1A1-2

ARMAモデルベース時系列クラスタイング

ARMA Model Based Time Series Clustering

末松 伸朗　　林 朗
Nobuo Suematsu　Akira Hayashi

*1広島市立大学大学院情報科学研究科
Graduate School of Computer Sciences, Hiroshima City University

In this work, we devise an ARMA model based time series clustering method based on a Dirichlet process mixture (DPM) model. DPM models enable full Bayesian analysis of mixture modeling and have been applied model based clustering tasks in recent years. However, the application of DPM model is not straightforward when we cannot use a conjugate prior as the base measure of the Dirichlet process. Since there is no conjugate prior for ARMA models (and most generative models for time series), we have to cope with the problems due to the non-conjugacy to realize sampling for the DPM of ARMA (DPM-ARMA). Our Markov Chain Monte Carlo algorithm for DPM-ARMA manages Metropolis-Hastings chains each of whose stationary distribution is the posterior of ARMA parameter vector given each time series observation. By using these Metropolis-Hastings chains, a Gibbs sampling for the DPM model can be performed.

1. はじめに

ディリクレ過程混合 (DPM) モデルに基づく ARMA モデルベース時系列クラスタリング法を開発する．モデルベースクラスタリングは，従来，事前に定めた数の要素モデルからなる混合モデルを EM アルゴリズム等でデータに当てはめることで行われてきた．有限混合 ARMA モデルによる時系列クラスタリングも提案されており，良好な結果が報告されている [Xiong 04]．これに対し，DPM モデルを用いるとクラスタ数も含め，ベイズ解析を行うことができる．ただし，解析的な取り扱いは困難でありマルコフ連鎖モンテカルロ (MCMC) 法によりサンプルを生成して解析を行うことになる [Neal 00]．

本研究では，ディリクレ過程混合 ARMA モデルに対する MCMC 法を提案し，時系列クラスタリングを実現する．

2. ディリクレ過程混合 (DPM) モデル

観測データ $\{z_1, \ldots, z_n\}$ に対する DPM モデルは，

$$\begin{aligned} z_i|\psi_i &\sim F(\psi_i) \\ \psi_i|G &\sim G \\ G &\sim DP(G_0, \alpha). \end{aligned} \tag{1}$$

と書ける．ここで，$F(\psi_i)$ はパラメータ ψ_i を持つ分布，G はそのパラメータ空間上の分布で，$DP(G_0, \alpha)$ は，基底測度 G_0，集中度 $\alpha > 0$ のディリクレ過程であり，G の事前分布である．

クラスタリングのために，事後分布 $p(\psi_1, \ldots, \psi_n|z_1, \ldots, z_n)$ を解析したいがその閉じた表式は得られないため，MCMC によりこの事後分布に従うサンプルを生成して解析に用いる．条件付密度関数

$$p(\psi_i|\psi_{-i}, z) = \frac{\alpha p(z_i)}{n-1+\alpha} p(\psi_i|z_i) + \sum_{j=1, j\neq i}^{n} \frac{p(z_j|\psi_i)}{n-1+\alpha} \delta(\psi_i - \psi_j) \tag{2}$$

からのサンプリングが可能であれば，$i = 1, \ldots, n$ について順次 ψ_i をサンプリングすることでモデル (1) に対するギブスサンプラーを実現できる．ただし，$\psi_{-i} = \{\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_n\}$ であり，$\delta(\cdot)$ はディラックのデルタ関数である．また，$p(z_i)$ はデータ z_i の周辺尤度，$p(\psi_i|z_i)$ はデータ z_i が与えられたときのパラメータの事後密度，$p(z_j|\psi_i)$ はパラメータ ψ_i のデータ z_j に対する尤度である．

3. ディリクレ過程混合 ARMA モデル

ARMA(p, q) モデルは p 次の自己回帰モデルと q 次の移動平均モデルを合わせたモデルであり，ARMA(p, q) に従う時系列 $\{Z_t\}$ は

$$Z_t + \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} = U_t + \theta_1 U_{t-1} + \cdots + \theta_q U_{t-q} \tag{3}$$

によって定義される．ここで，U_t は平均 0, 分散 σ^2 の正規分布に従う独立なノイズである．したがって，ARMA(p, q) のパラメータ $\psi = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ であり，観測時系列データ $z_i = \{z_{i,1}, \ldots, z_{i,T}\}$ は $\{Z_t\}$ の実現値である．モデル (1) の分布 $F(\psi_i)$ をこの ARMA モデルとすることで，ディリクレ過程混合 ARMA（DPM-ARMA）モデルとなる．

ARMA モデルのように共役事前分布を持たない要素モデルを用いる場合, 条件付密度関数 (2) からのサンプリングを直接的に行うことができない．周辺尤度 $p(z_i) = \int p(z_i|\psi_i)p(\psi_i)d\psi_i$ と事後分布 $p(\psi_i|z_i)$ が解析的に得られないため，前者の評価と後者からのサンプリングを直接的に行えないのがその原因である．

[Neal 00] に示されたアルゴリズムには，非共役 DPM モデルに対するものも含まれる．しかし，そのアルゴリズムでは，一つのクラスタに含まれるデータ集合が与えられたときのパラメータの事後分布からのサンプリング法が利用できることを仮定しているが，その効率的な実現方法については述べられていない．

本研究では，以下の 2 つの対応により，式 (2) からのサンプリングを可能にする．

周辺尤度は，パラメータの事前分布が決まれば各データについて一度計算しておけばよい．パラメータの次元が比較的

連絡先: 末松伸朗, 広島市立大学大学院情報科学研究科, 広島市安佐南区大塚東 3-4-1, suematsu@hiroshima-cu.ac.jp

低い場合にはモンテカルロ積分で十分であるが，本研究では Annealed Importance Sampling[Neal 01] を応用して周辺尤度を推定する．

また，事後分布 $p(\psi_i|z_i)$ からのサンプリングは，この分布に対する MH 法に基づくマルコフ連鎖をシミュレートすることにより行う．すなわち，データ数と同じ数の MH 連鎖を管理し，$p(\psi_i|z_i)$ からのサンプリングが必要になる度に対応する MH 連鎖のシミュレーションをある一定のステップ数だけ進めてサンプルを得るのである．

4. クラスタリング

ディリクレ過程混合 ARMA モデルに対する MCMC 法ができたので，事後分布 $p(\psi_1, \ldots, \psi_n|z_1, \ldots, z_n)$ に従うサンプルを多数生成することができる．各サンプルは，それぞれ一つのクラスタリング結果を与えるが，それらの情報を統合して一つの結果へまとめるには様々な方法が考えられる．

本研究では，データ z_i, z_j が同じクラスタに属す事後確率 $P(\psi_i = \psi_j|z_1, \ldots, z_n)$ を MCMC サンプル中の頻度から推定して用いる．その推定値を$\hat{}$付で示すと，z_i と z_j の距離は

$$d(z_i, z_j) = 1 - \hat{P}(\psi_i = \psi_j|z_1, \ldots, z_n) \qquad (4)$$

により定義される．そして，この距離に基づいて凝集型の階層的クラスタリングを行う．

非階層的クラスタリング結果が必要な場合，MCMC サンプルにおけるクラスタ数の相対頻度により階層数の事後分布 $P(k|z_1, \ldots, z_n)$ を推定し，事後確率最大のクラスタ数となるよう階層的クラスタリング結果の切断を行うのが素直なアプローチであろう．

5. 実験

提案手法の妥当性を実験で確認する．実験には [Kalpakis 01, Xiong 04] で用いられている実データの一つを使う．このデータセットは，1929 年から 1999 年までのアメリカの 25 の州における一人当たりの個人収入の推移である．[Kalpakis 01] では，個人収入の伸び率の高い東海岸の 17 州と，伸び率の低い中西部の 8 州の 2 つのグループに分けるのを正解と設定している[*1]．前処理や ARMA モデルの次数は [Kalpakis 01] に従い AR(1) モデルを要素モデルに用いる．

提案した MCMC 法によりパラメータ集合 $\{\psi_i\}_{i=1}^{25}$ の 2×10^6 サンプルを生成し，後半の 1×10^6 を解析に用いた．4. に述べた方法で得た 25×25 の距離行列を図 1 に示す．式 (4) で定義される距離は 0 から 1 の値をとるので，それを黒から白のグレースケールに対応付けて表示している．東海岸の 17 州が 1 から 17 の行（列）に対応している．図において概ね 2 グループに分かれる傾向が見られるが，どちらとも言えないデータも少なくないことが分かる．この距離行列から得られた樹形図を図 2 に示す．階層的クラスタリングのクラスタ間距離には平均距離を用いた．図中，略称の右肩に $*$ を付されたのが中西部の 8 州である．また，略称横に示した数字は図 1 で対応する行（列）番号である．表 1 に示したクラスタ数の相対頻度から分かるように，提案手法の結果は 3 クラスタとするのが適当であることを比較的強く示している．クラスタ数が 3 となるよう図 2 の破線で切断すると，中西部のグループが 2 つに分かれるが，東海岸のグループは一つにまとまっており，妥当な結果と言えるだろう．

*1 ただし，この正解がデータをよく調べた上で選ばれた訳ではなさそうである．

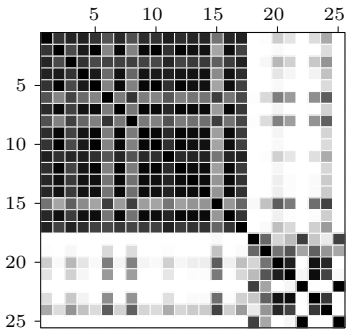

図 1: 距離行列

表 1: クラスタ数の相対頻度

k	相対頻度 [%]
2	0.28
3	86.89
4	12.67
5	0.17
6	0.01 未満

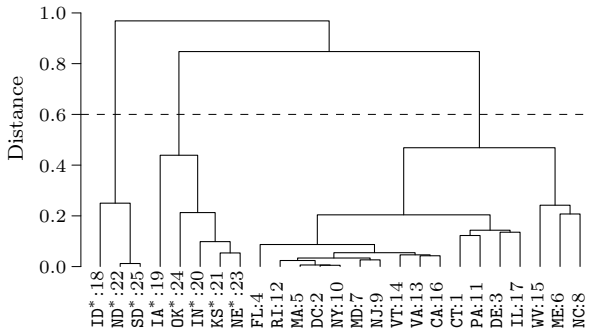

図 2: 樹形図

6. まとめ

ディリクレ過程混合モデルに基づいた ARMA モデルベース時系列クラスタリングを開発した．クラスタリング法実現のため，非共役事前分布を用いる場合のディリクレ過程混合モデルに対するマルコフ連鎖モンテカルロを，データと同じ数のメトロポリス・ヘイスティングス連鎖をシミューレトすることにより構成する方法を提案し利用した．そして，実データに提案手法を適用し，概ね妥当な結果が得られることを確認した．

今後，多くの実験を実施し，提案したサンプリング手法で得られるクラスタリング結果の質などについて詳しく調べる必要がある．また，混合率が低いであろうと思われるギブスサンプラーに基づいているので，収束性の評価や，混合率を高める工夫について検討する必要があるだろう．

参考文献

[Kalpakis 01] Kalpakis, K., Gada, D., and Puttagunta, V.: Distance Measures for Effective Clustering of ARIMA Time-Series, in *Proceedings of the IEEE International Conference on Data Mining*, pp. 273 – 280 (2001)

[Neal 00] Neal, R. M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models, *Journal of Computational and Graphical Statistics*, Vol. 9, No. 2, pp. 249–265 (2000)

[Neal 01] Neal, R. M.: Annealed Importance Sampling, *Statistics and Computing*, Vol. 11, No. 2, pp. 125–139 (2001)

[Xiong 04] Xiong, Y. and Yeung, D.-Y.: Time Series Clustering with ARMA Mixtures, *Pattern Recognition*, Vol. 37, pp. 1675 – 1689 (2004)