様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25 年　3 月　31 日現在

機関番号：25403
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：22500261
研究課題名（和文）　ディリクレ過程混合 ARMA モデルによる時系列クラスタリング

研究課題名（英文）　Time series clustering with Dirichlet process mixtures of ARMA models

研究代表者
末松 伸朗　（SUEMATSU NOBUO）
広島市立大学・情報科学研究科・准教授
研究者番号：70264942

研究成果の概要（和文）：ARMA モデルは非常に簡潔な時系列データの確率モデルである．時系列データの集合が与えられたとき，ARMA モデルの混合モデルから生じたと見なし，その混合モデルをデータへ当てはめると時系列データのクラスタリングが実現できる．このとき，ディリクレ過程を使うと，クラスタの数についても推論が行える．本研究では，このようなモデルのデータへの当てはめを可能とするマルコフ連鎖モンテカルロアルゴリズムを開発した．

研究成果の概要（英文）：ARMA models are parsimonious stochastic models for time series. Given a set of time series, we can cluster them by regarding that they were drawn from a mixture of ARMA models and by fitting the model to them, where if the mixture model is a Dirichlet process mixture, the number of the clusters can be simultaneously estimated. In this work, we have developed a Markov Chain Monte Carlo method to fit a Dirichlet process mixture of ARMA models to a set of time series.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,200,000	360,000	1,560,000
2011 年度	500,000	150,000	650,000
2012 年度	400,000	120,000	520,000
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：時系列解析

１．研究開始当初の背景

ノンパラメトリックベイズ解析の利用が急速な拡大を見せていた．ベイズ統計は，古くはパラメトリックモデルに対して用いられてきた．しかし，ベイズ統計の枠組みは，ノンパラメトリックな確率密度や関数を推定する問題へ適用可能であり，理論の整備やコンピュータの発展，そして，数値計算アルゴリズムの開発により，それが可能となって来たため，急激に利用が広がったのである．

ディリクレ過程混合モデルは，ノンパラメトリックベイズ解析の代表的な例として，よく研究されていた．このモデルは，実質的には加算無限個の要素モデルからなる混合モデルを考えることに相当する．そして，実際にデータを生み出すのに使われる要素モデルの数がデータのクラスタ数に当たり，その数の推定が同時に行えることがディリクレ過程混合モデルの重要な特長の一つである．

ディリクレ過程混合モデルは，様々な対象

に適用されたが，時系列モデルは要素モデルが複雑になるため，ほとんど使われていなかった．ARMA モデルは，時系列モデルとしては比較的簡潔なモデルであるが，やはりディリクレ過程混合モデルに使われた例は存在しておらず，本研究は先進的な取り組みであった．

２．研究の目的

ディリクレ過程混合（Dirichlet Process Mixture; DPM）モデルを用いたモデルベース時系列クラスタリングの研究を行う．

モデルベース時系列クラスタリングでは，有限混合モデルが広く用いられて来たが，DPM モデルを用いると，有限混合モデルでは困難な，クラスタ数やクラスタ構造に関するベイズ推論も可能となる．

ただし，DPM モデルを用いるには，適切なマルコフ連鎖モンテカルロ（MCMC）法によるサンプリングアルゴリズムの開発が必要となる．

本研究課題では，ディリクレ過程混合 ARMA モデルに対して効率の良いサンプリングアルゴリズムを開発することを目指す．

MCMC 法を利用する場合，得られたサンプルからどのように必要な情報を抽出するのかも常に問題となる．そこで，クラスタ解析のために有効な MCMC サンプルからの情報抽出法についても研究を行う．

３．研究の方法

ディリクレ過程混合 ARMA モデルのための MCMC アルゴリズムを，一般の DPM モデルで使われるギブスサンプリングに基づいて開発する．ただし，ARMA モデルの場合に問題となる部分があるが，その部分については，他の近似法や MCMC 法を援用することにより解決する．その後，シミュレーション実験により，開発した手法の有効性を確認する．

４．研究成果

(1) 問題の定義

時系列データの集合を $\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ とすると，このデータ集合に対する DPM モデルは右図のように表される．図中の ψ_i は，ARMA モデルのパラメータ集合を表しており，G は，パラメータを生み出す未知の分布である．また，図から分かるように，各時系列データ $\mathbf{z}_i$ に独自のパラメータ ψ_i が対応しているが，G は離散分布であり，同じパラメータを複数回生成する確率がある．そして，パラメータを共有する時系列の集合がクラスタを形成することになる．

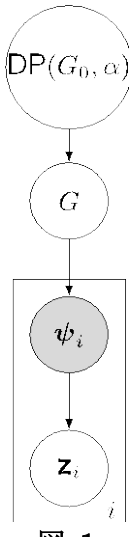

図 1

(2) MCMC アルゴリズムの開発

図１に示した DPM モデルに対する最も基本的な MCMC 法は，ギブスサンプリングに基づくものである．しかし，それをこの問題で使うためには次の２つの問題を解決しなければならなかった．

一つは，時系列の周辺尤度 $p(\mathbf{z}_i)$ の評価である．ARMAモデルのように共役事前分布を持たないモデルを使う場合，これが解析的に得られない．この問題に対しては，モンテカルロ積分で近似する方法，すなわち，パラメータの事前分布 G_0 から多数のサンプル $\{\psi^{(1)},\ldots,\psi^{(N)}\}$ を生成し，

$$p(\mathbf{z}_i) \simeq \frac{1}{N}\sum_{n=1}^{N} p(\mathbf{z}_i \mid \psi^{(n)})$$

を評価する方法と，Annealed Importance Sampling（Neal 2001）と呼ばれる方法を試したが，いずれの場合も良好な結果が得られている．なお，この計算は，事前に一度行っておけばよいものであり，計算の中心となるマルコフ連鎖のシミュレーション時には計算結果の数値を利用するだけである．

二つ目の問題は，一つの時系列データが与えられたときのパラメータの事後分布 $p(\psi|\mathbf{z}_i)$ からサンプリングが行えないことである．この問題に対しては，メトロポリス・ヘイスティングス(MH)法によりサンプルを生成する方法を提案した．本研究によるサンプリング法が効率のよいものとなっているのはこの部分の実現方法にある．

この $p(\psi \mid \mathbf{z}_i)$ からのサンプルを一つ得るために，長い MH 法の連鎖を行ったのでは，非常に多くの計算量を要する．そこで，本研究では，この事後分布が，時系列と同数しかないことに着目し，それだけの数の MH 法の連鎖を並行して走らせるのである．ギブスサンプリングの部分で実際に $p(\psi \mid \mathbf{z}_i)$ からのサンプルが必要になれば，その並行して存在する連鎖からサンプルを取り出し，連鎖をある回数進めるのである．サンプルが必要とされない間は，その連鎖はシミュレーションが停止した状態にあり，計算量に対する負荷は存在しない．

以上により，ギブスサンプリングを実現する上での二つの問題が解決された．

(3) MCMC サンプルからのクラスタ構成法

開発した MCMC アルゴリズムにより，パラメータ集合 $\{\psi_1,\ldots,\psi_n\}$ の多数のサンプルが得られるが，それらは多様なクラスタリングの可能性を含むものであり，実際にクラスタリング結果を定めるには，そこから情報を読み取らなければならない．

本研究では，二つの時系列データ $\mathbf{z}_i$，$\mathbf{z}_j$

間の距離を

$$d(z_i, z_j) = 1 - \hat{P}_{i,j}$$

と定義し，階層クラスタリング法を利用する方法を提案した．ここで，$\hat{P}_{i,j}$ は，$\mathbf{z}_i$ と $\mathbf{z}_j$ が同じクラスタに所属する確率の推定値であり，MCMC サンプルの中で ψ_i と ψ_j が一致する割合により求められる．上記の時系列間の距離としては，他にも様々な定義が考えられるが，クラスタリング結果への影響は非常に小さかった．

また，クラスタリング法としては，階層クラスタリング以外に Partition Around Medoids (PAM) も試したが，階層クラスタリングとの差異はあまり見られなかった．

(4) 例題

実例に従って，開発した時系列クラスタリング法の流れを説明する．この例で使用するデータは，アメリカの 25 の州における一人当たりの個人収入の 1929 年から 1999 年までの推移である．その間の各州の経済状況に基づき，東海岸の 17 州と中西部の 8 州は分けられるだろうとされている．ARMA モデルを適用するために，定常時系列と見なせるよう前処理を施したデータを図２，３に示す．

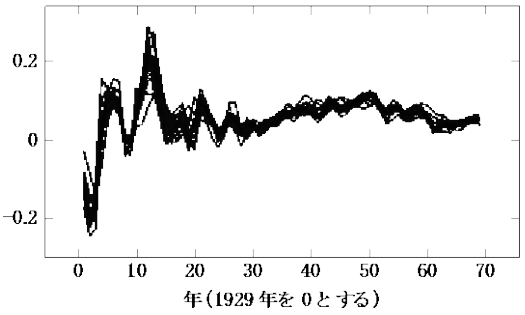

図 2　東海岸の 17 州

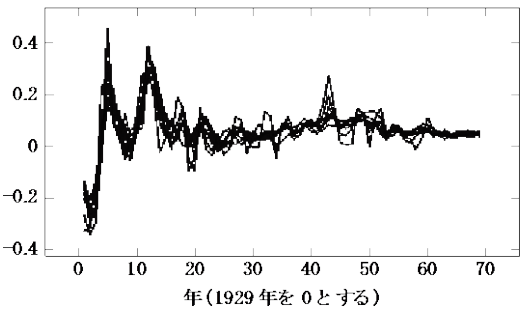

図 3　中西部の 8 州

これらの図から，２グループの間に明確な差を見つけるのは難しいことが分かるだろう．

これらの時系列データ集合に対して，開発した MCMC アルゴリズムにより，200 万ステップのシミュレーションを行い，後半の 100 万ステップに対応する 100 万サンプルをクラスタリングに用いた．

まず，サンプル中のクラスタ数の相対頻度を下表に示す．

表１　クラスタ数の相対頻度

クラスタ数	相対頻度[%]
2	0.28
3	86.89
4	12.67
5	0.17
6	0.01 未満

この表より，この結果では３クラスタへ分割すべきことを強く示している．

次に，時系列データ間の距離を濃淡で表示する図を示す．

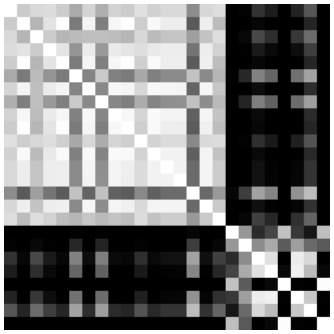

図 4　距離行列

この図では，近い距離ほど明るい色で示されている．この図から，インデックスの小さい方の過半数の時系列が，相互の距離が比較的近いグループを形成していることが分かる．実際には，このグループは，東海岸の 17 州からなっている．

次に，図４の距離行列を使って，階層クラスタリングを行って得られた樹形図を図５に示す．

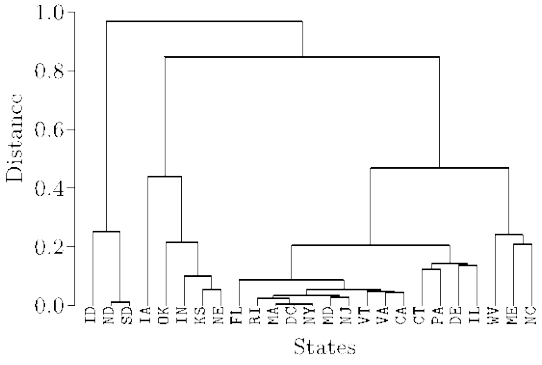

図 5 樹形図

この樹形図において，クラスタ数の相対頻度が示したクラスタ数３となるように分割したとき，右側に位置するクラスタが東海岸の 17 州からなっている．残りの 8 州が中西部の州であるが，この結果では２グループに分かれている．アイダホ，ノースダコタ，サ

ウスダコタの３州が最も左のグループで，アイオワ，オクラホマ，インディアナ，カンザス，ネブラスカの５州が左から２番目のグループである．各州の配置を考えると，この分割は納得できるものであり，むしろ本手法の有効性を指示する結果となっている．

(5) 今後の展望

本研究は，ディリクレ過程混合モデルで本格的な時系列モデルを取り扱った最初の事例である．ディリクレ過程混合モデルは，現在も急速にその適用範囲を広げつつある中で，本研究で達成された内容は重要な意味を持つと考える．

しかし，本研究およびそこから発展する研究には，今後解決すべき課題が多く残されている．それらの課題を以下に述べる．

① ARMA モデルの次数の推定

本研究では，ARMA モデルの次数の推定は行っていない．従って，前もって時系列データを分析し，適当な次数を判定しなければならない．ARMA モデルの表現には多くの冗長性があるため，次数の推定を組み入れるのは容易ではない．しかし，次数の異なる ARMA モデルを含む混合モデルも当然考えられるので，この問題の解決は重要である．

② ARMA モデルによる制約

本研究では，定常 ARMA モデルで表される時系列を対象としている．しかし，そのままで定常な時系列データは比較的少なく，実際には適切な差分を取るなどして定常時系列として扱えるような前処理を行っている．このことは，時系列データの非定常性の部分に特徴がある場合，適切にクラスタリング出来ないことを意味する．

この問題を抜本的に解決するには ARIMA モデルの混合モデルへと拡張する必要があるだろうが，それも上のモデルの次数の同時推定が出来なければ不可能である．

③ MCMC アルゴリズムの改善

本研究で開発した MCMC アルゴリズムは，最も基本的なギブスサンプリングを基礎としている．このアプローチでは，比較的容易にサンプリングを実現できるが，実現されるマルコフ連鎖の混合率はあまり高くならないという問題がある．従って，的確なクラスタリングを実現するには，非常に多くのサンプルを生成する必要があることになる．

パラメータの新しいサンプルを，現在行っているようなデータ単位ではなく，クラスタ単位に生成する方法があり，その方法に基づくマルコフ連鎖の方が高い混合率が実現でき，同じサンプリングステップ数に対して，実質的なデータ数は増えることになる．このようなより洗練されたサンプリング法をベースにしたアルゴリズムを開発することができれば，非常に大規模な時系列データへの適用も容易となるなど，多くのメリットが得られるだろう．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

① 秋本　真治，　末松　伸朗，　林　朗，岩田　一貴，ガウス過程に基づくノンパラメトリックベイズ時系列整列，電子情報通信学会論文誌 D，査読有，Vol.J96-D，No.3，pp.587-595，2013.

② 金子 悟士，林　朗，末松　伸朗，岩田一貴，階層的時系列データのための識別モデル，電子情報通信学会論文誌 D，査読有，Vol.J96-D，No.2，pp.306-315，2013.

③ Katsutoshi Ueaoki，Kazunori Iwata，Nobuo Suematsu，Akira Hayashi，Matching Handwritten Line Drawings with Von Mises Distributions，査読有，IEICE Transactions on Information and Systems，Vol.E94-D，No.12，pp.2487-2494，2011.

④ 玉田　寛尚，林　朗，末松　伸朗，岩田一貴，階層隠れCRF，査読有，電子情報通信学会論文誌 D，Vol.J93-D，No.12，pp.2610-2619，2010.

〔学会発表〕（計６件）

① 栗栖　昂勢，末松　伸朗，岩田　一貴，林　朗，ガウス過程事前分布を用いた空間変化混合モデルによる画像分割，情報処理学会第 75 回大会全国大会、2013 年 3 月 8 日.

② Nobuo Suematsu，Akira Hayashi，Time Series Alignment with Gaussian Processes，Proc. the 21st International Conference on Pattern Recognition，査読有，pp.2355-2358，Nov. 14，2012，Japan.

③ 秋本　真治，末松　伸朗，林　朗，岩田一貴，ガウス過程事前分布を用いた時系列整列，電子情報通信学会 NC研究会，2012 年 3 月 15 日.

④ Satoshi Kaneko，Akira Hayashi，Nobuo Suematsu and Kazunori Iwata，Hierarchical Hidden Conditional Random Fields for Information Extraction，Proc. the Learning and Intelligent Optimization，査読有，Jan. 2011，Italy.

⑤ Shinji Akimoto，Nobuo Suematsu，A Nonparametric Bayesian Approach to Time Series Alignment，Proc. Second World Congress on Nature and

Biologically Inspired Computing, 査読有, pp.655-660, Dec.17, 2010, Japan.
⑥ 秋本 真治, 末松 伸朗, 林 朗, 岩田一貴, ガウス過程事前分布を用いた時系列多重整列法, 第9回情報科学技術フォーラム, 2010年9月7日.

６．研究組織

(1)研究代表者

末松 伸朗 (SUEMATSU NOBUO)
広島市立大学・情報科学研究科・准教授
研究者番号：70264942