H-004

ガウス過程事前分布を用いた時系列多重整列法
Multiple Alignment of Time Series with Gaussian Process Priors

秋本　真治 †　末松　伸朗 †　林　朗 †　岩田　一貴 †
Shinji Akimoto　Suematsu Nobuo　Akira Hayashi　Kazunori Iwata

1　はじめに

実数値をとる時系列に対する，ガウス過程事前分布を利用した多重整列法を提案する．本手法では，ガウス過程事前分布を用いることで，一貫してベイズアプローチをとることを目指している．そのため，例えば，動的時間伸縮で見られるような特異な対応 [1] を適切に回避したり，時系列間の対応に複数の解釈が存在する場合には，適切にそれらに対応する解を発見できる等の優れた特徴を持つ整列法の実現が期待できる．

多重整列は，遺伝子やタンパク質のアミノ酸配列に対して盛んに研究されて来ているが，実数値時系列に対する研究は比較的限られている．たとえば，[2] では，隠れマルコフモデルに基づく手法が提案されている．

一般的に多重整列では，基本となる時系列（標準時系列）に何らかの変異が生じて観測系列が得られるという生成過程を仮定する．本研究では，標準時系列に対する非線形時間変換とノイズの付加により観測時系列が得られるというモデルを考え，標準時系列と非線形時間変換関数の事前分布としてガウス過程をおくモデルを提案する．そして，マルコフ連鎖モンテカルロ法により，事後分布に従う標準時系列と時間変換関数のサンプルを生成することでベイズ解析を実現する手法を提案する．

2　時系列生成モデル

時系列データが n 本あるとき，i 番目の時系列データ $\mathbf{y}_i = (y_{i,1},...,y_{i,T})$ は標準時系列 $f(t)$ に時間変換が加わったものを時刻 $\mathbf{t} = (t_1,...,t_T)$ で観測することで得られると考える．すなわち，

$$y_{i,j} = f(s_i(t_j)) + \varepsilon_j.$$

ここで $s_i(t)$ は時間変換関数であり，ε_j は分散 σ_n^2 の正規分布に従うノイズである．ベイズ解析を実現するために，未知の関数 $f(\cdot)$ と $s_i(\cdot)$ の事前分布としてガウス過程をおく．ガウス過程については次章で説明する．また，簡単のため，全ての時系列は $\mathbf{t}$ で観測されたものとし，$t_1 = 0, t_T = 1, s_i(0) = 0, s_i(T) = 1$ と仮定する．

3　ガウス過程 [3]

確率的に定まる関数 $f(x)$ が平均関数 $m(x)$, 共分散関数 $k(x_p,x_q)$ のガウス過程に従うとき，

$$f(x) \sim GP(m(x),k(x_p,x_q))$$

と書く．このとき，任意の $\mathbf{x}=(x_1,...,x_n)'$ に対して，$\mathbf{f} = (f(x_1),...,f(x_n))'$ は平均 $\mathbf{m}=(m(x_1),...,m(x_n))'$, 共分散行列 $K(\mathbf{x},\mathbf{x})$ の正規分布に従う．ここで，$K(\mathbf{x},\mathbf{x}) = [k(x_i,x_j)]_{n\times n}$ である．提案手法で述べる $f(\cdot)$ と $\mathbf{s}_i$ のサンプリングを行う際にガウス過程回帰法を用いるので，以下の節でそれを説明しておく．

3.1　観測値にノイズがない場合の事後分布

未知の関数 $f(x)$ の事前分布が $GP(m(x),k(x_p,x_q))$ であるときに，ノイズを含まない観測値 $\{(x_i,f(x_i))\}_{i=1}^n$ が与えられたとすると $f(x)$ の事後分布は

$$f(x) \mid \{x_i,f(x_i)\}_{i=1}^n \sim GP(m_*(x),k_*(x_p,x_q)) \tag{1}$$

である．ここで，

$$m_*(x) = K(x,\mathbf{x})K(\mathbf{x},\mathbf{x})^{-1}f(x),$$
$$k_*(x_p,x_q) = K(x_p,x_q) - K(x_p,\mathbf{x})K(\mathbf{x},\mathbf{x})^{-1}K(\mathbf{x},x_q).$$

3.2　観測値にノイズが含まれる場合の事後分布

ノイズを含む観測値 $\{(x_i,y_i)\}_{i=1}^n$ が与えられたとする．ここで $y_i = f(x_i) + \varepsilon_i$ である．このとき $f(x)$ の事後分布は

$$f(x) \mid \{x_i,y(x_i)\}_{i=1}^n \sim GP(m_*(x),k_*(x_p,x_q)) \tag{2}$$

である．ここで，

$$m_*(x) = K(x,\mathbf{x})[K(\mathbf{x},\mathbf{x}) + \sigma_n^2 I]^{-1}f(x),$$
$$k_*(x_p,x_q) = K(x_p,x_q) - K(x_p,\mathbf{x})[K(\mathbf{x},\mathbf{x}) + \sigma_n^2 I]^{-1}K(\mathbf{x},x_q).$$

4　提案手法

推測したい $f(\cdot)$ と $\{\mathbf{s}_i\}_{i=1}^n$ の事後分布に従うサンプルをマルコフ連鎖モンテカルロ法で生成し解析を行う．$f(\cdot)$ のサンプリングにはギブス・サンプラーを用い，$\mathbf{s}_i$ のサンプリングにはメトロポリス・ヘイスティングス法 (以下 MH 法) を用いる．以下の節で提案アルゴリズムとそれぞれのサンプリング法を説明する．

4.1　提案アルゴリズム

アルゴリズム 1 に提案アルゴリズムをまとめる．3 行目で $f(\cdot)$ のサンプリングを行い，4~9 行目で $\mathbf{s}_i$ のサンプリングを行う．それぞれのサンプリング法は次節以降で説明する．

4.2　$f(\cdot)$ のサンプリング

$f(\cdot)$ のサンプリングにはギブス・サンプラーを用いる．$f(\cdot)$ の事後分布は $\{\{(s_{i,j},y_{i,j})\}_{j=1}^T\}_{i=1}^n$ をノイズ付き観測値と見なして式 (2) を用いることで得られる．ただし，実際に関数 $f(\cdot)$ をサンプリングすることはできない．本研究では，サンプル点数を十分多く取り，スプライン補間を用いて $f(\cdot)$ のサンプリングを実現している．

† 広島市立大学大学院情報科学研究科

アルゴリズム 1 提案アルゴリズム

1: $\mathbf{s}_1,...,\mathbf{s}_n$ の初期値を $s=\mathbf{t}$ とする．
2: **for** $k=1:M$ **do**
3: $f(\cdot)\leftarrow$ 4.2 節で述べる事後分布からサンプリング．
4: **for** $i=1:n$ **do**
5: **for** $b=1:B$ **do**
6: 4.3 節で述べる提案分布より候補 $\mathbf{s}'_{i,J_b}$ を生成．
7: $\mathbf{s}'_{i,J_b}$ を確率的に受理し，$\mathbf{s}_{i,J_b}\leftarrow\mathbf{s}'_{i,J_b}$
8: **end for**
9: **end for**
10: **end for**

4.3 $\mathbf{s}_i$ のサンプリング

$\mathbf{s}_i$ のサンプリングでは，$\mathbf{s}_i$ 上に B 個のブロックを定義し，ブロック毎に MH 法を適用する．ブロック b に含まれる $\mathbf{s}_i$ の添字集合を J_b とすると，$\mathbf{s}_{i,J_b}=\{s_{i,j}\mid j\in J_b\}$ に対する提案分布は，$\{(t_j,s_{i,j})\mid j\notin J_b\}$ をノイズなしの観測値と見なして，式 (1) を用いることで得られる．$\mathbf{s}_i$ のサンプル点数 $T=10$，ブロックサイズ $|J_b|=4$ の場合の提案分布の模式図を図 1 に示す．$s_i(0)=0, s_i(T)=1$ と仮定しているので，各ブロックは $J_1=\{2,3,4,5\}, J_2=\{3,4,5,6\},...,J_5=\{6,7,8,9\}$ となる．図 1 はブロック J_2 に対する提案分布を示しており，色が濃い部分ほど確率が高いことを表している．

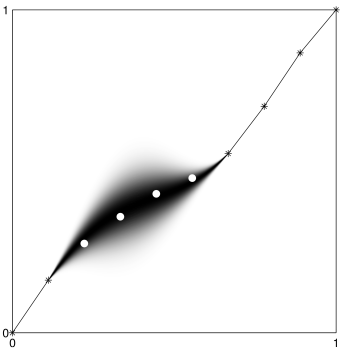

図 1　提案分布の模式図

5　実験

以上で提案した手法を用いて，多重整列実験を行い，その有効性を検証する．標準時系列 $f(t)=\sin(5\pi t)$ とし，時間変換関数 $\mathbf{s}_i$ をガウス過程によりランダムに生成し，ノイズ (分散 $\sigma_n^2=0.04$) を加えて時系列データを 5 本作成した．また，1 時系列のサンプル点数 $T=20$ とした．

生成された 5 本の時系列に対して，アルゴリズム 1 を適用し，サンプリングを 6000 回行った．

各サンプルに対する対数尤度

$$\log p(\{\mathbf{y}_i\}_{i=1}^n\mid f,\{\mathbf{s}_i\}_{i=1}^n,\mathbf{t},\sigma_n^2)$$

を図 2 に示す．図 2 をみると，MCMC が定常状態になっている様に見える．

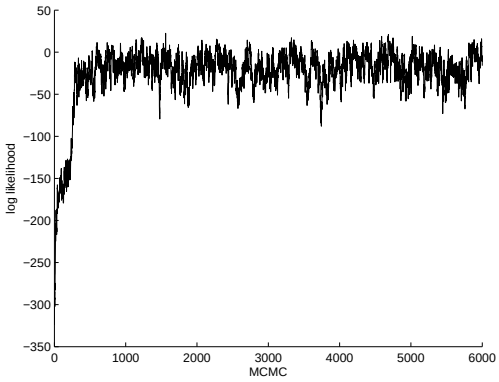

図 2　尤度の変化

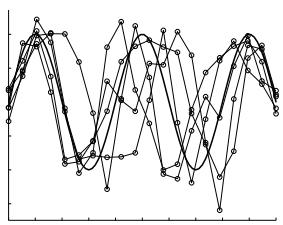

図 3　整列前

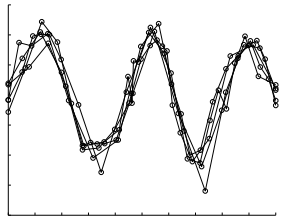

図 4　整列後

図 3 に標準時系列（太線:$f(t)=\sin(5\pi t)$）と整列前のデータ，図 4 に対数尤度が最大のときの $\{\mathbf{s}_i\}_{i=1}^n$ を用いて $\{(s_{i,j},y_{i,j})\mid j=1,...,T,\ i=1,...,n\}$ をプロットしたものを示す．図 4 をみると提案した手法により，時系列の整列ができていることがわかる．

6　まとめ

ガウス過程事前分布を用いた時系列の生成モデルを考案し，マルコフ連鎖モンテカルロ法を用いてベイズ解析を行う手法を提案した．提案手法では，尤度の高いサンプルに基づいて多重整列を実現する．また，人工データを用いた実験を行い，提案手法の有効性を示した．

参考文献

[1] KEOGH, E. J. and PAZZANI, M. J. Derivative Dynamic Time Warping, First SIAM International Conference on Data Mining(SDM'2001) (2001).

[2] LISTGARTEN, J., NEAL, R. M., ROWEIS, S. T. and EMILI, A. Multiple Alignment of Continuous Time Series, Advances in Neural Information Processing Systems 17 (eds.Saul, L. K., Weiss, Y. and Bottou, L.), MIT Press, Cambridge, MA (2005), 817–824.

[3] RASMUSSEN, C. E. and WILLIAMS, C. K. I. *Gaussian Processes for Machine Learning*, MIT Press (2006).