ数 K が十分小さい場合に，各クラスタに十分なデータが含まれると仮定すると，同様に $\lambda(K)$ も同じ主要項をもつことがわかる．このことは人工データを用いてデータ数 n を変えた数値実験により確かめることができる [16]．また，**4.** で用いた二項分布（及びその混合）においても，N が大きいとき，中心極限定理から同様の考察が成り立ち，ペナルティパラメータはデータ数 n に対し $\ln n$ のオーダーでスケールすることを実験的に確かめることができる．

6. むすび

本研究では，DP-means 法におけるペナルティパラメータはレートひずみ理論における最大ひずみと同様の意味をもつことを明らかにした．これにより，最大ひずみ基準のレートひずみ理論に対応して，データ数とデータの次元数を上げたとき，クラスタ数の曲線は最大ひずみ基準のレートひずみ曲線に近づくことを示した．また，データ数が有限であるため，**4.1** の実験方法では，ひずみが小さい領域では，この検証を行うことは困難であることがわかった．ひずみが大きい部分においては，DP-means 法のクラスタ数の曲線と最大ひずみは，少し差異があるが，アルゴリズムを改変することで，この差異を小さくすることができることを確認した．本研究の結果から，次元が十分に大きいデータに対して DP-means 法を行う場合，データの分布に対応する最大ひずみ基準のレートひずみ曲線からペナルティパラメータを指定することが有効であると考えられる．実際のパラメータ設定法の構成には，レートひずみ曲線の有限データ数での解析 [10] 及び有限次元での解析 [17] を組み合わせることが重要であると考えられる．

謝辞 本研究の一部は科学研究費助成事業 25120014, 15K16050, 16H02825 の助成を受けた．

文　献

[1] 持橋大地，“最近のベイズ理論の進展と応用 (III)—ノンパラメトリックベイズ，” 信学誌，vol.93, no.1, pp.73–79, Jan. 2010.

[2] B. Kulis and M.I. Jordan, “Revisiting k-means: New algorithms via Bayesian nonparametrics,” Proc. International Conference on Machine Learning, pp.513–520, 2012.

[3] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” J. Machine Learning Research, pp.1705–1749, 2005.

[4] K. Jiang, B. Kulis, and M.I. Jordan, “Small-variance asymptotics for exponential family Dirichlet process mixture models,” Advances in Neural Information Processing Systems, pp.3158–3166, 2012.

[5] X. Pan, J. Gonzalez, S. Jegelka, T. Broderick, and M.I. Jordan, “Optimistic concurrency control for distributed unsupervised learning,” Advances in Neural Information Processing Systems, pp.1403–1411, 2013.

[6] O. Bachem, M. Lucic, and A. Krause, “Coresets for nonparametric estimation – the case of DP-means,” Proc. International Conference on Machine Learning, pp.209–217, 2015.

[7] D. Bruno, S. Calinon, and D.G. Caldwell, “Learning autonomous behaviours for the body of a flexible surgical robot,” Autonomous Robots, vol.41, no.2, pp.333–347, 2017. 10.1007/s10514-016-9544-6

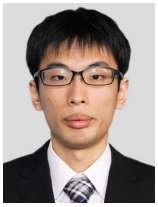
[8] M. Comiter, M. Cha, H.T. Kung, and S. Teerapittayanon, “Lambda means clustering: automatic parameter search and distributed computing implementation,” Proc. International Conference on Pattern Recognition, 2016.

[9] 韓　太舜，情報理論における情報スペクトル的方法，培風館，1998.

[10] T. Linder, “On the training distortion of vector

quantizers," IEEE Trans. Inf. Theory, vol.46, no.4, pp.1617–1623, 2000.
[11] R.M. Gray and D.L. Neuhoff, "Quantization," IEEE Trans. Inf. Theory, vol.44, no.6, pp.2325–2383, 1998.
[12] T. Berger, Rate Distortion Theory: A Mathematical Basis for Data Compression, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[13] Ya.G. シナイ，シナイ確率論入門コース，丸善，2016.
[14] A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu, "An information theoretic analysis of maximum likelihood mixture estimation for exponential families," Proc. International Conference on Machine Learning, pp.57–64, 2004.
[15] M.R. Leadbetter, G. Lindgren, and H. Rootzén, Extremes and Related Properties of Random Sequences and Processes, Springer, 1983.
[16] 上遠野貴広，"ディリクレ過程平均法におけるクラスタ数推定法と推定精度の解析，" 電子情報通信学会東海支部卒業研究発表会予稿集，2015.
[17] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," IEEE Trans. Inf. Theory, vol.58, no.6, pp.3309–3338, 2012.

（平成 29 年 6 月 19 日受付）



小林真佐大　（学生員）

平 28 豊橋技科大工学部情報・知能工学課程卒業．現在，同大大学院博士前期課程在学中．統計的機械学習に関する研究に従事．

渡辺　一帆　（正員）

平 18 東工大大学院総合理工学研究科知能システム科学専攻了．博士（工学）．日本学術振興会特別研究員，東大大学院学術研究支援員・特任助教，奈良先端大助教を経て現在，豊橋技科大情報・知能工学系講師．統計的機械学習に関する研究に従事．