GTM を用いた時系列データの可視化
Visualization of Time Series Data Using GTM

山口 暢彦 (PY)†
Nobuhiko Yamaguchi(PY)
† 佐賀大学理工学部知能情報システム学科
yamag@is.saga-u.ac.jp

Abstract— The object of this paper is to visualize time series data using GTM (generative topographic mapping). The standard GTM algorithm assumes that the data are independent, identically distributed samples. For time series, however, the i.i.d. assumption is a poor approximation. In this paper we propose the extension of the GTM to time series.

Keywords— **GTM, Visualization, Time Series, AR Model**

1 はじめに

観測データの裏にひそむ本質的な構造を探る手法として，潜在変数を用いて観測データの分布をモデル化する手法がある．その代表的な例として，観測データを潜在変数の線形変換を用いて表現することにより構造を探る因子分析等が挙げられる．これに対し，文献 [1] では，観測データを潜在変数の非線形変換を用いて表現することにより構造を探る Generative Topographic Mapping(GTM) が提案され，データの可視化やクラスタリング等において応用が試みられている．GTM は，同様にデータの可視化等を目的とする自己組織化マップと比べ，確率論的にモデル化されている等の幾つかの有用な特性が示されている [1].

本論文では，GTM を時系列データに適用する方法 GTM-AR について提案を行い，GTM-AR を用いて時系列データの可視化を行う．本論文の構成は以下の通りである．2 では，時系列データの基本モデルである AR(Autoregressive) モデル [2] について述べる．3 では，GTM-AR の提案を行う．4 は計算機実験である．

2 AR モデル

時点 t における確率過程の値 X_t が直近 p 個の過去の値 $X_{t-1}, \cdots, X_{t-p}$ を用いて式 (1) と表される時系列モデルを AR モデルと呼ぶ.

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t \tag{1}$$

ただし，ここで $\phi_0, \cdots, \phi_p$ は AR 係数であり，$\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$ である．この時，T 個の観測値 $\boldsymbol{x} = (x_1, \cdots, x_T)^T$ の条件付き尤度関数は式 (2) で与えられる [2].

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{\phi}) &= \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2} e_t^2\right), \\ e_t &= x_t - \phi_0 - \sum_{i=1}^{p} \phi_i x_{t-i} \end{aligned} \tag{2}$$

ここで，$\boldsymbol{\phi} = (\phi_0, \cdots, \phi_p)^T$ は全ての AR 係数を並べた列ベクトルである．

3 GTM-AR

3.1 潜在変数を用いた時系列データの生成モデル

GTM-AR では，時系列データ $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ を潜在変数 $\boldsymbol{u} = (u_1, \cdots, u_L)^T$ を用いて表現することを考える．このため，潜在変数 $\boldsymbol{u}$ を AR 係数 $\boldsymbol{\phi}$ に写す関数 $\boldsymbol{y}(\boldsymbol{u}; \boldsymbol{W})$ を定義する．

$$\boldsymbol{\phi} = \boldsymbol{y}(\boldsymbol{u}; \boldsymbol{W}) = \boldsymbol{W}\boldsymbol{\psi}(\boldsymbol{u})$$

ここで，$\boldsymbol{W}$ は関数 $\boldsymbol{y}$ を決定するパラーメータ行列であり，$\boldsymbol{\psi} = (\psi_1, \cdots, \psi_M)^T$ は M 個の固定された基底関数である．又，潜在変数 $\boldsymbol{u}$ が与えられたとした際，時系列データ $\boldsymbol{x}$ は AR モデルにより生成されていると考え，$\boldsymbol{x}$ の分布 $p(\boldsymbol{x}|\boldsymbol{u})$ を式 (3) と定義する．

$$p(\boldsymbol{x}|\boldsymbol{u}) = \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(\frac{-1}{2\beta^2} e_t^2\right) \tag{3}$$

図 1 に，GTM-AR における時系列データ $\boldsymbol{x}$ の生成モデルを示す.

ここで，潜在変数 $\boldsymbol{u}$ の分布 $p(\boldsymbol{u})$ を定義した場合，時系列データ $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ は式 (4) により得られる．

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{u}) p(\boldsymbol{u}) d\boldsymbol{u} \tag{4}$$

潜在変数 $\boldsymbol{u}$ の分布 $p(\boldsymbol{u})$ は，計算の容易さと自己組織化マップとの類似性を考慮し，潜在変数空間上に規則的に並べられた格子点 $\boldsymbol{u}_k$ を中心とする K 個のデルタ関数の和と定義する．

$$p(\boldsymbol{u}) = \frac{1}{K} \sum_{k=1}^{K} \delta\left(\boldsymbol{u} - \boldsymbol{u}_k\right) \tag{5}$$

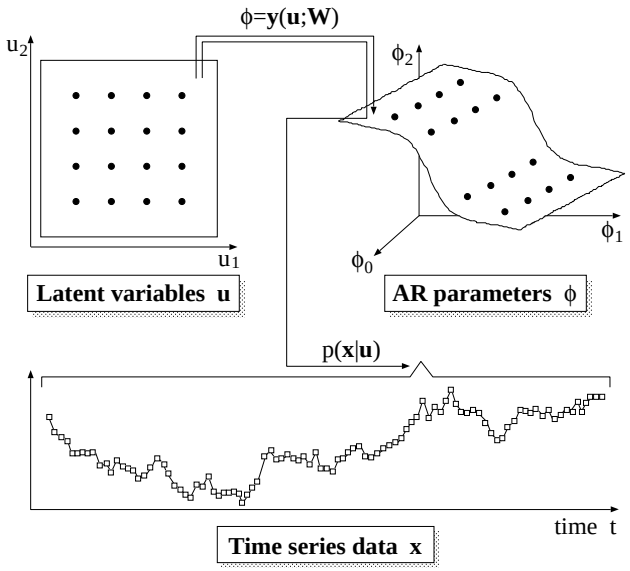

図 1: GTM-AR における時系列データの生成モデル ($L = 2, p = 2$ の場合)

式 (3) と式 (5) を式 (4) に代入することにより，時系列データ $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ は式 (6) と得られる．

$$p(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=p+1}^{T}\frac{1}{\sqrt{2\pi\beta^2}}\exp\left\{\frac{-1}{2\beta^2}e_{t,k}^2\right\},$$
$$e_{t,k} = x_t - \phi_{k,0} - \sum_{i=1}^{p}\phi_{k,i}x_{t-i} \qquad (6)$$

ここで，$\phi_{k,i} = y_i(\boldsymbol{u}_k; \boldsymbol{W})$ である．式 (6) は，混合係数 π_k を全て $1/K$ とした混合 AR 分布である．つまり，GTM-AR では時系列データ $\boldsymbol{x}$ の生成モデルとして制約条件付きの混合 AR モデルを仮定していると言える．

式 (6) より，時系列データ集合 $D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ の尤度関数 L は式 (7) と求めることができる．GTM-AR では，尤度関数 L を EM アルゴリズムを用いて最大化することにより，行列 $\boldsymbol{W}$, 分散 β^2 の推定を行う．

$$L(\boldsymbol{W}, \beta^2) = \prod_{n=1}^{N} p(\boldsymbol{x}_n) \qquad (7)$$

3.2 時系列データの可視化

GTM-AR では，時系列データ $\boldsymbol{x}$ が与えられたとした際の潜在変数 $\boldsymbol{u}$ の期待値を用いて可視化を行う．

$$\langle \boldsymbol{u}|\boldsymbol{x}\rangle = \int \boldsymbol{u}p(\boldsymbol{u}|\boldsymbol{x})d\boldsymbol{u} = \sum_{k=1}^{K}\boldsymbol{u}_k p(\boldsymbol{u}_k|\boldsymbol{x})$$

ここで，分布 $p(\boldsymbol{u}_k|\boldsymbol{x})$ はベイズの定理を用いることにより式 (8) と計算する．

$$p(\boldsymbol{u}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{u}_k)}{\sum_{k'=1}^{K} p(\boldsymbol{x}|\boldsymbol{u}_{k'})} \qquad (8)$$

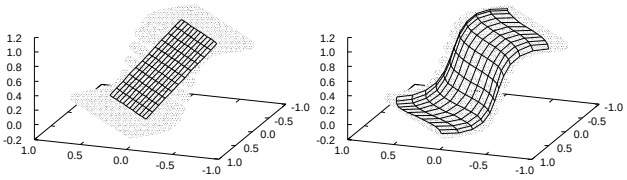

図 2: 関数 $\boldsymbol{y}(\boldsymbol{u}; \boldsymbol{W})$ を用いて 12×12 個の潜在変数 $\boldsymbol{u}_k$ を AR 係数空間上に写した結果 (左図：主成分分析による初期化後，右図：学習後)

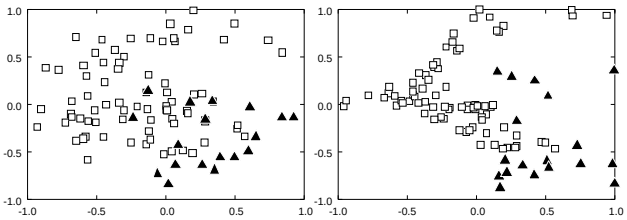

図 3: GTM-AR による MIT-BIH 不整脈データの可視化 (左図：主成分分析による初期化後，右図：学習後)

4 計算機実験

本実験では，潜在変数 $\boldsymbol{u}$ は $[-1, 1] \times [-1, 1]$ の 2 次元の領域に均等に $K = 12 \times 12$ 個配置し，基底関数の個数は $M = 19$ と設定した．

4.1 人工データを用いた計算機実験

以下で示される AR 係数を持つ 1000 個の AR(2) モデルを用意し，各 AR モデルから 1 個づつ合計で 1000 個の時系列データを用いて実験を行った．各 AR モデルの係数は，$\phi_0 = 1/(1 + \exp(10 \times (\phi_1 + \phi_2)))$ と設定し，係数 ϕ_1, ϕ_2 は $(0, -1)$, $(1, 0)$, $(0, 1)$, $(-1, 0)$ の 4 点を結ぶ菱形の領域から均等に設定した．

図 2 に，関数 $\boldsymbol{y}(\boldsymbol{u}; \boldsymbol{W})$ を用いて 12×12 個の潜在変数 $\boldsymbol{u}_k$ を AR 係数空間上に写した結果を実線で示す．図 2 より，AR 係数の真値（灰色の領域）を覆うように $\boldsymbol{y}(\boldsymbol{u}_k; \boldsymbol{W})$ が形成されていることが分かる．

4.2 心電図データを用いた計算機実験

MIT-BIH 不整脈データベース (`http://www.physionet.org/physiobank`) に含まれる心電図データを用いて実験を行った．図 3 に，正常波形 (▲)，不整脈波形 (□) を可視化した結果を示す．図 3 より，主成分分析を用いて行列 $\boldsymbol{W}$ を初期化 [1] した直後と比べ，学習後の方がよりデータの重なりを低く抑えていることが分かる．

参考文献

[1] M. Svensén (1998) “GTM: the generative topographic mapping.” PhD thesis, Aston University.

[2] G. Edward, P. Box, and G. M. Jenkins (1994) “Time Series Analysis: Forecasting and Control.” Prentice Hall PTR.