情報処理学会論文誌　データベース　Vol.16　No.1　1–13 (Jan. 2023)

時系列データの自動ネットワーク構造検出アルゴリズム

小幡 紘平[1,2,a)]　松原 靖子[1]　川畑 光希[1]　中村 航大[1,2]　櫻井 保志[1]

受付日 2022年6月8日, 採録日 2022年10月3日

概要：本論文では，ネットワーク構造を持つ多次元時系列データのためのパターン検出手法である NGL について述べる．NGL は，時間変化するネットワーク構造を持つ多次元時系列データが与えられたときに，その時系列データの中から重要なネットワーク構造を発見し，それらの情報を要約，表現する．具体的に，提案手法は，(a) 多次元時系列データからネットワーク構造に基づいた解釈性の高いクラスタを発見する．(b) その際に最適な分割点とクラスタ数を自動的に決定する．すなわち，事前情報の付与が必要ない．そして，(c) 自動決定アルゴリズムにより高精度なクラスタリングを実現する．人工データを用いた精度評価実験では最新の既存手法と比較して提案手法が大幅な精度向上を達成していることを明らかにした．また，実データを用いた実験では NGL が解釈性の高いクラスタを発見していることを確認した．

キーワード：時系列データ，ネットワーク構造，グラフィカルラッソ

Automatic Network Structure-based Clustering of Multivariate Time Series

Kohei Obata[1,2,a)]　Yasuko Matsubara[1]　Koki Kawabata[1]　Kota Nakamura[1,2]
Yasushi Sakurai[1]

Received: June 8, 2022, Accepted: October 3, 2022

Abstract: In this paper we present NGL, pattern mining algorithm for multiple time series data with underlying network structures. Our method has the following properties: (a) Interpretable: it provides interpretable network structures for the data; (b) Automatic: it determines the optimal cut points and the number of clusters automatically; (c) Accurate: it provides reliable clustering performance thanks to the automated algorithm. We evaluate our NGL algorithm on synthetic datasets, outperforming state-of-the-art baselines in terms of accuracy. And extensive experiments on real datasets demonstrate that NGL does indeed obtain interpretable network structure clusters.

Keywords: time series, network structure, graphical lasso

1. まえがき

車両走行センサ [1]，生体信号，ソーシャルネットワーク [2]，株価に代表される金融データ [3] など，さまざまなアプリケーションにおいて時系列データが生成される．これらの応用では，さまざまな特徴量を多次元時系列データとして扱い，データの構造理解や予測に有用な特徴量間の相関関係，すなわち，ネットワーク構造に基づくパターンを発見することが非常に重要な課題である．

一般に，実際に生成される時系列データは，複数の異なるネットワーク構造を持つことが多い．たとえば，車両走行センサデータの走行パターンはいくつかの代表的な運転行動（直進，右折，左折，減速，急ブレーキ，急旋回など）から構成される．ネットワーク構造の表現に効果的なグラフ理論 [4] によると，各センサをノード，センサ間の相関関係の強さをエッジとして表現することができる．右左折では，ハンドル角と左右加速度にエッジが，減速ではブレーキペダルストロークと前後加速度にエッジが形成されるだろう．しかし，ネットワーク構造の変化点や種類が

[1] 大阪大学産業科学研究所産業科学 AI センター
SANKEN, Osaka University, Ibaraki, Osaka 567–0047, Japan
[2] 大阪大学情報科学研究科
IST, Osaka University, Ibaraki, Osaka 567–0047, Japan
[a)] obata88@sanken.osaka-u.ac.jp

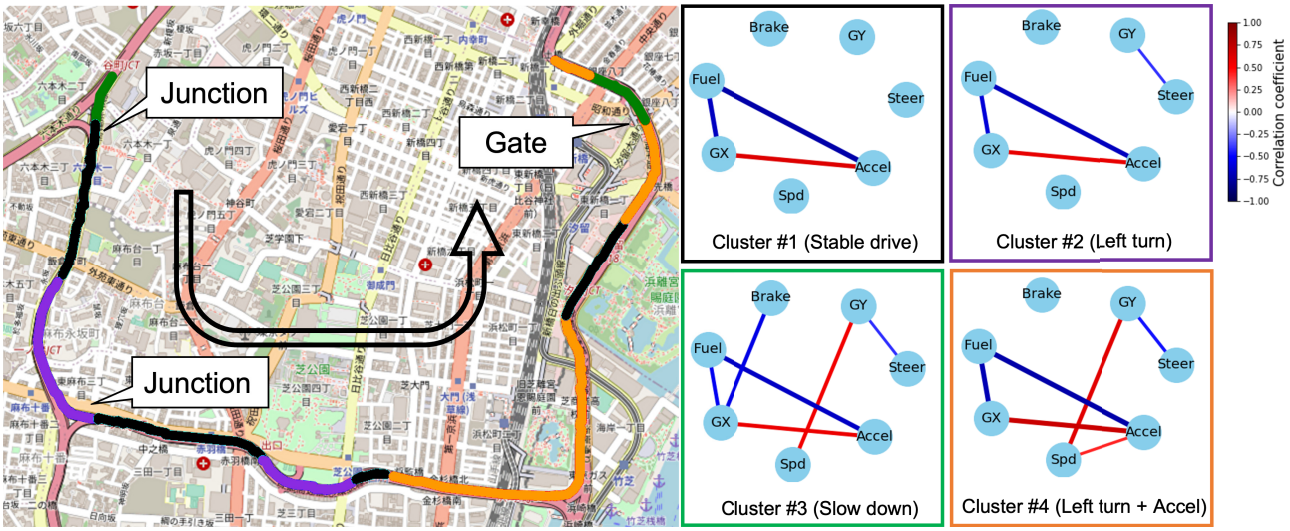


図 1　車両走行センサデータ：高速道路における NGL の出力例（GY = 左右加速度，Steer = ハンドル角，Accel = アクセルペダルストローク，Spd = 速度，GX = 前後加速度，Fuel = 燃費，Brake = ブレーキペダルストローク）

Fig. 1　Clustering result of NGL using highway automobile datasets.

既知であることは稀であり，人手で設定した閾値などによるパターン分割は難しく，現実的でない．また，データが本来持つネットワーク構造を無視したパターン分割はパターン変化点や依存関係の誤検出につながる．そこで本研究では，大規模多次元時系列データの中から典型的なネットワーク構造を自動的に検出するためのクラスタリングアルゴリズムである NGL を提案する．

本論文で扱う問題は以下のとおりである．

問題：大規模多次元時系列データ X が与えられたとき，X を表現する動的ネットワーク構造を抽出する．

より具体的には，(1) X の中のネットワーク構造に基づいた分割点を発見し，部分シーケンス集合（セグメント）に分割する．(2) 解釈性が高く，共通のネットワーク構造を持つセグメント集合（クラスタ）を見つけ，最適なクラスタ数を自動的に決定する．

具体例．図 1 は車両走行センサデータと NGL の出力結果例である．この車両走行センサデータは高速道路を走行した際の 7 つのセンサデータ値から構成される．図 1 左部は地図にクラスタをプロットしたものであり，同一のクラスタに含まれるセグメントは同一の色で表現されている．左部は各クラスタのネットワーク構造を示しており，ノードが変数，エッジが変数間の相関係数を表している．提案手法は，安定走行，カーブ，減速区間のパターンを発見していることが地図とネットワーク構造から分かる．ここで最も重要なこととして，NGL はこれらの走行パターンに関する事前知識を必要とせず，クラスタから運転行動を推測できるネットワーク構造を出力することができる．実験結果についての詳細は，6 章において示す．

1.1　自動ネットワーク構造抽出手法の重要性

時系列データを対象とした教師なしクラスタリング手法は，数多く存在する [5]．しかし，先行研究の中にネットワーク構造と基にしたクラスタリングを行い，かつ最適なクラスタ数を自動で発見する手法はない．多くの手法は実値の距離に基づいたクラスタを発見するため，クラスタから得られる情報は少ない [6], [7]．そのため，事前知識なしではクラスタの解釈が難しい．一方，ネットワーク構造に基づいたクラスタリングでは，データの背後に潜む変数間の相互関係が明らかになり，変数間の関係性をモデル化することが可能となる．さらに，多くの時系列クラスタリング手法はクラスタ数を事前に指定する必要がある [8], [9]．しかしながら，クラスタ数を指定するには事前知識が必要であり，未知のデータやビッグデータの解析には適していない．また，クラスタ数を事前に指定する必要があるモデルでは，既存クラスタで表現することができない観測値に対し，新たなクラスタを動的に生成できないため，クラスタの解釈性が低下する懸念がある．提案手法はデータに応じてクラスタを動的に生成することができる．最適なクラスタ数を自動で決定することで，リアルタイム処理への拡張や，予期せぬクラスタの発見ができ，解釈性が高いクラスタの検出が可能となる．

1.2　本論文の貢献

本論文では，大規模多次元時系列データから動的ネットワーク構造を抽出するための効果的なアルゴリズムである NGL を提案する．NGL は以下の特長を持つ．

(a) 類似したネットワーク構造を持つ時系列パターン（クラスタ）の個数と種類を把握し，データから解釈性の高いクラスタを発見する．

(b) モデルを表現する新しい符号体系を用いることで，データについての事前情報の付与を必要とせず，時系列パターンの最適な分割点とクラスタ数を自動で発見する．
(c) 効率的かつ効果的なセグメント分割アルゴリズムを提案し，人工データを用いた実験において，提案手法が最新の既存手法より高精度に時系列パターンを検出することを示す．

2. 関連研究

関連研究は以下の 2 つに分類される．

パターン発見．時系列データの解析に関する研究はさまざまな分野で進められている [10], [11], [12]．中でも，時系列サブシーケンスのクラスタリングはデータを理解するために有用である．時系列データの教師なしクラスタリングの代表的な技術である，DTW（Dynamic Time Warping）[13] と K-menas は距離に基づいたクラスタリングを行い，データの構造よりも実値を比べることに焦点を置いている．Li ら [14] が提案した，DynaMMo は線形動的システム（LDS: Linear Dynamical System）に基づく手法で欠損を含む大規模時系列データ集合から時系列のパターンを発見できる．Wang ら [15] による pHMM（pattern-based hidden Markov model）は隠れマルコフモデル（HMM: Hidden Markov model）に基づく手法であり，時系列のセグメント化とクラスタリングのための動的モデルである．Mastubara ら [16] は多階層 HMM モデルを使ったパラメータフリーの手法として AutoPlait を提案している．これらの手法は，時系列の複雑な動的パターンを表現する能力はあるが，その一方で，ネットワーク構造を考慮していないため，クラスタの解釈には困難がともなう．

時系列ネットワーク推定．時系列情報を加味したネットワーク推定は経済データ，生体信号データの解析手法として研究されている [17]．グラフィカルラッソ [4] は静的なネットワーク推定手法であり，損失関数に ℓ_1 正則化項を加味することで解釈が容易なスパースなネットワーク構造が推定できる [18]．Hallac ら [19] は文献 [4] に時系列情報を考慮したネットワーク推定手法である TVGL（Time Varying Graphical Lasso）を提案し，Harutyunyan ら [20] は文献 [21] を改良した共分散行列推定手法として T-CorEx を提案した．Tomasi ら [22] は文献 [23] を時系列データに適応し，潜在状態を考慮した動的なネットワーク構造を推定する手法である，LTGL（Latent variable Time-varying Graphical Lasso）を提案した．これらの手法は，ネットワーク構造の時系列変化をモデル化しており，前後のネットワーク構造を比較することで変化点の検知は可能だが，クラスタリングする能力はない．ネットワーク構造を基にしたクラスタリング手法として Hallac ら [8] が提案した，TICC（Toeplitz Inverse Covariance-based Clustering）と Tozzo ら [9] が提案した，TAGM（Time Adaptive Gaussian Model）がある．TICC はマルコフランダムフィールド（MRF: Markov Random Field）とテプリッツ行列を用い変数間に内在する関係をとらえる手法であり，TAGM は HMM と混合ガウスモデル（GMM: Gaussian Mixture Model）を融合した手法である．これらの手法は各サブシーケンスのネットワーク構造に応じたクラスタを発見する．これにより，クラスタに解釈性を持たせ，ほかの従来のクラスタリング手法では発見できなかったパターンを発見することができる．両者とも，モデルにグラフィカルラッソを組み込み変数間の相互作用を基にクラスタリングしているが，事前情報としてクラスタ数を指定しなければならない．つまり，提案手法のみが解釈性の高いクラスタリングと最適なクラスタ数を自動的に見つけるという特長を持つ．

3. 事前準備

ここでは本論文で必要な概念について定義を行う．また，**表 1** に主な記号と定義を示す．

3.1 問題定義

T 個の連続した観測値集合 $X = \{x_1, x_2, \ldots, x_T\}$ からなる p 次元時系列データの各時刻 i において，$|x_i| \geq 1$ 個の異なる観測値があるとする．$x_i \in \mathbb{R}^p$ は i 番目の観測値ベクトルであり，K 個の多変量正規分布のいずれか $x_i \sim N(0, \Sigma_k)$ から取得される．ネットワーク構造はグラフと同値であり，多次元データ $x_i \in \mathbb{R}^p$ から多変量正規分布 $x_i \sim N(0, \Sigma_k)$ が得られたとき，各変数がノードを表し，共分散行列 Σ_k がエッジを表す．共分散行列の対角成

表 1　主な記号と定義
Table 1　Symbols and definitions.

記号	定義
シーケンス	
T	時系列の長さ
x_i	時刻 i における観測値集合
$\|x_i\|$	時刻 i における観測値数
p	時系列の次元数
X	時系列データ
モデル	
θ_k	クラスタ k の逆共分散行列
Θ	逆共分散行列集合
$\mathcal{F}_k$	クラスタ k への割当て
$\mathcal{F}$	割当て集合
M_k	k 番目のクラスタのモデルパラメータ
M	K 個のクラスタのモデルパラメータ集合
K	クラスタ数
コスト関数	
$< M >$	M のモデル表現コスト
$< X\|M >$	M による X の符号化コスト
$< X; M >$	M による X の総コスト

分がすべて 1 であるとき，非対角成分の値は変数間の相関係数と等しい．本研究の目的は X を K 個のクラスタ集合に分割する割当て $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_K\}$ を発見することである．ここで，$\mathcal{F}_k$ は $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ に基づいた $X_k \subset X\,(s.t.\, k = 1, 2, \ldots K)$ のクラスタ k への割当てである．それゆえ，クラスタ k に属する X_k の簡潔な記述モデルを $M_k = \{\theta_k, \mathcal{F}_k\}$ とすると，パラメータ集合は $M = \{M_1, M_2, \ldots, M_K\}$ となる．

3.2　グラフィカルラッソ

まずはじめに，静的なネットワーク構造 θ_i を推定するグラフィカルラッソについて述べる．グラフィカルラッソは，多変量正規分布 $N(0, \Sigma)$ を仮定した対数尤度関数に基づく損失関数に，ℓ_1 正則化項を加えた関数を最小化することによりスパースな逆共分散行列を推定する手法である．具体的には，以下の式を最適化する：

$$\begin{aligned} &\text{minimize}_{\,\theta \in S^p_{++}} \; -ll(x_i, \theta_i) + \lambda ||\theta_i||_{od,1}, \qquad (1)\\ &ll(x_i, \theta_i) = |x_i|(\log\det\,\theta_i - Tr(S_i\theta_i)), \end{aligned}$$

ただし，$||\cdot||_{od,1}$ は対角成分を除いた ℓ_1 ノルムである．正則化ハイパーパラメータ $\lambda \geq 0$ により損失関数と ℓ_1 正則化項のバランスを調整することで，スパース性を制御する．$ll(x_i, \theta_i)$ は対数尤度関数である．θ_i は正定値対称行列（S^p_{++}）である．そして，S は観測値から計算される共分散行列 $(1/|x_i|)\sum_{j=1}^{|x_i|} x_j x_j^T$ であり，x_j は各観測値である．

3.3　TVGL 問題

TVGL [19] は動的なネットワーク構造を推定するために上記の問題 (1) を拡張し，時間情報を加味した逆共分散行列集合 Θ を推定する手法である．TVGL は以下の問題を最適化する：

$$\begin{aligned} &\text{minimize}_{\,\theta_i \in S^p_{++}} \sum_{i=1}^{T} -ll(x_i, \theta_i) + \lambda ||\theta_i||_{od,1}\\ &+ \beta \sum_{i=2}^{T} \psi(\theta_i - \theta_{i-1}), \qquad (2) \end{aligned}$$

ただし，λ はネットワークのスパース性を制御するハイパーパラメータである．β は隣接する逆共分散行列の類似度を決定するハイパーパラメータである．罰則項に含まれる関数 ψ は θ_i と θ_{i-1} の類似性を定義する．多種の ψ によってネットワーク構造の時間変化の類似性を制御することができる．具体的には，ℓ_1 罰則項：$\psi(X) = \sum_{i,j} |X_{i,j}|$，$\ell_2$ 罰則項：$\psi(X) = \sum_j ||X_j||_2$，ラプラシアン罰則項：$\psi(X) = \sum_{i,j} X^2_{i,j}$，などがあげられる．以下，本論文では，ラプラシアン罰則項を用いる．式 (2) は凸最適化問題の最適化法である交互方向乗数法（ADMM: Alternating Direction Method of Multipliers）によって解ける．まず，補助変数 $Z = \{Z_0, Z_1, Z_2\} = \{(Z_{1,0}, \ldots, Z_{T,0}), (Z_{1,1}, \ldots, Z_{T-1,1}), (Z_{2,2}, \ldots, Z_{T,2})\}$ を用意し，式 (2) の変数を補助変数に置き換えることで変数を分離する：

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{T} -ll(x_i, \theta_i) + \lambda ||Z_{i,0}||_{od,1}\\ &+ \beta \sum_{i=2}^{T} \psi(Z_{i,2} - Z_{i-1,1})\\ &\text{subject to } Z_{i,0} = \theta_i,\ \theta_i \in S^p_{++} \text{ for } i = 1, \ldots, T\\ &\qquad (Z_{i-1,1}, Z_{i,2}) = (\theta_{i-1}, \theta_i) \text{ for } i = 2, \ldots, T. \end{aligned}$$

すると，拡張ラグランジュ関数は次のようになる：

$$\begin{aligned} &L_\rho(\Theta, Z, U)\\ &= \sum_{i=1}^{T} -ll(x_i, \theta_i) + \lambda ||Z_{i,0}||_{od,1}\\ &\quad + \beta \sum_{i=2}^{T} \psi(Z_{i,2} - Z_{i-1,1})\\ &\quad + (\rho/2) \sum_{i=1}^{T} \left(||\theta_i - Z_{i,0} + U_{i,0}||^2_F - ||U_{i,0}||^2_F \right)\\ &\quad + (\rho/2) \sum_{i=2}^{T} (||\theta_{i-1} - Z_{i-1,1} + U_{i-1,1}||^2_F - ||U_{i-1,1}||^2_F\\ &\quad + ||\theta_i - Z_{i,2} + U_{i,2}||^2_F - ||U_{i,2}||^2_F), \end{aligned}$$

ただし，双対変数 $U = \{U_0, U_1, U_2\} = \{(U_{1,0}, \ldots, U_{T,0}), (U_{1,1}, \ldots, U_{T-1,1}), (U_{2,2}, \ldots, U_{T,2})\}$, ADMM の罰則項 $\rho > 0$ とする．r を更新数とすると，ADMM の更新式は以下で表される：

$$\begin{aligned} \Theta^{r+1} &:= \arg\min_{\theta \in S^p_{++}} L_\rho(\Theta, Z^r, U^r)\\ Z^{r+1} &:= \arg\min_{\theta \in S^p_{++}} L_\rho(\Theta, Z^r, U^r)\\ U^{r+1} &:= \arg\min_{\theta \in S^p_{++}} L_\rho(\Theta, Z^r, U^r) \end{aligned}$$

詳細は文献 [19] を参照されたい．TVGL は，前後の逆共分散行列 θ_i と θ_{i-1} を比較することで分割点を発見することが可能であるが，クラスタを発見することはできない．本手法では各時刻の逆共分散行列を求める際，TVGL を最適化手法として用いる．

4.　提案手法

前章では TVGL がどのように逆共分散行列集合 Θ を求めるかを取り扱った．本章では，(a) グラフィカルラッソモデルの分割基準をどのように設定するか，(b) 最適な分割点はどのように決定すればよいか，(c) 最適なクラスタはどのように決定すればよいか，を解決するモデルを提案

する．提案モデルは以下の 3 つのアイデアに基づく．

- モデル表現コスト：最適な分割点とクラスタの発見のために，最小記述長（MDL: minimum description length）の概念を用いる．MDL は情報理論に基づくモデル選択基準の 1 つであり，直感的には，データをより圧縮できれば良いモデルと見なすことができる．本論文の目的を解決するために，新しい符号体形をグラフィカルラッソモデルに対して定義する．
- CutPointSearch：一般的なボトムアップアルゴリズム [24] を改良し，時系列データを扱うために適したアルゴリズムを提案する．最初に設定した分割点による小サイズのセグメントをコスト制限を満たす隣接セグメントと反復的にマージすることで最適な分割点を求める．
- NGL：EM アルゴリズムを用い，CutPointSearch で発見したセグメントを最適なクラスタに割り当てる．また，コスト関数に基づき最適なクラスタ数を自動的に決定する．

4.1　特徴抽出とデータ圧縮

ここでは，大規模時系列データを表現するための符号化スキームを導入する．簡潔に表すと，MDL を用いてデータを表現するために必要なグラフィカルラッソモデルの最小数を求めることを目標とする．データ X が与えられたときのモデルのよさは次の式で表現できる：$< X; M >= \alpha \cdot < M > + < X|M >$．ここで，$< M >$ はモデル M を表現するためのコストを示し，$< X|M >$ は M が与えられたときの X の符号化のコストを示す．ハイパーパラメータ $\alpha > 0$ によってモデル表現コストと符号化コストのバランスを調整し，モデルの複雑さを制御する．

4.1.1　モデル表現コスト

モデル M の表現コストは以下の要素の総和から構成される．

- クラスタの総数 K：$\log^*(K)$ [*1]
- 各クラスタの観測値数：$\sum_{k=1}^{K} \log^*(|\mathcal{F}_k|)$
- 各クラスタの平均値 $p \times 1$：$\sum_{k=1}^{K}(p \times c_F)$
- 各クラスタの逆共分散行列 $p \times p$：$\sum_{k=1}^{K} |\theta_k|_{\neq 0}(2\log(p) + c_F) + \log^*(|\theta_k|_{\neq 0})$

ここで，$|\cdot|_{\neq 0}$ は行列の非 0 要素の数を，c_F は浮動小数点のコストを示す[*2]．

4.1.2　データ記述長

先述のとおり，本論文ではグラフィカルラッソモデルを用いてデータ X のパターンを表現するが，ここで重要なのは，推定したモデルが X を正しく表現しているかを判断する指標の導入である．ハフマン符号 [25] を用いた情報圧縮では，モデル M が与えられた際の X の符号化コストを負の対数尤度を用いて次のように表現することができる：

$$< X|M >= -\log_2 P(X|M).$$

ここで $P(X|M)$ は X の尤度を表す．したがって X と K 個のクラスタのモデルパラメータ M が与えられたとき，符号化コストは次のように表される：

$$< X|M > = -\sum_{k=1}^{K} ll(X_k, \theta_k).$$

4.1.3　符号化コスト関数

まとめると，モデルパラメータ集合 M が与えられたとき X の符号長は以下のようになる：

$$< X; M >= \alpha \cdot < M > + < X|M > . \quad (3)$$

したがって，本論文の次の目標は上記のコスト関数 $< X; M >$ を最小化するようなモデルパラメータ集合 M を発見することである．

4.2　セグメント分割アルゴリズム

前項では，モデルパラメータ集合 M を与えられたときのデータ X を表現するためのコスト関数として式 (3) を示した．続いての問題は式 (3) を最小化する最適な分割点を発見することである．同じモデルで表現するべきセグメントの候補は多数あり，すべてを検討することは組合せ爆発が起こるため難しい．本項では，ボトムアップアルゴリズムを改良することで少数の候補から効率良く最適な分割点を発見する．具体的には，以下の 2 つのアルゴリズムを提案する．

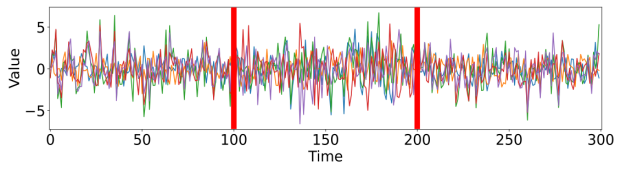
(1) MergeSegment（inner loop）：分割点が与えられたときに，コスト関数に基づき隣接セグメントとマージし，分割点を更新する．
(2) CutPointSearch（outer loop）：分割点の更新が止まるまで，与えられた分割点に基づいた逆共分散行列を計算する．

図 2 は CutPointSearch の処理のながれである．データ X は 2 個の多変量正規分布から取得され，正しい分割点は図 2 (a) である．CutPointSearch は多数の初期分割点，図 2 (b) から開始し，図 2 (c) のようにコスト関数に基づき隣接セグメントをマージしながら，最終的にデータ X の最適な分割点，図 2 (d) を発見する．各イテレーションにおいて局所的にコスト関数が減少するように隣接セグメントをマージする．

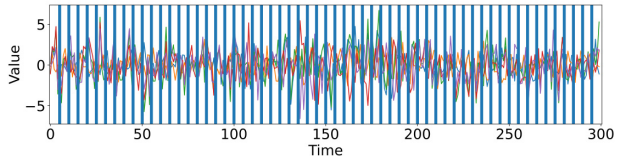
4.2.1　MergeSegment

隣接するセグメントが同じクラスタに属する傾向があると仮定して，MergeSegment によって分割点を更新する問題を考える．アルゴリズム 1 は MergeSegment の処理を示す．分割点 $cp = \{c_0, c_1, \ldots, c_m\}$ と各セグメントの逆共分散行

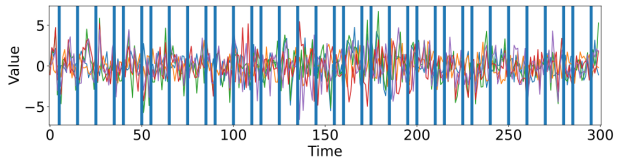
[*1] ここで，$\log^*$ は整数のユニバーサル符号長を表す．
[*2] 本論文では 4 × 8 ビットとする．



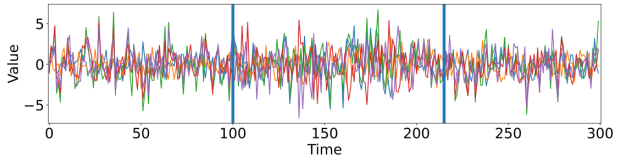
(a) K: 2, pattern: ”1, 2, 1”



(b) Iteration: 0, # segments: 60



(c) Iteration: 1, # segments: 36



(d) Iteration: 6, # segments: 3

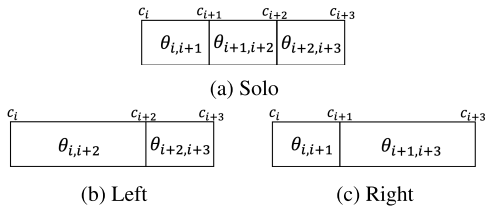
図 2　CutPointSearch の概要図

Fig. 2　Overview of the workflow of CutPointSearch.

Algorithm 1 MergeSegment$(\Theta_S, \Theta_E, \Theta_O, cp, X)$

Input: Each covariance Θ, initial cut point set cp, and bundle X
Output: Updated cut point set cp_{new}
$id = 0$, $cp_{new} = \phi$;
while $id < length(X)$ **do**
  if id is even **then**
    $\theta_{Left} = \theta_O$; $\theta_{Right} = \theta_E$;
    $id_{Left} = \lfloor id/2 \rfloor$; $id_{Right} = \lfloor id/2 \rfloor + 1$;
  else if id is odd **then**
    $\theta_{Left} = \theta_E$; $\theta_{Right} = \theta_O$;
    $id_{Left} = \lfloor id/2 \rfloor + 1$; $id_{Right} = \lfloor id/2 \rfloor + 1$;
  end if
  $C_{solo} =< X; \Theta_S[id] > + < X; \Theta_S[id+1] > + < X; \Theta_S[id+2] >$;
  $C_{left} =< X; \Theta_{Left}[id_{Left}] > + < X; \Theta_S[id+2] >$;
  $C_{right} =< X; \Theta_S[id] > + < X; \Theta_{Right}[id_{Right}] >$;
  if $min(C_{solo}, C_{left}, C_{right}) = C_{solo}$ **then**
    $cp_{new} = cp_{new} \cup cp[id]$; $id+ = 1$;
  else if $min(C_{solo}, C_{left}, C_{right}) = C_{left}$ **then**
    $cp_{new} = cp_{new} \cup cp[id+1]$; $id+ = 2$;
  else if $min(C_{solo}, C_{left}, C_{right}) = C_{right}$ **then**
    $cp_{new} = cp_{new} \cup cp[id], cp[id+2]$; $id+ = 3$;
  end if
end while
return cp_{new};

列集合 $\Theta_S = \{\theta_{,c_0}, \theta_{c_0,c_1}, \ldots, \theta_{c_m,}\}$ が与えられたとする．m は分割点の個数で，$m+1$ はセグメント数となる．そして，



(a) Solo　(b) Left　(c) Right

図 3　MergeSegment で比較される 3 つの分割点の候補の概略図．これらの分割点の候補の MDL コストを比較する

Fig. 3　Illustration of the three candidates of cut points. We compare each MDL cost of these candidates of cut points.

奇数/偶数番目の分割点からなる各セグメントの逆共分散行列集合 $\Theta_O = \{\theta_{,c_1}, \theta_{c_1,c_3}, \ldots\}$ と $\Theta_E = \{\theta_{,c_0}, \theta_{c_0,c_2}, \ldots\}$ が与えられたとする．X_{c_i,c_j}，M_{c_i,c_j}，θ_{c_i,c_j}，はそれぞれ，分割点 c_i から c_j までのセグメントにおける，観測値，パラメータ集合，逆共分散行列とする．ここでの目標はセグメントを隣接セグメントとマージするかを決定することである．シーケンスを $id = 0$ から時系列順に処理していく．新しい分割点の候補は，つねに 3 つ存在する．**図 3** に $id = i$ の場合の新しい分割点の候補を示す．(a) Solo は 3 つのセグメントから，(b) Left，(c) Right は片一方のセグメントがマージされた 2 つのセグメントから構成される．それぞれのコスト関数は以下になる：

$$\begin{aligned}
&\text{(a) } \alpha \cdot (< M_{c_i,c_{i+1}} > + < M_{c_{i+1},c_{i+2}} > + < M_{c_{i+2},c_{i+3}} >) \\
&\quad + < X_{c_i,c_{i+1}} | M_{c_i,c_{i+1}} > + < X_{c_{i+1},c_{i+2}} | M_{c_{i+1},c_{i+2}} > \\
&\quad + < X_{c_{i+2},c_{i+3}} | M_{c_{i+2},c_{i+3}} >, \\
&\text{(b) } \alpha \cdot (< M_{c_i,c_{i+2}} > + < M_{c_{i+2},c_{i+3}} >) \\
&\quad + < X_{c_i,c_{i+2}} | M_{c_i,c_{i+2}} > + < X_{c_{i+2},c_{i+3}} | M_{c_{i+2},c_{i+3}} >, \\
&\text{(c) } \alpha \cdot (< M_{c_i,c_{i+1}} > + < M_{c_{i+1},c_{i+3}} >) \\
&\quad + < X_{c_i,c_{i+1}} | M_{c_i,c_{i+1}} > + < X_{c_{i+1},c_{i+3}} | M_{c_{i+1},c_{i+3}} > .
\end{aligned}$$

上記の 3 つのコスト関数を比較し，最小コストを示したものを新しい分割点として更新する．(b) が最小コストを示した場合，c_{i+2} が新しい cp に加えられる．(a) が最小コストを示した場合，cp は更新前後で変化がない．この処理をすべてのセグメントについて繰り返す．

4.2.2　CutPointSearch

ここで扱う問題は，データ X の最適な分割点を発見する問題である．アルゴリズム 2 に CutPointSearch の手順を示す．データ X と初期分割点 cp が与えられたとする．初期分割点はユーザが任意の特定間隔で設定する．TVGL の最適化手法を用い，各セグメントと奇数/偶数番目の分割点からなる各セグメントの逆共分散行列集合 Θ_S，Θ_O，Θ_E を計算する．すべての Θ を得たら，MergeSegment アルゴリズムにより，分割点を更新する．この処理を分割点の更新が止まるまで繰り返す．

Algorithm 2 CutPointSearch(X, cp)

Input: Bundle X and initial cut point set cp
Output: Optimal cut point set cp
repeat
  $\Theta_{Single} = \text{TVGL}(X, cp)$;
  $\Theta_{Even} = \text{TVGL}(X, cp[0::2])$; /* even-numbered */
  $\Theta_{Odd} = \text{TVGL}(X, cp[1::2])$; /* odd-numbered */
  $cp = \text{MergeSegment}(\Theta_{Single}, \Theta_{Even}, \Theta_{Odd}, cp, X)$;
until convergence;
return cp;

Algorithm 3 NGL(X, cp)

Input: Bundle X, initial cut point set cp
Output: Cluster parameters Θ and cluster assignments $\mathcal{F}$
$cp_{opt} = \text{CutPointSearch}(X, cp)$; $K = 1$;
while improving the total cost $< X; M >$ **do**
  $\Theta = \text{ModelInitialization}(cp_{opt}, K)$;
  repeat
    $\mathcal{F} = \text{AssignToCluster}(X, \Theta, cp_{opt})$; /* E-step, Equation (4) */
    $\Theta = \text{GraphicalLasso}(X, \mathcal{F})$; /* M-step */
  until convergence;
  Compute $< X; M >$; // $M = \{\Theta, \mathcal{F}\}$
  $K = K + 1$;
end while
return $M = \{\Theta, \mathcal{F}\}$;

4.2.3　理論的な分析

補助定理 1　提案手法の計算コストは最小で $O(m \cdot p^3)$，最大で $O(m^2 \cdot p^3)$ である．

証明 1　NGL の計算コストの大部分は CutPointSearch のイテレーション回数と共分散行列集合を推定する計算コストによる．データの次元数が p，初期分割点の数が m 個あり最終的に分割点の数が 0 個になる場合を考える．このとき，共分散行列行列を推定する計算コストは $O(m \cdot p^3)$ である [19]．CutPointSearch アルゴリズムのイテレーション回数は，最大 m 回であり，この場合 m が 1 つずつ減少する．最小は $\log_2 m$ 回であり，この場合 m が半減していく．よって，NGL の計算コストは最小で $O(m \cdot p^3)$，最大で $O(m^2 \cdot p^3)$ となる．

4.3　クラスタリングアルゴリズム

本論文の最終目標は，大規模時系列データの中から適切な数のネットワーク構造を自動的に抽出することである．最後に，最適な分割点から構成されるセグメントを適切なクラスタに割り当てる手法である NGL について述べる．

NGL は，各セグメントをクラスタに割り当てるために EM アルゴリズムを用いる．アルゴリズム 3 に NGL の手順を示す．最大のクラスタ数をセグメントの数とし，$K = 1, 2, 3, \ldots$ と変化させ，コスト関数 (3) を最小化する最適な K を求める．具体的には，各 K において以下の問題を最小化することで割当てを求める：

$$\arg\min_{\mathcal{F},\Theta} \sum_{k=1}^{K} -ll(X_k, \theta_k) + \lambda ||\theta_k||_{od,1}. \tag{4}$$

ここでは，すでに最適な分割点が得られており，各セグメントにはモデルを形成するための観測値数が十分にあるため複雑なアルゴリズムは必要ない．E ステップでは，対数尤度が最小となるよう，セグメントを適切なクラスタに割り当てる．M ステップでは，各クラスタについて割り当てられたデータの逆共分散行列を求める．ここで，各クラスタのパラメータの学習には式 (2) の $\beta = 0$（式 (1) と同等）とした TVGL の最適化法を用いる．

5.　評価実験

本章では，人工データに対する NGL のクラスタリング精度と計算コストの検証を行う．クラスタリング精度比較に用いられる典型的な実データはネットワーク構造に基づいた正解ラベルが与えられていない．一方で，人工データでは明確なネットワーク構造のあるデータが生成可能で，ネットワーク構造に基づいたクラスタリング精度の比較が可能である．人工データの生成，実験設計は文献 [8], [26] に従った．K 個のクラスタを持ち，各クラスタが多変量正規分布 $X \sim N(0, \Sigma)$ に従う，$X \in \mathbb{R}^5$ の人工データをランダムグラフに基づき生成した．ネットワーク構造に基づいたクラスタリング精度を評価するため，各クラスタの平均値は $\vec{0}$ とした．以下の手順で，各クラスタの逆共分散行列を作成した [26].

(1) 隣接行列 $A \in \mathbb{R}^{5\times5}$ を Erdös-Rényi モデルに従って作成する．全ノードペアについて，確率 20%でエッジを形成する．
(2) A の選ばれたエッジについて，$A_{ij} \sim Unif([-0.6, -0.3] \cup [0.3, 0.6])$ を設定する．また，A は対称行列 $A_{ij} = A_{ji}$ とする．
(3) Σ^{-1} を正定値行列とするために，$\Sigma^{-1} = A + (0.1 + |c|)I$ とする．$c = \lambda_{\min}(A)$ は A の最小固有値で，I は $p \times p$ の単位行列である．

次のような異なるセグメントの組合せの 4 つのデータセットについて実験を行った [8]（“1, 2, 1”,“1, 2, 3, 2, 1”,“1, 2, 3, 4, 1, 2, 3, 4”,“1, 2, 2, 1, 3, 3, 3, 1”）．それぞれのデータセットにつき 10 回実験を行い，$macro\text{-}F_1$ スコアの平均と標準偏差を記録した．$macro\text{-}F_1$ スコアは，適合率（Precision）と再現率（Recall）の調和平均を各クラスタについて求め，平均したもので，1 に近い値は高いクラスタリング精度を意味する[*3]．サンプル数について述べる場合は，各セグメントごとのサンプル数と同意である（たとえば，“1, 2, 1” において，サンプル数を 100 とした場合，100 サンプルごとに θ_1，θ_2 からサンプルが生成された，計 300

[*3] $macro\text{-}F_1 = \frac{1}{K}\sum_i^K \frac{1}{1/presicion_i + 1/recall_i}$

表 2　4 つの異なるデータセットにおける NGL と比較手法の *macro-F*$_1$ スコアによるクラスタリング精度（高いほど高精度）

Table 2　Macro-F1 score of clustering accuracy for four different temporal sequences, comparing NGL with state-of-the-art methods (higher is better).

Model	**NGL**	TAGM（KDD'21）	TICC（KDD'18）	AutoPlait（SIGMOD'14）	NGL no-cps
1, 2, 1	**0.93 ± 0.05**	0.83 ± 0.25	0.85 ± 0.26	0.41	0.62 ± 0.13
1, 2, 3, 2, 1	**0.96 ± 0.03**	0.74 ± 0.21	0.89 ± 0.18	0.20	0.66 ± 0.15
1, 2, 3, 4, 1, 2, 3, 4	**0.94 ± 0.03**	0.78 ± 0.26	0.82 ± 0.21	0.11	0.66 ± 0.11
1, 2, 2, 1, 3, 3, 3, 1	**0.93 ± 0.05**	0.89 ± 0.17	0.83 ± 0.26	0.19	0.62 ± 0.07

サンプルのデータとなる）．

5.1　提案手法のクラスタリング精度

はじめに，与えられた人工データに対する提案手法のクラスタリング精度を検証するために，最新の時系列クラスタリング手法と比較する．TICC [8]，および TAGM [9] はネットワーク構造に基づいたクラスタリングを行う手法である．TICC にはスパース性を制限するハイパーパラメータ λ と隣接ポイントを同じクラスタに割り当てない際の罰則コストであるハイパーパラメータ β がある．TAGM にはスパース性を制限するハイパーパラメータ λ のみ存在する．さらに，これらの手法はクラスタ数の指定が必要であるため，正しいクラスタ数を与え実験した．AutoPlait [16] は多階層 HMM ベースの自動クラスタリングアルゴリズムである．また，CutPointSearch の効果を検証するため，NGL から CutPointSearch を除いた NGL no-cps とも比較した．NGL と NGL no-cps のハイパーパラメータは人工データでの実験を通して，初期分割点の幅を 5，$\alpha = 1$ と設定した．なお，各手法のハイパーパラメータは，本実験に使用したデータセットとは別に用意したサンプル数 100 のデータセットを用いた実験において，平均精度が最も高いものを使用した．

5.1.1　多種類の人工データにおけるクラスタリング精度

4 つのデータセットについてサンプル数 100 で実験し，クラスタリング精度を *macro-F*$_1$ スコアで比較した結果を**表 2** に示す．本手法がすべてのデータセットにおいて最も高い平均精度であり，最も低い標準偏差を記録した．AutoPlait はネットワーク構造を考慮しないため，クラスタを発見できなかった．TICC と TAGM は正しいクラスタ数を与えられたにもかかわらず，平均精度において NGL より 10%以上低かった．NGL no-cps の結果から，CutPointSearch で大きな塊のセグメントを発見することが重要であることが分かる．つまり，CutPointSearch により，初期分割点時の隣接セグメントが同一クラスタに割り当てられやすくなるため，クラスタリング精度が上がる．

5.1.2　サンプル数を変化させたときのクラスタリング精度

サンプル数の増加に対して精度を維持することは大規模時系列データを扱うにあたり重要である．“1, 2, 3, 4, 1, 2, 3, 4” を例にとりサンプル数 25～10,000 まで増加させサンプル数による精度への影響を評価した．**図 4** はサンプル数に対する *macro-F*$_1$ スコアをプロットした結果である．NGL は他手法と比較して，サンプル数が比較的少ない 25～100 において安定して高い精度を示している．これは MDL の定式化によって，動的なクラスタ数の推定が可能となる MergeSegment による効果であると考えられる．一方で，NGL はサンプル数が 2,500 を超えると精度の低下が見られる．サンプル数の増加により，正解クラスタとのモデルの差が大きい部分シーケンスが多く生成されることで，余分なクラスタが多く生成されるためである．NGL-20 は初期分割点の幅をサンプル数の 20 分の 1 に設定した手法である．この設定により，サンプル数の増加にともない，1 クラスタのサンプル数の下限が増加する．これにより，サンプル数 2,500 以上の範囲においても安定した精度を示している．これは，各セグメントにおいてネットワーク構造がより正確に推定されるため，余分なクラスタが生成されにくくなるためである．このように，初期分割点の幅を調節することで，提案手法は大規模時系列データにおいても高精度なクラスタリングをすることができる．

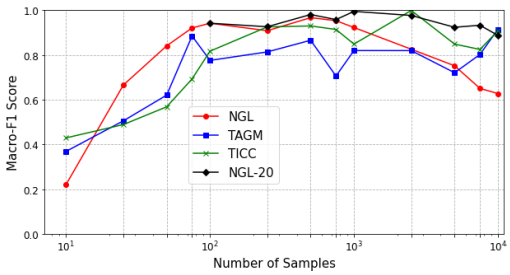


図 4　NGL と 2 つの比較手法におけるサンプル数に対する *macro-F*$_1$ スコア

Fig. 4　Plot of clustering accuracy macro-F1 score vs. number of samples for NGL and two other state-of-the-art methods.

5.1.3　次元数を変化させたときのクラスタリング精度

本手法の符号化コスト関数はデータの次元数による影響を受ける．“1, 2, 3, 4, 1, 2, 3, 4” を例にとりサンプル数 100 で次元数 p を 5～50 まで増加させ，次元数による精度への影響を評価した．**図 5** は次元数に対する *macro-F*$_1$ ス

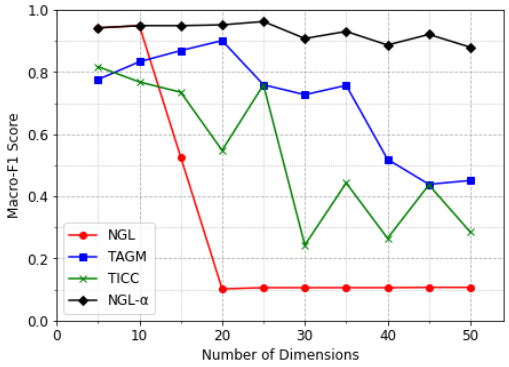


図 5　NGL と 2 つの比較手法における次元数に対する *macro-F*$_1$ スコア

Fig. 5　Plot of clustering accuracy macro-F1 score vs. number of dimensions for NGL and two other state-of-the-art methods.

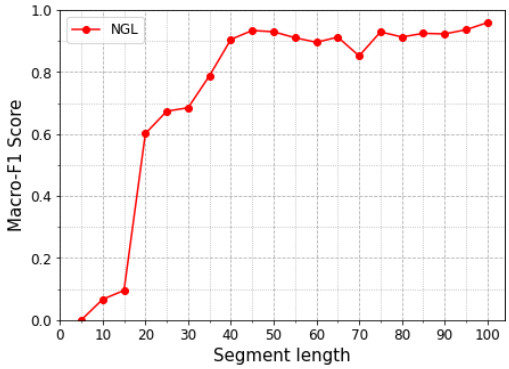


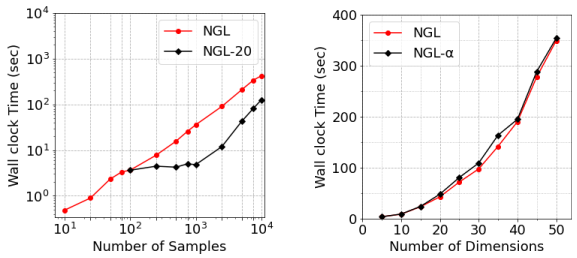
図 6　NGL の検出可能クラスタのサンプル数に対する *macro-F*$_1$ スコア

Fig. 6　Plot of clustering accuracy macro-F1 score vs. number of target cluster samples for NGL.

コアをプロットした結果である．NGL では次元数 15 以上の場合において精度が低下している．次元数の増加にともない，クラスタ間のモデルの差が小さくなるため，クラスタの検出が困難になる．そのため，NGL-α ではモデルの複雑さを制御するため，次元数 10 以上において $\alpha = 100/p^2$ に設定した．これにより，上記の生成方法による人工データでは，次元数 50 までの範囲において比較的高い精度を示している．このように，α を適切に設定することで次元数の増加に対してロバストにモデルの推定を行うことができる．

5.1.4　クラスタ検出可能なサンプル数の検討

提案手法は他手法と比較して，サンプル数 25〜100 において特に有効であると示した．サンプル数を 1 つのクラスタのみ減らすことで，クラスタとして検出するのに必要なサンプル数を検討する．“1, 2, 1” を例にとりクラスタ 1 のサンプル数を 300，クラスタ 2 のサンプル数 5〜100 まで増加させ，精度を評価した．**図 6** はクラスタ 2 のサンプル数に対するクラスタ 2 の *macro-F*$_1$ スコアをプロットした結果である．クラスタ 2 のサンプル数 15 以下ではクラスタ 2 をほとんど検出できていないが，クラスタ 2 のサンプル数 40 以上では高精度で検出している．よって，提案手法が上記の人工データのクラスタを高精度で検出するのに必要なサンプル数は 40 以上であることが分かる．



(a) Sample　(b) Dimension

図 7　NGL のサンプル数 (a) と次元数 (b) に対する計算コスト

Fig. 7　Plot of wall clock time vs. number of samples (a) and dimensions (b) for NGL.

5.2　提案手法の計算コスト

ここでは，提案手法の計算コストについて検証する．補助定理 1 では NGL の計算コストが最小で $O(m \cdot p^3)$，最大で $O(m^2 \cdot p^3)$ であることを示した．**図 7** は “1, 2, 3, 4, 1, 2, 3, 4” のサンプル数 n，次元数 p を変化させたときの NGL の計算コストを示す．NGL は初期分割点の幅が一定のため，サンプル数と初期分割点の数 m は比例する．図 7 (a) から NGL の計算コストがサンプル数に対し線形であることが分かる．また，NGL の計算コストが最大になることは稀であり，多くの場合において最小に近くなると考えられる．NGL-20 は初期分割点の数が一定である．NGL-20 の計算コストは $n \leq 1000$ において一定だが，$n > 1000$ においてはオーバヘッドのため線形になっている．一方，次元数の増加は共分散行列集合を推定する計算コストに影響を与える．提案手法の計算コストは，図 7 (b) から，次元数 p に対し $O(p^3)$ となっている．これは補助定理 1 において示した結果と一致する．

6.　ケーススタディ

本章では，実データを対象とした実験により，NGL が意味のあるネットワーク構造を教師なしで発見可能であることを示す．

6.1　ハイパーパラメータ選択基準

本章で扱う実データには正解クラスタが存在しない．また，データによって特徴量の次元数やクラスタ間のモデルの差が異なる．このような場合においては，ユーザの直感に合う，期待どおりの分割結果が求められる．提案手法は特定のデータに対してモデルの符号化コストを適切に設定することで，クラスタ数の自動決定が可能となる．その一方で，提案手法にはさまざまなデータに対応するために，

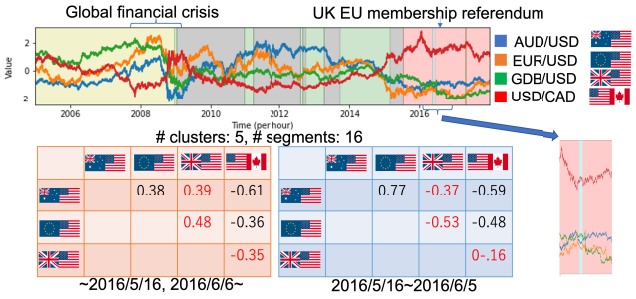


図 8　外貨交換レートデータにおける NGL の出力例
Fig. 8　Clustering result of NGL using currency datasets.

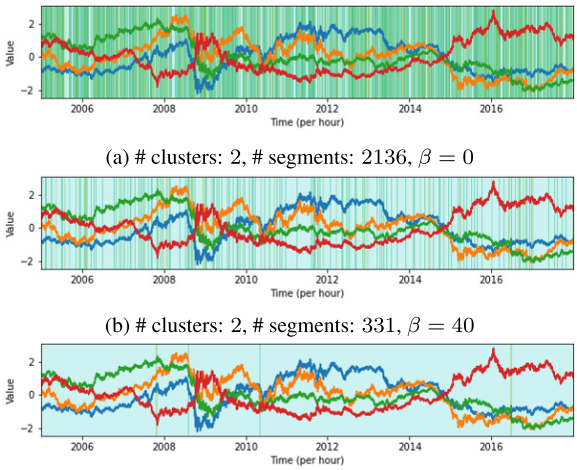
符号化コスト関数を制御するハイパーパラメータとして，α，λ が存在する．α はモデルの複雑さを制御する．α を小さくすると，モデル表現コストの影響が小さくなるため，クラスタ間のモデルの差が小さいクラスタが発見されるようになる．λ はモデルのスパース性を制御する．λ を大きくすると，モデルがスパースになるためモデル表現コストの値が小さくなり，クラスタ間のモデルの差が小さいクラスタが発見されるようになる．クラスタ間のモデルの差が異なる多種の実データに対応するには，これらのハイパーパラメータの設定は不可欠である．本章では，分割結果を一意に決定するにあたり，大きなクラスタが形成されているか，クラスタ数は解釈できる数に収まっているか，という 2 点を重要視した．分割結果が直感にそぐわない場合は適宜ハイパーパラメータを調整した．そして，分割結果からデータの背景（社会情勢，地図情報）を考慮することで，クラスタを解釈した．

6.2　金融データ

一般的に，株式，国債，外貨データは互いに相関している．時系列金融データを解析することで，経済ネットワークの関係性を推定することができる．投資においてネットワーク構造を知ることは，ポートフォリオ形成の際，高相関のプロダクトを避けるなどのリスク回避行動がとれるため重要である．2005 年から 2018 年の 1 時間ごとに取得された外貨交換レートデータ（AUD/USD，EUR/USD，GBP/USD，USD/CAD）を用い実験した[*4]．サンプル数は 82,882，次元数は 4 である．ネットワーク構造が 1 週間変化しないと仮定し，初期分割点を 1 週間（123 サンプル前後）に設定した．また，データを 1 週ごと正規化することで，ネットワーク構造の変化のみをとらえることに焦点を置いた．これにより，得られるクラスタの逆共分散行列の非対角成分の値は各変数の相関係数となる．

図 8 上部は経済データのクラスタリング結果を示す．NGL が 2007 年中期から 2009 年初期にかけて起こった世界金融危機付近で相関関係の変化をとらえていることが，黄色クラスタから灰色クラスタへの変化で分かる．図 8 下部は 2016 年 5 月 16 日から 2016 年 6 月 5 日におけるネットワーク構造の急激な変化を示している．赤字で示した行列の値から分かるようにイギリスに関する相関係数が大きく変化している．これは，2016 年 6 月 23 日に控えたイギリスの欧州連合離脱是非を問う国民投票による，国民の関心や不安が反映されたものと思われる．

[*4] https://github.com/FutureSharks/financial-data



(a) # clusters: 2, # segments: 2136, $\beta = 0$

(b) # clusters: 2, # segments: 331, $\beta = 40$

(c) # clusters: 2, # segments: 8, $\beta = 200$

図 9　外貨交換レートデータにおける TICC のクラスタ割当て
Fig. 9　Clustering result of TICC using currency datasets.

6.2.1　比較手法 TICC との差

TICC [8] はクラスタ数を事前に指定する必要があり，クラスタ数を 1 つずつ増加させた結果に BIC を用いることでクラスタ数の決定が可能である．**図 9** に同じ条件の経済データにおいて TICC を用い，割り当てられたクラスタを時間に対してプロットした実験結果を示す．図 9 の (a)，(b)，(c) は各 β において最良の BIC を示した結果である．TICC はハイパーパラメータを調整したにもかかわらず，大きなクラスタを発見できていないことが分かる．これは TICC がクラスタ間のモデルの差がある程度以上存在する場合のみクラスタを発見することができるからである．本実験のデータは 1 週間ごとに正規化されているため，クラスタ間のモデルの差が人工データと比較して小さい．TICC はクラスタ間のモデルの差を制御するハイパーパラメータを持たないため，クラスタ割当ての初期化に失敗し，大きなクラスタを発見できないと考えられる．一方で，提案手法はボトムアップ型のアルゴリズムにより，隣接データを同一のセグメントに組み込みやすくなり大きなセグメントを形成しやすい．また，符号化コストを制御することで，クラスタ間のモデルの差が小さい場合でも大きなクラスタを発見できる．これは α を調整した車両走行センサデータの例からも分かる．

6.3　車両走行センサデータ：高速道路

先に例示したように，車両走行センサは互いに相関しており，運転行動によりネットワーク構造は異なる．図 1 は

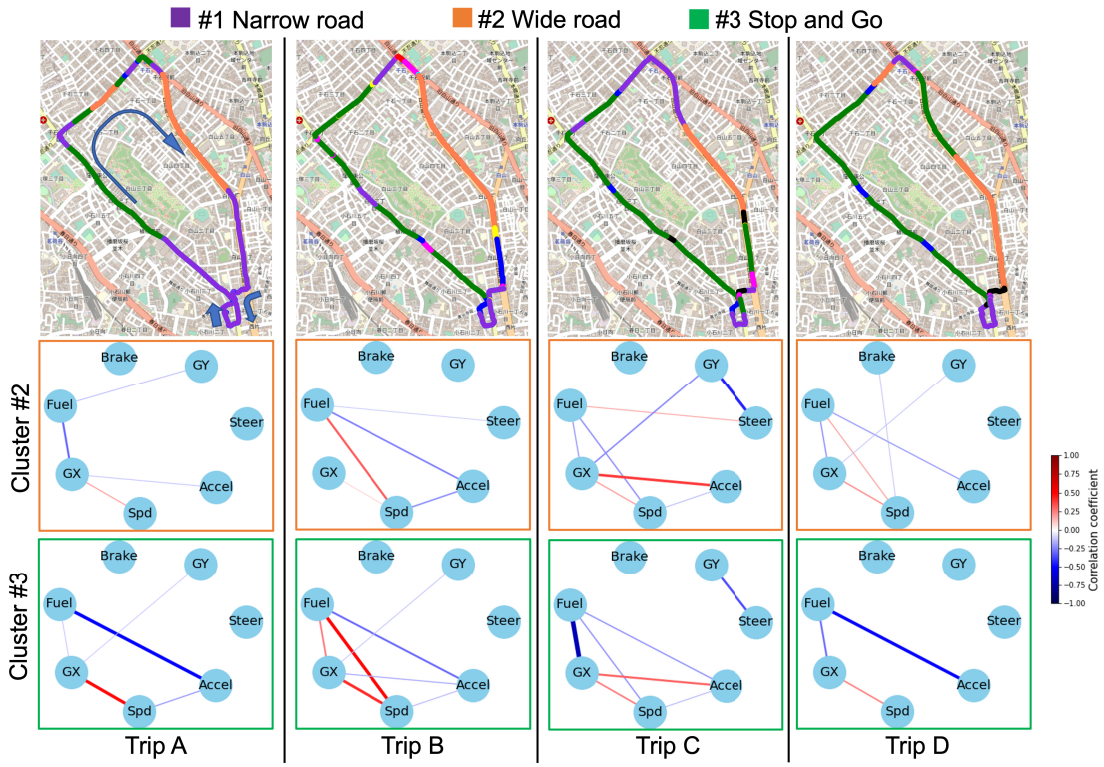


図 10　車両走行センサデータ：市街地における NGL の出力例
Fig. 10　Clustering result of NGL using city automobile datasets.

高速道路を約 3,500 m 走行した車両走行センサデータの解析結果である．複数のドライバーが同コースを計 40 周回し，7 つのセンサデータ（左右加速度，ハンドル角，アクセルペダルストローク，速度，前後加速度，燃費，ブレーキペダルストローク）が 10 m 間隔で取得された．サンプル数は 14,000，次元数は 7 である．ある地点を走行する際のネットワーク構造を発見することを目的とするため，初期分割点の幅を周回数と同じ 40 に設定し，初期セグメントごとに正規化した．

図 1 右部は NGL により発見された 4 つのクラスタのネットワーク構造を示している．すべてのクラスタに共通した特徴として，燃費とアクセルペダルストロークの負の相関，燃費と前後加速度の負の相関，アクセルペダルストロークと前後加速度の正の相関が見られた．そして，ネットワーク構造と地図を用いることですべてのクラスタを解釈することができた．クラスタ#1 は直線を安定して走行している区間であると解釈できる．クラスタ#2 にはクラスタ#1 のネットワークに加え左右加速度とハンドル角に負の相関が見られることから，左折していることが分かる．クラスタ#3 では唯一，ブレーキペダルストロークと前後加速度の相関が見られる．これはクラスタ#3 付近に合流や料金所があるために加減速を繰り返したためだと考えられる．クラスタ#4 は左右加速度とハンドル角に負の相関に加え，アクセルペダルストロークと速度に正の相関が見られることから加速しながら左折をしていることが分かる．実際に地図を確認することで合流などがなく，アクセルを踏めそうなカーブを左折していると確認できる．

6.4　車両走行センサデータ：市街地

図 10 は市街地を約 3,260 m 走行した車両走行センサデータの解析結果である．複数のドライバが同コースを 4 走行したデータで，7 つのセンサデータが 5 Hz で取得されたものをそれぞれ NGL で解析した．サンプル数はそれぞれ 5,464，5,991，6,410，5,800，次元数は 7 である．今回は，単一走行のネットワーク構造を発見することを目的とするため，初期分割点を 2 秒間隔に設定し，初期セグメントごとに正規化した．

市街地コースは主に道幅が狭く右左折が多い区間と，信号が多く混雑した区間と，道幅が広くスピードを出せる区間から構成される．図 10 左部の地図では色とクラスタ番号が対応しており，NGL により，すべての走行において類似した割当てがされたことが分かる．地図情報から判断すると，クラスタ#1 は道幅が狭い区間．クラスタ#2 は道幅が広い区間，そしてクラスタ#3 は混雑した区間に対応すると考えられる．また，クラスタ#2 と#3 のネットワーク構造をそれぞれ走行 A から D の間で比較することで走行の差やクラスタの特徴が分かる．クラスタ#2 では，道幅が広い区間では走行ごとに異なったネットワーク構造を持つことが分かる．これは広い道路では信号や前方車の有無などの複雑な外的要因，ドライバの技量により，走行

ごとに異なる操作をしていることを示している．クラスタ#3 ではすべてのネットワーク構造に前後加速度と速度の正の相関と，燃費とアクセルペダルストロークの負の相関が共通して見られることが分かる．加減速を繰り返す区間では，この 2 つの相関が強く出ることが分かる．

7. むすび

本論文では，ネットワーク構造に基づいたパターンを検出する手法として，NGL を提案した．NGL は与えられた多次元時系列データに関する事前知識を必要とせずネットワーク構造に基づいた解釈性の高いクラスタを発見できる．また，NGL はデータの最適な分割点とクラスタ数を自動的に発見することができる．人工データを用いた実験により，NGL は最新の既存手法と比べてより高い精度を持つことを示した．さらに，さまざまな種類の実データを用いた実験では，NGL が解釈性の高いネットワーク構造を持つクラスタを発見することを示した．

謝辞　本研究の一部は JSPS 科研費，JP20H00585，JP21H03446，JP22K17896，国立研究開発法人情報通信研究機構委託研究 03501，総務省 SCOPE JP192107004，JSTAIP 加速課題 JPMJCR21U4，ERCA 環境研究総合推進費 JPMEERF20201R02 の助成を受けたものです．

参考文献

[1] Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K. and Itakura, F.: Driver modeling based on driving behavior and its evaluation in driver identification, *Proc. IEEE*, Vol.95, No.2, pp.427–437 (2007).

[2] Hanneke, S., Fu, W. and Xing, E.P.: Discrete temporal models of social networks, *Electronic Journal of Statistics*, Vol.4, pp.585–605 (online), DOI: 10.1214/09-EJS548 (2010).

[3] Chiang, T.C., Jeon, B.N. and Li, H.: Dynamic correlation analysis of financial contagion: Evidence from Asian markets, *Journal of International Money and Finance*, Vol.26, No.7, pp.1206–1228 (online), DOI: 10.1016/j.jimonfin.2007.06.005 (2007).

[4] Friedman, J., Hastie, T. and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, Vol.9, No.3, pp.432–441 (2008).

[5] Li, L., Prakash, B.A. and Faloutsos, C.: Parsimonious linear fingerprinting for time series, Technical Report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science (2010).

[6] Kawabata, K., Matsubara, Y. and Sakurai, Y.: Automatic Sequential Pattern Mining in Data Streams, *CIKM*, pp.1733–1742 (2019).

[7] Honda, T., Matsubara, Y., Neyama, R., Abe, M. and Sakurai, Y.: Multi-aspect Mining of Complex Sensor Sequences, *ICDM*, pp.299–308 (2019).

[8] Hallac, D., Vare, S., Boyd, S.P. and Leskovec, J.: Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data, *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.215–223, ACM (online), DOI: 10.1145/3097983.3098060 (2017).

[9] Tozzo, V., Ciech, F., Garbarino, D. and Verri, A.: Statistical Models Coupling Allows for Complex Local Multivariate Time Series Analysis, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event*, Zhu, F., Ooi, B.C. and Miao, C. (Eds.), pp.1593–1603, ACM (online), DOI: 10.1145/3447548.3467362 (2021).

[10] Kumar, S., Zhang, X. and Leskovec, J.: Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks, *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2019).

[11] Wen, Q., Gao, J., Song, X., Sun, L., Xu, H. and Zhu, S.: RobustSTL: A robust seasonal-trend decomposition algorithm for long time series, *Proc. AAAI Conference on Artificial Intelligence*, Vol.33, pp.5409–5416 (2019).

[12] Hallac, D., Bhooshan, S., Chen, M.H., Abida, K., Sosic, R. and Leskovec, J.: Drive2Vec: Multiscale State-Space Embedding of Vehicular Sensor Data, *21st International Conference on Intelligent Transportation Systems, ITSC 2018*, Zhang, W., Bayen, A.M., Medina, J.J.S. and Barth, M.J. (Eds.), pp.3233–3238, IEEE (online), DOI: 10.1109/ITSC.2018.8569550 (2018).

[13] Berndt, D.J. and Clifford, J.: Using Dynamic Time Warping to Find Patterns in Time Series, *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop, Technical Report WS-94-03*, Fayyad, U.M. and Uthurusamy, R. (Eds.), pp.359–370, AAAI Press (1994).

[14] Li, L., McCann, J., Pollard, N.S. and Faloutsos, C.: DynaMMo: Mining and summarization of coevolving sequences with missing values, *KDD*, pp.507–516 (2009).

[15] Wang, P., Wang, H. and Wang, W.: Finding semantics in time series, *SIGMOD*, pp.385–396 (2011).

[16] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: AutoPlait: Automatic Mining of Co-evolving Time Sequences, *SIGMOD* (2014).

[17] Mohan, K., Chung, M., Han, S., Witten, D., Lee, S.-I. and Fazel, M.: Structured learning of Gaussian graphical models, *Advances in Neural Information Processing Systems*, Vol.25 (2012).

[18] Tomasi, F., Tozzo, V., Verri, A. and Salzo, S.: Forward-Backward Splitting for Time-Varying Graphical Models, *Proc. Ninth International Conference on Probabilistic Graphical Models*, Kratochvíl, V. and Studený, M. (Eds.), Vol.72, pp.475–486, PMLR (online), available from ⟨https://proceedings.mlr.press/v72/tomasi18a.html⟩ (2018).

[19] Hallac, D., Park, Y., Boyd, S.P. and Leskovec, J.: Network Inference via the Time-Varying Graphical Lasso, *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.205–213, ACM (online), DOI: 10.1145/3097983.3098037 (2017).

[20] Harutyunyan, H., Moyer, D., Khachatrian, H., Steeg, G.V. and Galstyan, A.: Efficient Covariance Estimation from Temporal Data, arXiv preprint arXiv:1905.13276 (2019).

[21] Steeg, G.V., Harutyunyan, H., Moyer, D. and Galstyan, A.: *Fast Structure Learning with Modular Regularization*, Curran Associates Inc. (2019).

[22] Tomasi, F., Tozzo, V., Salzo, S. and Verri, A.: Latent Variable Time-varying Network Inference, *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, Guo, Y. and

Farooq, F. (Eds.), pp.2338–2346, ACM (online), DOI: 10.1145/3219819.3220121 (2018).
[23] Chandrasekaran, V., Parrilo, P.A. and Willsky, A.S.: Latent variable graphical model selection via convex optimization, *2010 48th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*), pp.1610–1613 (online), DOI: 10.1109/ALLERTON.2010.5707106 (2010).
[24] Keogh, E.J., Chu, S., Hart, D.M. and Pazzani, M.J.: An Online Algorithm for Segmenting Time Series, *Proc. 2001 IEEE International Conference on Data Mining*, Cercone, N., Lin, T.Y. and Wu, X. (Eds.), pp.289–296, IEEE Computer Society (online), DOI: 10.1109/ICDM.2001.989531 (2001).
[25] Böhm, C., Faloutsos, C., Pan, J.-Y. and Plant, C.: Ric: Parameter-free noise-robust clustering, *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), Vol.1, No.3, pp.10–es (2007).
[26] Mohan, K., London, P., Fazel, M., Witten, D. and Lee, S.-I.: Node-Based Learning of Multiple Gaussian Graphical Models, *J. Mach. Learn. Res.*, Vol.15, No.1, pp.445–488 (2014).

小幡 紘平

2020 年名古屋大学農学部応用生命化学科卒業．2021 年大阪大学大学院修士課程進学．時系列データマイニングの研究に従事．

松原 靖子　（正会員）

2007 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学院博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．2012 年 NTT コミュニケーション科学基礎研究所 RA．2013 年日本学術振興会特別研究員（PD）．2014 年熊本大学大学院自然科学研究科助教．この間，カーネギーメロン大学客員研究員．2016 年国立研究開発法人科学技術振興機構さきがけ研究者．2019 年大阪大学産業科学研究所准教授．2016 年度日本データベース学会上林奨励賞，情報処理学会山下記念研究賞．2018 年度 IPSJ/ACM Award for Early Career Contributions to Global Research，2020 年度情報処理学会マイクロソフト情報学研究賞，電気通信普及財団第 36 回テレコムシステム技術賞，令和 4 年度科学技術分野の文部科学大臣表彰若手科学者賞等受賞．2018～2019 年度日本データベース学会理事．大規模時系列データマイニングに関する研究に従事．ACM，電子情報通信学会，日本データベース学会各会員．

川畑 光希　（正会員）

2016 年熊本大学工学部情報電気電子工学科卒業．2018 年同大学院博士前期課程修了．2021 年大阪大学大学院情報科学研究科情報システム工学専攻博士後期課程修了．博士（情報科学）．2019 年大阪大学情報科学研究科日本学術振興会特別研究員（DC2）．2021 年大阪大学産業科学研究所助教．DEIM Forum 2016 最優秀論文賞，WebDB Forum 2018 最優秀論文賞，学生奨励賞，企業賞，2019 年度コンピュータサイエンス領域奨励賞（データベースシステム），等受賞．データマイニング，データストリーム処理の研究に従事．日本データベース学会会員．

中村 航大　（学生会員）

2020 年熊本大学工学部情報電気電子工学科卒業．2022 年大阪大学大学院情報科学研究科情報システム工学専攻博士前期課程修了．現在，同大学院博士後期課程に在籍．2022 年度大阪大学情報科学研究科賞，DEIM2022 最優秀インタラクティブ賞，2022 年度情報処理学会山下記念研究賞，等受賞．データマイニング，大規模複合時系列データ処理，データストリーム処理の研究に従事．日本データベース学会学生会員．

櫻井 保志　（正会員）

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013～2019 年熊本大学大学院自然科学研究科教授．2019 年大阪大学産業科学研究所産業科学 AI センターセンター長・教授．2022 年日本学術振興会学術システム研究センター主任研究員．本会平成 18 年度長尾真記念特別賞，平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010）等受賞．データマイニング，データストリーム処理，センサーデータ処理，Web 情報解析技術の研究に従事．ACM，IEEE，電子情報通信学会，日本データベース学会各会員．

（担当編集委員　山本 修平）