ベイズ法を用いたノイズに頑健な高次元データの推定

Bayesian Robust Clustering for High Dimensional Data

星　野　　力

要　約　大量なデータを中心とするシステムアーキテクチャはビックデータと呼ばれ，現在大きな注目を集めている．データ解析において，大部分の実データには不確実な要素が含まれ，そのモデリング手段として確率統計的な処理が広く用いられている．データに比較的単純な分布（指数分布族）が仮定できる場合には，十分な基盤が整備されているが，隠れた階層や構造を持つ分布に対しては，近年になってベイズ統計の有用性が理論的にも実験的にも示されている．本論文では，典型的な階層モデリングである高次元のクラスタリングを取りあげ，データに対する仮説を確率モデルで表現する手段と，ベイズ法を用いたデータからの推定アルゴリズムを紹介する．実データでの適用実験を行った結果，従来法と同等もしくは，優越することがわかった．

Abstract　In many situations, the obtained data have essential uncertainty caused by many random effects under generative and observational processes of data analysis. A probabilistic approach is a main tool for uncertainty and successfully applied in many fields. Recently, for the distributional hypothesis that is beyond the simple distribution (such as exponential family) and that has hierarchical structure with hidden variables, the advantage of Bayesian method was shown by experimental and theoretical studies. In this report, we investigate high dimensional clustering, which is a typical application of hierarchical modeling. We introduce the method of translating data generation hypothesis into statistical models and derived Bayesian inference algorithm by Markov Chain Monte Carlo. In the real data experiments, the proposed method outperforms conventional methods in some data sets.

1. は　じ　め　に

大量なデータの流れをモデリングする，データ中心のシステムアーキテクチャはビックデータと呼ばれ，現在注目を浴びている．これからの情報産業にとって，データを蓄積，収集，分析するフレームワークを確固とした基盤として確立し，さらに，その基盤を用いて個別問題へと適用していく中で，様々な知見を蓄積し，深化させていくことが急務になっていることは疑いない．また，この方向性は短期的なものではなく，定型作業の自動化と並んで計算機活用の柱と考えられている知的処理（特に脳を補完もしくは代替する機能）が，センサー，ネットワーク，記憶媒体，計算速度，アルゴリズム等の発達をもとに，ある種の質的転換をおこしつつある中から自然に生じているものと考えられる．

実際のデータに関しても，センサーやメディアの多様化にともなって，従来からの数値のほか，テキストや音声，画像，動画，DNA や蛋白質の構造，ソーシャルグラフ等，これまでのデータ処理では扱ったことのなかったものが対象になっている．これらのデータは，人為的に作り出した確定したデータでないため，観測できない変数の影響や，測定機器の誤差などさまざま

な条件からデータにゆらぎが生じて，本質的に不確実な性質を持つ．不確実性に対処する手法としては，従来から確率統計を用いた手法が探求されてきた．特にデータの分布に指数分布族（正規分布や多項分布等）が仮定できる場合は，基礎理論が構築され大きな成果をあげてきた．

ところが，テキストや画像，蛋白質などのデータは，高次元であり，かつ内部に文法などの複雑な構造を持っていたり，測定自体の難しさからくる外れ値を含むものが多い．そのため，単純な分布族ではデータを十分に表現しきれず，階層性や文法などの構造を持つ確率モデルを仮定する必要がある．さらに，確率モデルを設定した後にはデータをもとに分布を推定することになる．分布の推定には様々な手法があるが，その中でベイズ推定が，特に階層性など隠れた構造を持つ分布の推定に対して，精度およびモデル自体の妥当性の検証の両面において理論的に優れていることが，近年示されつつある[1]．それらをふまえると，今後のデータ解析の基盤の一つとして，対象データの構造をうまく表現する確率モデルの設計手法と，ベイズ統計を用いた効率的な分布の推定を軸として考えることができる．

2. 仮説と確率モデル

2.1 生成モデル

データの性質に関する仮説を立て，解析者が行いたいタスクを反映するよう，仮説を表現する確率モデル（尤度関数やエネルギー関数）を，生成モデルと呼ぶ．例えば，テキストの解析において，生のデータは単なる文字列で与えられるが，我々はその背後に，形態素や文節単位としての区切りや，文節間の依存関係があることを知っている．一方で，与えられる文字列には，単語の区切りや，文節の関係のデータが付与されていないので，そこになんらかの構造を仮定することになる．そのような場合，生成モデルを用いたアプローチでは，見えないラベルに対応する構造を隠れ変数として導入する．テキストの例では，文字列の切れめや，文節間の依存関係が隠れた変数としてモデルに取り込まれる．実問題においても，隠れマルコフモデルや，確率文脈自由文法と呼ばれるモデルが考案され，音声認識や，バイオインフォマティクスなどで大きな成果をあげている．生成モデルは，陽に認識しない場合にも，分布関数を指定したことで導入されている．例えば平均と分散だけを使って推論や判断を行う場合は正規分布を暗に仮定している．生成モデルを用いた手法はその仮定のプロセスを前面に出して積極的に複雑な構造をモデリングするアプローチである．

2.2 分布の推測とモデル選択

仮説として生成モデルを設計した後には，データを用いたパラメータの推定や，複数の仮説である生成モデル間でどのモデルが最もデータを説明できるか比較するモデル選択の問題に取り組むことになる．パラメータの推定には，最尤推定や MAP 推定，ベイズ推定など様々な手法があるが，その中で仮説モデルに階層性が含まれる場合には，ベイズ推定を用いると，パラメータの推定精度，数学的に妥当なモデル選択の両面から他の手法を優越することが示されている．特に，汎用的な事後分布の推論アルゴリズムである Markov Chain Monte Carlo 法と，モデル選択基準である WAIC を併用することにより，理論的な基盤のもと分布推測とモデル選択が実行可能となった[1]．

2.3　高次元データのクラスタリング

データに対する仮説を生成モデルで書き下す手法と，分布推測のアルゴリズムを導出する具体的な対象として，データのクラスタリングを選択する．クラスタリングは与えられたラベルなしデータから，データの分布を反映して各データを自動的に分類しラベリングする技術である．観測されないラベルを隠れ変数として表現する混合モデルを対象に考えると，クラスタリングは典型的な階層モデリングの例となる．

本稿の目的を高次元でかつノイズに頑健なデータのクラスタリングに対する確率モデルの設計とする．そこで，モデリングのもとになるデータの生成に以下の三つの仮定を置く．

1. 各データはそれぞれ一つのクラスタに所属する．
2. データ自体は高次元であるが，共分散を考慮すると，本質的にはクラスタごとに異なる低次元の空間に圧縮され分布している．
3. 各クラスタは，クラスタ内の共分散を反映した楕円上に分布するが，複数の外れ値を持つ．

これらの仮定の，確率モデルでの表現の一つとして，

1. データの分布は混合分布で表現する．
2. 各クラスタの共分散構造は，因子分析を使った低次元で表現する．
3. 各クラスタは，異なる自由度を持つ，Student-t 分布で表現する．

がある．仮定 1 の混合分布のクラスタラベルと，仮定 3 の Student-t 分布のデータごとの分散の広がりは観測データから隠れていることに注意する．また，混合分布の仮定から，この分布は指数分布族では表現できず，特異モデルのクラスに所属する[1]．

3.　定式化

3.1　モデルと事前分布

2 章の確率モデルに対する仮定は下記の尤度関数で表現される．

$$
\begin{aligned}
&\sum_{k=1}^{K}\int_0^{\infty}\int_{-\infty}^{\infty} a_k \mathcal{N}(\mathbf{y}\mid \mathbf{W}_k\mathbf{x}+\boldsymbol{\mu}_k, (u\boldsymbol{\Lambda}_k)^{-1})\mathcal{N}(\mathbf{x}\mid 0, (u\mathbf{I}_M)^{-1})\, d\mathbf{x}\mathcal{G}\left(u\mid \frac{\nu_k}{2}, \frac{\nu_k}{2}\right) du \\
&= \sum_{k=1}^{K}\int_0^{\infty} a_k \mathcal{N}\left(\mathbf{y}\mid \boldsymbol{\mu}_k, \frac{1}{u}(\boldsymbol{\Lambda}_k^{-1}+\mathbf{W}_k\mathbf{W}_k^T)\right)\mathcal{G}\left(u\mid \frac{\nu_k}{2}, \frac{\nu_k}{2}\right) du \\
&= \sum_{k=1}^{K} a_k S\,(\mathbf{y}\mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}+\mathbf{W}_k\mathbf{W}_k^T, \nu_k)
\end{aligned}
$$

ただし，$a_k \geq 0, \sum_{k=1}^{K} a_k = 1$ をみたし，Λ_k は対角行列である．S，$\mathcal{N}$，$\mathcal{G}$ はそれぞれ Student-t 分布，Gauss 分布，ガンマ分布を表わし，その密度関数は

$$S(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}, \nu) = \frac{\Gamma\left(\frac{d+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{d}{2}}} |\boldsymbol{\Lambda}|^{\frac{1}{2}} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{d+\nu}{2}}$$

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\mathcal{G}(u \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} u^{\alpha-1} \exp(-\beta u) \quad (u > 0)$$

と書ける．また，分布の推定にベイズ統計を用いるため，パラメータの事前分布を設定する．パラメータ $\theta \equiv (\mathbf{a}, \boldsymbol{\Lambda}_k, \mathbf{W}_k, \nu_k)$ に対する事前分布としては共役事前分布を仮定し，$\mathbf{W}_k$ の事前分布の分散には階層事前分布を仮定する．

$$p(\mathbf{a}) = \mathcal{D}(\mathbf{a} \mid \phi_0)$$
$$p(\lambda_{kj}) = \mathcal{G}(\lambda_{kj} \mid a_0, b_0)$$
$$p(\mathbf{w}_{kj.} \mid \boldsymbol{\alpha}_k) = \mathcal{N}(\mathbf{w}_{kj.} \mid 0, \lambda_{kj}\, \mathrm{diag}(\boldsymbol{\alpha}_k))$$
$$p(\alpha_{kl}) = \mathcal{G}(\alpha_{kl} \mid c_0, d_0)$$

ただし，$\mathbf{w}_{kj.}$ は $\mathbf{W}_k$ の j 行を表わし，$\mathcal{D}$ はディリクレ分布

$$\mathcal{D}(\mathbf{a} \mid \phi_0) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} a_k^{\phi_0 - 1}$$

である．t 分布の自由度 ν_k には，数値計算上の安定性を考慮して，$\nu_k \geq 1$ の制約を入れた事前分布を用いる．

$$p(\nu_k \mid \eta_0) = \eta_0 \exp(-\eta_0(\nu_k - 1))\, I_{[1,\infty]}(\nu_k)$$

ただし，$I_{[a,b]}(x)$ は x のサポートが区間 $[a, b]$ であることを示す．

3. 2　分布推定のアルゴリズム

これらの仮定のもと，N 点の観測データ $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ が与えられたとき，隠れ変数までも含めた完全データ $CD^N \equiv ((\mathbf{y}_1, \mathbf{x}_1, u_1, \mathbf{z}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N, u_N, \mathbf{z}_N))$（ただし，$\mathbf{y}_i \in R^D$, $\mathbf{x}_i \in R^M$, $u_i \in R$, $\mathbf{z}_i \in \{1, 0\}^K$, $\sum_{k=1}^{K} z_{ik} = 1$ をみたす）の完全対数尤度は以下の式で与えられる．

$$\prod_{i=1}^{N} p(\mathbf{y}_i, \mathbf{x}_i, u_i, \mathbf{z}_i \mid \mathbf{W}, \boldsymbol{\Lambda}, \mu) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left\{\mathcal{N}(\mathbf{y}_i \mid \mathbf{W}_k\mathbf{x}_i + \mu_k, (u_i\boldsymbol{\Lambda}_k)^{-1}) \mathcal{N}(\mathbf{x}_i \mid 0, (u_i\mathbf{I}_M)^{-1}) \mathcal{G}(u_i \mid \frac{\nu_k}{2}, \frac{\nu_k}{2})\right\}^{z_{ik}}$$

このモデルのもとでは事後分布を解析的に得ることができないので，何らかの近似が必要となるが，本稿では，Markov Chain Monte Carlo（MCMC）法の一種である Gibbs Sampler を用いて事後分布の近似を行う．そのとき，Gibbs Sampler に必要な条件付き分布は以下のように構成される．

始めに，隠れ変数の条件付き分布からのサンプリングを考える．まず，クラスタの選択変数 $\mathbf{z}_i$ の条件付き事後分布は，$(\mathbf{x}_i, u_i)$ に関して周辺化を行い

$$p(\mathbf{z}_i \mid \mathbf{y}_i, \theta) = \frac{a_k S(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1} + \mathbf{W}_k \mathbf{W}_k^T, \nu_k)}{\sum_{k=1}^{K} a_k S(\mathbf{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1} + \mathbf{W}_k \mathbf{W}_k^T, \nu_k)}$$

と表わすことができる．

次に，サンプルごとの分散の広がり u_i は，$\mathbf{z}_i$ が与えられたもと，

$$\begin{aligned} p(u_i \mid \mathbf{y}_i, \mathbf{z}_i, \theta) &= \mathcal{G}(u_i \mid e, f) \\ e &= \frac{D + \nu_k}{2} \\ f &= \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{\Lambda}_k^{-1} + \mathbf{W}_k \mathbf{W}_k^T)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) + \frac{\nu_k}{2} \end{aligned}$$

と書ける．

次に，圧縮された空間での座標 x_i は，$(\mathbf{z}_i, u_i)$ が与えられたもと，

$$\begin{aligned} p(\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{z}_i, u_i, \theta) &= \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_{x_i}, \boldsymbol{\Sigma}_{x_i}) \\ \boldsymbol{\mu}_{x_i} &= (\mathbf{I}_M + \mathbf{W}_k^T \boldsymbol{\Lambda}_k \mathbf{W}_k)^{-1} \mathbf{W}_k^T \boldsymbol{\Lambda}_k (y_i - \boldsymbol{\mu}_k) \\ \boldsymbol{\Sigma}_{x_i} &= \frac{1}{u_i} (\mathbf{I}_M + \mathbf{W}_k^T \boldsymbol{\Lambda}_k \mathbf{W}_k)^{-1} \end{aligned}$$

と表わすことができる．これらの条件付き分布から隠れ変数をサンプリングすることで，各サンプルに対する完全データが得られる．

次に，上で得られた完全データを使って，パラメータの条件付き分布を考える．準備として完全データを使った十分統計量を定義しておく．

$$\begin{aligned} \left\langle n \right\rangle_k &\equiv \sum_{i=1}^{N} z_{ik} \\ \left\langle \mathbf{yy} \right\rangle_k &\equiv \sum_{i=1}^{N} z_{ik} u_{ik} (\mathbf{y}_i \circ \mathbf{y}_i^T) \\ \left\langle \mathbf{yx} \right\rangle_k &\equiv \sum_{i=1}^{N} z_{ik} u_{ik} (\mathbf{y}_i \mathbf{x}_i) \\ \left\langle \mathbf{xx} \right\rangle_k &\equiv \sum_{i=1}^{N} z_{ik} u_{ik} (\mathbf{x}_i \mathbf{x}_i^T) \\ \left\langle u \right\rangle_k &\equiv \sum_{i=1}^{N} z_{ik} (u_{ik} - \log u_{ik}) \end{aligned}$$

ただし，$\circ$ は要素ごとの積をあらわす．混合パラメータ $\mathbf{a}$ の条件付分布は，

$$\begin{aligned} p(\mathbf{a} \mid CD^N, \phi_0) &= \mathcal{D}(\mathbf{a} \mid \phi_1, \ldots, \phi_K) \\ \phi_k &= \left\langle n \right\rangle_k + \phi_0. \end{aligned}$$

分散パラメータ λ_{kj} の条件付分布は，

$$
\begin{aligned}
&p(\lambda_{kj} \mid CD^N, \boldsymbol{\alpha}_k) = \mathcal{G}(\lambda_{kj} \mid a, b) \\
&a = \frac{1}{2}\left\langle n \right\rangle_k + a_0 \\
&b = \frac{1}{2}\left\{ \left\langle yy \right\rangle_{kj} - \left\langle \mathbf{yx} \right\rangle_{kj}^T (\mathrm{diag}(\boldsymbol{\alpha}_k) + \left\langle \mathbf{xx} \right\rangle_k)^{-1} \left\langle \mathbf{yx} \right\rangle_{kj} \right\} + b_0
\end{aligned}
$$

となり，重みパラメータ $\mathbf{w}_{kj}$ の条件付分布は，

$$
\begin{aligned}
&p(\mathbf{w}_{kj} \mid CD^N, \boldsymbol{\alpha}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\mathbf{w}_{kj} \mid \boldsymbol{\mu}_{w_{kj}}, \boldsymbol{\Sigma}_{w_{kj}}) \\
&\boldsymbol{\mu}_{w_{kj}} = (\mathrm{diag}(\boldsymbol{\alpha}_k) + \left\langle \mathbf{xx} \right\rangle_k)^{-1} \left\langle \mathbf{yx} \right\rangle_{kj} \\
&\boldsymbol{\Sigma}_{w_{kj}} = \frac{1}{\lambda_{kj}} (\mathrm{diag}(\boldsymbol{\alpha}_k) + \left\langle \mathbf{xx} \right\rangle_k)^{-1}
\end{aligned}
$$

となる．重みの階層事前分布 $\boldsymbol{\alpha}_k$ の条件付分布は，

$$
\begin{aligned}
&p(\alpha_{kl} \mid \mathbf{W}_k) = \mathcal{G}(\alpha_{kl} \mid c, d) \\
&c = \frac{D}{2} + c_0 \\
&d = \frac{1}{2}\sum_{j=1}^{D} w_{kjl} + d_0
\end{aligned}
$$

となる．t 分布の自由度 ν_k の条件付き分布は，

$$
\begin{aligned}
&p(\nu_k \mid CD^N, \eta_0) \propto \left(\frac{\nu_k}{2}\right)^{\frac{\left\langle n \right\rangle_k \nu_k}{2}} \Gamma\left(\frac{\nu_k}{2}\right)^{-\left\langle n \right\rangle_k} \exp(-\phi\nu_k)\, \mathcal{I}_{[1,\infty]}(\nu_k) \\
&\phi = \frac{1}{2}\left\langle u \right\rangle_k + \eta_0
\end{aligned}
$$

となるが，この分布からは簡単にサンプルが取れないので，文献[3]の手法を利用して，棄却サンプリングを用いる．提案分布 g を

$$
g(\nu_k; \delta) = \delta \exp(-\delta(\nu_k - 1))\, \mathcal{I}_{[1,\infty]}(\nu_k)
$$

とし，パラメータ δ については，条件付き分布と一次のモーメントを合わせる条件である，

$$
\frac{\left\langle n \right\rangle_k}{2}\left(\log\left(\frac{1+\delta}{2\delta}\right) + 1 - \boldsymbol{\Psi}\left(\frac{1+\delta}{2\delta}\right)\right) + \delta - \phi = 0
$$

を解いて決める（$\boldsymbol{\Psi}$ は，*digamma* 関数）．その時，棄却率は

$$
p = \left(\frac{\Gamma\left(\frac{1+\delta}{2\delta}\right)}{\Gamma\left(\frac{\nu_k}{2}\right)}\right)^{\left\langle n \right\rangle_k} \left(\frac{\nu_k}{2}\right)^{\frac{\left\langle n \right\rangle_k \nu_k}{2}} \left(\frac{1+\delta}{2\delta}\right)^{\frac{-\left\langle n \right\rangle_k (1+\delta)}{2\delta}} \exp\left((\nu_k - 1)(\delta - \phi) + \frac{\phi}{\delta} - 1\right)
$$

となる．これら，隠れ変数とパラメータからのサンプリングを 1 ステップとして繰り返し計算を行う．

4. 数値実験

4.1 データセット

モデルの妥当性を考察するために，実データを用いた実験を行った．データは UCI Machine Learning Repository（http://archive.ics.uci.edu/ml/）から入手した Iris Data Set と Wine Data Set を使用した．それぞれのデータのサンプル数，次元，クラスタ数は表 1 のとおりである．

表 1　データセット

	サンプル数	次元	クラスタ数
Iris	150	4	3
Wine	178	13	3

4.2 実験条件

実験は，クラスタ数は既知として，正解セットのラベルを使わずにクラスタリングした．評価は，クラスタリング結果の各クラスタに含まれるクラスタ番号のうち最もサンプル数が大きいものを正解としたときに，正解が含まれる割合を正答率として算出した．データの前処理として，平均を 0，標準偏差が 1 になるように正規化を行った．圧縮する次元については予備実験をもとに Iris データでは $M=2$，Wine データでは $M=3$ とした．事前分布のハイパーパラメータはデータセット共通なものが $(\phi_0=1, c_0=d_0=\eta_0=10^{-3})$ で (a_0, b_0) に関しては，予備実験をもとに，Iris で $(a_0=10^{-3}, b_0=10^{-3})$，Wine で $(a_0=0.5, b_0=0.5)$ とした．この違いはデータセットによりクラスタ内の分散が大きく異なることを反映している．

MCMC の条件としては，初期値および，局所最適解の影響を少しでも緩和するために，交換モンテカルロ法を用いた[4]．ただし，ラベルスイッチングの問題を解決するのが困難なため，最終的には一つのモデルのみからのサンプルで評価を行った．具体的には，まず初めの 50,000 ステップは交換モンテカルロを通常通りに走らせ，残りは 50,000 ステップ目で温度が 1 であったモデルから 50,000 ステップのサンプリングを行った．各サンプル i について，$z_{ik}=1$ となったクラスタをカウントしてその数が最も大きいクラスタを最終的な所属クラスタとした．また最終結果は異なるシミュレーションからの 10 回の結果の平均と分散で測定することとした．

比較対象としては，次元の圧縮も外れ値の考慮もないガウス混合モデル（GM），次元圧縮はあるが外れ値の考慮はない混合因子モデル（MFA）[2]，提案手法である次元圧縮，外れ値共にモデルに組み込んだ混合 t 因子モデル（MTFA），および，参考として混合分布の標準的なパッケージである MClust を選択した．四つのモデルの比較によって，データに対する仮説を確率モデルに組む込む効果を検証するのが目的である．

4.3 結果

実験の結果は表 2 のとおり．ただし，GM と MClust の結果は文献[5]からの引用である．次元削減を考慮しないガウス混合分布が最も精度が悪く，提案手法である混合 t 因子分析は他のモデルと Iris データでは同程度，Wine データでは優位性があり，かつ完全なクラスタリングができていることが確認できる．また，交換モンテカルロ法の効果により，シミュレーションごとのばらつきが押さえられ，この実験の範囲内ではすべて同じクラスタが推定された．

表 2　正当率

	GM	MFA	MTFA	MClust
Iris	93.5 ± 1.3	**96.6** ± 0	**96.6** ± 0	**96.6** ± 0
Wine	96.6 ± 0.0	98.9 ± 0	**100.0** ± 0	97.1 ± 0

4. 4　考察

次元の高さに比べてデータ数が少ないケースではガウス混合分布のすべての共分散を推定するのは難しく，なんらかの次元圧縮を仮定したほうがクラスタリングの精度が上る可能性のあることがわかった．Iris データについては，正答率が 3 種のモデルで 96.6 と同一になっていることから，クラスタ間で本質的にデータが重なっているか，もしくは，モデルの仮定であるクラスタの楕円的な広がりが破られている可能性がある．Wine データについては，次元圧縮，外れ値への対処と仮説の導入につれて精度が上っていき，提案手法においては完全なクラスタリングを達成していることから，データの生成に対する仮説を確率モデルとしてある程度まで設計できたといえる．

5. お　わ　り　に

隠れた構造や，外れ値を持つ不確実データをモデル化するための方法として，データの生成過程に対する仮説を確率モデルとして表現し，ベイズ統計を利用してデータによるモデルの評価を行う手法を紹介した．さらに，高次元データのクラスタリングを例として，次元の圧縮と外れ値への頑健性を表現する方法を提案し，一部の実データで提案手法の有効性を確認した．

参考文献　［1］　渡辺澄夫，“ベイズ統計の理論と方法”，コロナ社，2012 年 3 月．
［2］　T. Hosino, “High Dimensional Non-linear Modeling with Bayesian Mixture of CCA”, Inter national Conefence of Neural Information Processing (ICONIP 2010), Novenver 2010, pp 446-453.
［3］　P. J. Deschamps, “A exible prior distribution for Markov switching autoregressions with Student-t errors”, Journal of Econometrics, July 2006, Volume 133, Issue 1, pp 153-190.
［4］　伊庭幸人，種村正美，“計算統計 2 マルコフ連鎖モンテカルロ法とその周辺（統計科学のフロンティア 12）”，岩波書店，2005 年 10 月．
［5］　C. Bouveyron, C. Brunet, “Simultaneous model-based clustering and visualization in the Fisher discriminative subspace”, Statistics and Comupting, January 2012, Volume 22, Issue 1, pp 301-324.

執筆者紹介　星　野　　力（Chikara Hoshino）
2000 年日本ユニシス入社．確率論とその応用に従事．工学博士．