卒業研究報告書

題目

ガウス過程に基づく
逐次抽出型 c-回帰モデルの検討

指導教員

濵砂 幸裕　講師

報告者

18–1–037–0205

武川 海斗

近畿大学理工学部情報学科

令和 4 年 1 月 24 日提出

概要

c-回帰モデル (c-Regression Models; CRM) とは，回帰と分類を同時に行うクラスタリング手法である．CRM には様々な応用手法があり，その 1 つとして，ガウス過程に基づく c-回帰モデル (Gaussian Process c-Regression Model; GPCRM) が提案されている．GPCRM は，ガウス過程回帰を CRM に応用することで，期待値と分散を考慮した非線形な回帰モデルに基づく c-回帰手法である．しかし，GPCRM には，クラスタ数を事前に指定する必要がある．そのため，得られるクラスタ構造はクラスタ数に依存するという問題がある．

本研究では，クラスタ数の自動推定が可能な表現力の高いクラスタリング手法として，ガウス過程に基づく逐次抽出型 c-回帰モデル (Gaussian Process Sequential c-Regresson Model; GPSCRM) を提案する．逐次抽出型は，クラスタ数を事前決定せず，逐次的にクラスタを抽出するクラスタリング手法である．GPCRM に逐次抽出型の考え方を応用することで，クラスタ数の自動推定が可能な回帰モデルを構築することが可能と考えられる．さらに，提案手法の性能を評価するために，既存の逐次抽出 c-回帰モデル (Sequential c-Regression Model; SCRM) との比較実験を行った．

数値実験より，提案手法による分割結果は，不均衡なデータに関しての分割が有効に働くことを確認した．しかし，非線形なクラスタ構造に対して，有効な回帰線が得られないことが確認された．その原因として，提案手法に用いる正則化パラメータ λ とカーネルパラメータ α の適切な値を求めることが困難であることが考えられる．そのため，これらのパラメータを自動推定する手法の研究が必要であることが示唆された．

目次

図目次

1 序論

センシング技術やストレージ技術の進歩，インターネットやビデオ監視などのアプリケーションの飛躍的な拡大により，膨大かつ高次元なデータが大量に生み出されている [1]．これらのデータの多くは構造化されていないため，分析が困難になっている．そのため，データを自動的に理解し，処理し，分類するための手法が注目されている．その中でも，膨大なデータから有益なデータを効率的に取り出す技術が研究されるようになっており．これらの手法はデータマイニングと呼ばれる [1][2]．データマイニングの中でも，データを自動で分類する方法として，クラスタリングがある．クラスタリングとは教師なし学習の 1 つであり，個体間に定義された類似度，もしくは非類似度に基づいて，クラスタと呼ばれるグループに分類する手法である．

本研究では，教師あり学習である回帰分析と，教師なし学習であるクラスタリングについて取り扱う．クラスタリングに回帰分析の要素を加えた手法として，c-回帰モデル (c-Regression Model; CRM) がある．CRM は，回帰モデルをクラスタごとに当てはめることで，データ全体の構造を表現することができるモデルである [3][4]．CRM は，カーネルを用いた手法やロバストな手法，ガウス過程に基づく手法など，様々な発展手法が提案されている [5][6][7]．特に，ガウス過程に基づく c-回帰モデル (Gaussian Process c-Regression Model; GPCRM) は，既存の CRM では表現できない非線形な回帰モデルを持つ [5]．

ここで，ガウス過程とは，関数の確率分布を推定する手法であり，表現力の高さから様々な研究や手法の提案が行われている [8][9]．GPCRM では，回帰係数が確率分布で表現される．その利点として，分散を用いた回帰線の信頼度の可視化が挙げられる．しかし GPCRM は，クラスタ数を事前に指定する必要があり，得られるクラスタ構造はクラスタ数に依存するという問題点がある．

クラスタ数を事前に指定する必要がない手法として，逐次抽出型クラスタリングがある．逐次抽出型とは，クラスタ数を指定せずに，逐次的にクラスタを抽出するクラスタリング手法である [10]．逐次抽出型を導入したモデルとして，逐次抽出型 c-means，逐次抽出型 c-回帰モデル，可能性クラスタリングなどが提案されている [6][11][12][13]．しかし，逐次抽出型の手法に関して，非線形なデータに関する検証は少ない．特に，逐次抽出型 c-means に関するカーネル化の研究は，良好な結果が得られていない [13]．他にも，別のアプローチの逐次抽出型のクラスタリング手法として，マウンテンクラスタリングが存在する [14]．マウンテンクラスタリングでは，次元が増えると計算量が増えるという問題点がある．

そこで，本研究では，表現力の高いガウス過程に基づく c-回帰モデルに対して，クラスタ数の自動推定を行う逐次抽出型を導入することで，新たなクラスタリング手法を提案する．さらに，提案手法の有効性を検証するために，人工データを用いて既存の逐次抽出 c-回帰モデルとの比較実験を行う．分類結果の比較として，Adjusted Rand Index(ARI) を用いる [15]．ARI とは，2 つの異なる有限集合の一致度を測る評価指標である．ARI は 0 以上 1 以下の値をとり，値が 1 に近づくほど正解ラベルとの一致度が高い分割であることを表し，良いクラスタリングが行えていると評価する．

本論文の構成は，次の通りである．第 2 節では，提案手法に用いる逐次抽出型 c-回帰モデル，ガウス過程 c-回帰モデルおよびガウス過程とその周辺情報に関しての説明を行う．第 3 節では，提案手法であるガウス過程に基づく逐次抽出型 c-回帰モデルの説明を行う．第 4 節では，提案手法の有効性を検証するために行った数値実験の内容と結果について述べる．さらに，その実験から得られた結果に関して考察を行う．第 5 節では，数値実験を通じて得られた結果をまとめ，今後の課題について説明する．

2　準備

2.1　カーネル法

カーネル法とは，非線形な構造を持つデータを扱う場合に用いられる手法である．しかし，非線形な構造を持つデータに対する最適化問題は，次元数が増えた場合や大規模データの場合に，解くことが困難となる．そこでカーネル法では，線形のモデルで非線形問題を解くという方法を用いる．そのために，入力データを高次元空間に写像し，高次元空間上で線形分類を行う．しかし，高次元空間には次元の呪いという問題があり，データ数が増えると容易に分類を行うことはできない．そこで，カーネル法では高次元空間における内積を入力データから直接計算する方法を考える．この内積計算に使用される関数がカーネル関数である．そのため，具体的な写像の内容が分からないまま内積を計算することができ，計算量も大幅に削減することができる．これはカーネルトリックと呼ばれる．カーネル法はさまざまな手法に対して適応可能であり，k-means，階層的クラスタリングなどのクラスタリング手法に対して用いられている [1][16][17].

よく用いられるカーネル関数には，ガウスカーネルがある．ここで，$\boldsymbol{x}$ と $\boldsymbol{x}'$ は入力データであり，$\boldsymbol{x}$ と $\boldsymbol{x}'$ によって値が決まる．ガウスカーネルは以下の式 (1) で表される．

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \alpha \exp\left(-\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2\right) \tag{1}$$

$\alpha \in \Re^+$ は，ガウスカーネルの性質を決めるためのパラメータである．

2.2　ガウス過程

2.2.1　ガウス分布

ガウス分布とは，正規分布とも呼ばれる連続型の確率変数である．つりがね状の形をした確率分布であり，パラメータに平均 μ と分散 σ^2 を持つ．1 次元のガウス分布の確率密度関数は以下の式 (2) で表される．

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2}$$

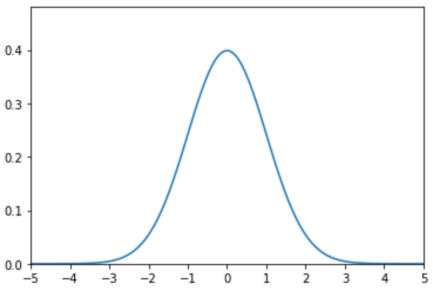

図 1: 1 次元ガウス分布

2.2.2　多変量ガウス分布

ガウス過程の導出のために，多変量のガウス分布について説明する．s 次元のベクトル $\boldsymbol{x} = (x_1, x_2, ...x_s)$ が平均 $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ のガウス分布に従っているとき，確率密度関数は以下の式 (3) で表される．

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^s \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3}$$

式 (3) で用いられる共分散行列の逆行列 $\boldsymbol{\Sigma}^{-1}$ は，精度行列と呼ばれ，$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ とする．$\boldsymbol{\Sigma}$ は $s \times s$ の共分散行列で，その $(i.j)$ 要素が x_i と x_j の共分散を表している．また，期待値のベクトル $\boldsymbol{\mu}$ および，共分散行列 $\boldsymbol{\Sigma}$ は以下の式 (4)，(5) で表される．

$$\boldsymbol{\mu} = E[\boldsymbol{x}] \tag{4}$$

$$\begin{aligned} \Sigma_{ij} &= E[(x_i - E[x_i])(x_j - E[x_j])] \\ &= E[x_i x_j] - E[x_i]E[x_j] \end{aligned} \tag{5}$$

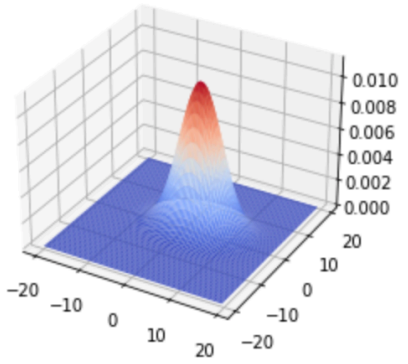

図 2: 2 次元ガウス分布

2.2.3　多変量ガウス分布の周辺化

周辺化とは，同時確率に対して一部の確率変数を積分することで周辺分布を求めることである．多変量ガウス分布には，周辺化した後の分布も多変量ガウス分布に従うという特徴がある．

ここで，ある確率変数 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ が与えられたする．このときの同時確率を $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ と表す．以下は $\boldsymbol{x}_1, \boldsymbol{x}_2$ の同時分布から得られる期待値と分散の式 (6) である．

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right) \tag{6}$$

式 (6) を $\boldsymbol{x_2}$ に関して周辺化した時の $\boldsymbol{x_1}$ は，以下の式 (7) で表される．

$$p(\boldsymbol{x}_1) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2)\, d\boldsymbol{x}_2 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \tag{7}$$

2.2.4 多変量ガウス分布の条件付き分布

多変量ガウス分布の条件付き分布とは、変数の一部を固定し、残りの変数について表した分布である．この時の条件付き確率は以下の式 (8) で表される．

$$p\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right)=N\left(\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right), \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}\right) \tag{8}$$

2.2.5 ガウス過程

ガウス過程 (Gaussian Process) は，入力空間 $\boldsymbol{\chi}$ 上のランダムな関数を求める確率過程である．確率過程とは，入力の集合 $\boldsymbol{X}=\{x_1, x_2, \cdots x_n\}$ に対応する確率変数の集合 $\boldsymbol{Y}=(y_1, y_2, ...y_n)$ に同時分布 $p(y_1, y_2, \cdots, y_n)$ を与える確率分布を指す．以下にガウス過程の定義を述べる [19].

どんな自然数 N についても，入力 $\boldsymbol{X}=\{x_1, x_2, \cdots x_n\} \in \chi$ に対応するベクトル

$$\boldsymbol{f}=\{f(x_1), f(x_2), \cdots f(x_N)\} \tag{9}$$

が平均 $\boldsymbol{\mu}=(\mu(x_1), \mu(x_2), \cdots, f(x_N))$，$K_{nn'}=k(x_n, x_{n'})$ を要素とする行列 $\boldsymbol{K}$ を共分散行列とするガウス分布 $N(\boldsymbol{\mu}, \boldsymbol{K})$ に従うとき，f はガウス過程に従うといい，これを

$$\boldsymbol{f} \sim GP\left(\boldsymbol{\mu}, \boldsymbol{K}\right) \tag{10}$$

と表す.

ここで，$\boldsymbol{f}$ がガウス過程である場合，平均関数 $\boldsymbol{\mu}(\boldsymbol{x})$ と正定値カーネル $\boldsymbol{K}$ が存在することを意味する．逆に，平均関数 $\boldsymbol{\mu}(\boldsymbol{x})$, 正定値カーネル $\boldsymbol{K}$ が存在する場合，対応するガウス過程 $\boldsymbol{f} \sim \mathrm{GP}(\boldsymbol{\mu}, \boldsymbol{K})$ が存在する [19]．このように，ガウス過程 $\boldsymbol{f} \sim \mathrm{GP}(\boldsymbol{\mu}, \boldsymbol{K})$ と、平均関数 $\boldsymbol{\mu}$ と正定値カーネル $\boldsymbol{K}$ のペアには、一対一の対応関係がある．そのため，入力データが観測されれば，$\boldsymbol{f}$ の事後分布を求めることができ，ガウス過程はベイズ統計の立場から見たカーネル法であると言える [19].

2.3 回帰モデル

2.3.1 カーネル回帰モデル

カーネル回帰モデルとは，線形回帰モデルの入力 $\boldsymbol{X}$ をカーネル関数に置き換えたもので，非線形なモデルを構築することができる [17]．ここで入力データを $\boldsymbol{X}=\{\boldsymbol{x}_k \mid x_k \in \Re^p, k=1 \sim n\}$，$\boldsymbol{Y}=\{y_k \mid x_k \in \Re, k=1 \sim n\}$ とし，$\boldsymbol{\beta}=\{\beta_1, \beta_2, \cdots, \beta_n\}$ とする．すると，カーネル回帰モデルでは，$f(\boldsymbol{x})$ について，式 (11) とする．

$$f(\boldsymbol{x})=\sum_{i=1}^{n} \beta_i k\left(\boldsymbol{x}_i, \boldsymbol{x}\right) \tag{11}$$

ここで，グラム行列，回帰パラメータを以下の式 (12),(13) とする．

$$\boldsymbol{K} = \begin{pmatrix} k(x_1, x_1) & k(x_2, x_1) & \cdots & k(x_n, x_1) \\ k(x_1, x_2) & k(x_2, x_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_1, x_n) & \cdots & \cdots & k(x_n, x_n) \end{pmatrix} \tag{12}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \tag{13}$$

カーネル回帰モデルによる二乗誤差 $R(\boldsymbol{\beta})$ は式 (14) のようになる．

$$\begin{aligned} R(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i - \sum_{i=1}^{n} \beta_i k(x_i, x) \right\}^2 \\ &= (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta}) \end{aligned} \tag{14}$$

従って，$R(\boldsymbol{\alpha})$ を最小化するような $\boldsymbol{\alpha}$ は以下の式 (15) になる．

$$\boldsymbol{\beta} = \left(\boldsymbol{K}^T \boldsymbol{K}\right)^{-1} \boldsymbol{K}^T \boldsymbol{y} = \boldsymbol{K}^{-1} \boldsymbol{y} \tag{15}$$

2.3.2 ガウス過程回帰モデル

ガウス過程回帰とは，回帰のためのベイズに基づいたノンパラメトリックな手法である．ガウス過程回帰により推定される関数は，確率分布で表される．そのため，出力結果として標準偏差を使ったデータの散らばりを表現することができる．

ここで，$\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$ が与えられるとする．$\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon$ の関係が成り立つと仮定し，誤差 ϵ はガウス分布 $N(0, \sigma^2)$ に従うとする．入力 $\boldsymbol{x}$ が与えられた時の $\boldsymbol{y}$ の分布は，カーネル関数を通じて，ガウス過程として表現できる．そのため，$\boldsymbol{f}$ は以下の式 (16) で表される．

$$\boldsymbol{f} \sim GP\left(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') + \sigma^2 \boldsymbol{I}\right) \tag{16}$$

次に，新たな入力 $\boldsymbol{x}$ での出力 $\boldsymbol{y}^*$ を予測するため，出力 $\boldsymbol{Y}$ と $\boldsymbol{y}^*$ の同時分布は以下の式 (17) で表される．

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{y}^* \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k}_{**} \end{pmatrix}\right) \tag{17}$$

この時，$\boldsymbol{k}_* = \left(k(\boldsymbol{x}^*, \boldsymbol{x}_1), \cdots, k(\boldsymbol{x}^*, \boldsymbol{x}_n)\right)^T$ はカーネル関数の値を並べたベクトルであり，観測データと予測点との内積を表す．また，$\boldsymbol{k}_{**} = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$ は予測したい点の共分散を表す．多変量ガウス分布の条件付き確率の式 (8) に基づき，以下の式 (18) で表す．

$$(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{D}) = N\left(\boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{y}, \boldsymbol{k}_{**} - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{k}_*\right) \tag{18}$$

この時，$\boldsymbol{y^*}$ の期待値は $\boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{y}$，分散は $\boldsymbol{k}_{**} - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{k}_*$ で与えられる．

2.4 c-回帰モデル

c-回帰モデル (c-Regression Model; CRM) は，回帰とクラスタリングを同時に行う手法である [3]．クラスタごとに線形回帰式を準備し，二乗誤差が小さい方にクラスタが属するように交互最適化を行うことでクラスタリングを行う．クラスタリング手法としてよく用いられる K-means 法は，球状のクラスタを抽出できるのに対して，CRM 法は超平面状のクラスタを抽出することができる [4]．そのため，株価のリターンの予測や，各国の GDP データに対するクラスタリングなどの超平面状のデータが予測されるデータに対して用いられることが多い [12][20]．

本節では，CRM を拡張した手法として，逐次抽出型 c-回帰モデル，ガウス過程に基づく c-回帰モデルについて説明を行う．

2.4.1 逐次抽出型 c-回帰モデル

逐次抽出型 c-回帰モデル (Sequential c-Regression Model; SCRM) とは，CRM 法をノイズクラスタリング [10] に拡張した回帰モデルである [12]．

個体集合を $X = \{\boldsymbol{x}_k \mid \boldsymbol{x}_k \in \Re^s, k = 1 \sim n\}$，$Y = \{y_k \in \Re, k = 1 \sim n\}$，クラスタ集合を $\boldsymbol{C} = \{C_i \mid i = 1 \sim c\}$ とする．ここで，i 番目のクラスタにおける k 番目のデータの残差 d_{ki} は以下の式 (19) で表される．

$$d_{ki} = \left(y_k - \sum_{j=1}^{p} \beta_i^j x_k^j + \beta_i^{p+1} \right)^2 \tag{19}$$

ここで，β は回帰パラメータである．また，SCRM の目的関数と制約条件は以下の式 (20) になる．

$$J_{\mathrm{SCRM}}(\boldsymbol{U}, \boldsymbol{B}) = \sum_{k=1}^{n} u_{ki} d_{ki} + \sum_{k=1}^{n} u_{k0} D \tag{20}$$

$$\mathcal{U}_{\mathrm{SCRM}} = \left\{ (u_{ki}) : u_{ki} \in \{0, 1\}, \sum_{i=1}^{c} u_{ki} = 1, {}^{\forall}k \right\} \tag{21}$$

ただし，$D > 0$ をノイズパラメータ，抽出クラスタを u_{k1}，ノイズクラスタを u_{k0} とする．逐次抽出法では，残差 d_{ki} とノイズパラメータ D を比較し，個体が抽出クラスタ，ノイズクラスタのどちらかに分類することで，データが密な領域を抽出することができる手法である．そして，抽出クラスタに属する要素を $\boldsymbol{X}$ から抽出し，再び残った個体でクラスタリングを行う．

また，個体の抽出を行うたびにデータのサイズが変わるため，ノイズパラメータ D の更新も必要となる [18]．ここで，D の初期値として，$D = 0$ を与える．ノイズパラメータ D の更新式は以下の式 (22) で表される．ここで，θ はノイズパラメータ D の大きさを決定するパラメータである．抽出クラスタに属するデータを l 個とし，ノイズクラスタに属するデータを $n - l$ 個とする．また，D' は，更新前の D の値である．

$$D = \theta \left[\frac{d_{ki} + \sum_{k=1}^{n-l} D'}{n} \right] \tag{22}$$

Algorithm 1 に SCRM のアルゴリズムを示す．

Algorithm 1 SCRM

SCRM1 θ の初期値を設定する.

SCRM2 ランダムに帰属度 u_{ki} を設定する.

SCRM3 収束するまで **SCRM3** を繰り返す.

SCRM3.1 β_i を計算する.

$$\boldsymbol{\beta}_i = \left(\sum_{k=1}^{n} u_{ki}\boldsymbol{z}_k\boldsymbol{z}_k^T\right)^{-1}\left(\sum_{k=1}^{n} u_{ki}y_k\boldsymbol{z}_k\right)$$

ただし，$\boldsymbol{z_k} = \left(x_k^1, \ldots, x_k^p, 1\right)$ である.

SCRM3.2 u_{ki} の帰属度を更新する.

$$u_{ki} = \begin{cases} i & (d_{ki} \leq D) \\ 1-i & (\text{ otherwise }) \end{cases} \quad (i = 0, 1)$$

SCRM4 ノイズパラメータ D を更新する.

SCRM5 $\{x_k \mid u_{ki} = 1\}$ の要素を $\boldsymbol{X}$ から抽出する.

SCRM6 $\boldsymbol{X} = \emptyset$ の場合アルゴリズムを終了する.

$\boldsymbol{X} \neq \emptyset$ の場合 **SCRM2** に戻る.

END SCRM.

2.4.2 ガウス過程に基づく c-回帰モデル

ガウス過程回帰モデル (Gaussian Process c-Regression Model; GPCRM) では，観測値 $\boldsymbol{Y}$ と入力データ $\boldsymbol{X} = \{x_k \mid x_k \in \Re^p, k = 1 \sim n\}$ が与えられた時のガウス過程回帰の予測分布との残差を考える [5].

ここで，i 番目のクラスタにおけるガウス過程回帰の予測分布は以下の式 (23) で表される.

$$P\left(y_k^{(i)*} \mid \boldsymbol{x}_k^{(i)*}, \boldsymbol{D}_i\right) = N(\boldsymbol{k}_*^{(i)T}\boldsymbol{K}_i^{-1}\boldsymbol{y}^{(i)}, k_{**} - \boldsymbol{k}_*^{(i)T}\boldsymbol{K}_i^{-1}\boldsymbol{k}_*^{(i)}) \tag{23}$$

k 番目の個体である x_k と i 番目のクラスタとの残差を考えると，GPCRM の目的関数と制約条件は以下の式 (24),(25) になる.

$$J_{\text{gpcrm}}(\boldsymbol{U}, \boldsymbol{K}) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}(y_k^{(i)} - \boldsymbol{k}_*^{(i)T}\boldsymbol{K}_i^{-1}\boldsymbol{Y_i})^2 \tag{24}$$

$$\text{s.t. } \sum_{i=1}^{c} u_{ki} = 1 (k = 1, \ldots, n) \tag{25}$$

ここで，ガウス過程はモデルの表現力が高く，カーネル関数のパラメータによっては過学習を起こすため，$\boldsymbol{K_i^{-1}} = (\boldsymbol{K_i} + \lambda\boldsymbol{I_{ci}})^{-1}$ として，正則化項を追加する.

GPCRM では，帰属度 $\boldsymbol{U}$ とカーネル行列 $\boldsymbol{K}$ を交互最適化をすることで最適化を行う．**Algorithm 2** に GPCRM のアルゴリズムを記す.

Algorithm 2 GPCRM

GPCRM1 クラスタ数 c，カーネルパラメータ α を設定
GPCRM2 データ $\boldsymbol{D}$ を c 個に分割
GPCRM3 収束するまで，**GPCRM3** を繰り 返す
　GPCRM3.1 $\boldsymbol{U}$ を固定，$\boldsymbol{K}$ を更新
　GPCRM3.2 $\boldsymbol{K}$ を固定，$\boldsymbol{U}$ を更新
GPCRM4 収束すれば終了，そうでなければ
　　GPCRM3 に戻る
END GPCRM.

3　ガウス過程に基づく逐次抽出型 c-回帰モデル

本節では，提案手法であるガウス過程に基づく逐次抽出型モデル (Gaussian Process Sequential c-Regression Model; GPSCRM) についての定式化について記す．

個体集合を $\boldsymbol{X} = \{\boldsymbol{x}_k \mid \boldsymbol{x}_k \in \Re^p, k = 1 \sim n\}$，$\boldsymbol{Y} = \{\boldsymbol{y}_k \mid y_k \in \Re, k = 1 \sim n\}$，クラスタ集合を $C = \{C_i \mid i = 1 \sim c\}$ とする．$D > 0$ をノイズパラメータ，抽出クラスタを u_{k1}，ノイズクラスタを u_{k0} とする．ここで，抽出クラスタに属するデータ点を $\boldsymbol{x}_k^{(1)}$，ノイズクラスタに属するデータ点を $\boldsymbol{x}_k^{(0)}$ と表記する．$\boldsymbol{x}_*$ は予測点であり，$\boldsymbol{k}_*^{(1)} = (k(\boldsymbol{x}_*^{(1)}, \boldsymbol{x}_1^{(1)}), k(\boldsymbol{x}_*^{(1)}, \boldsymbol{x}_2^{(1)}), \cdots, k(\boldsymbol{x}_*^{(1)}, \boldsymbol{x}_n^{(1)}))^T$ は抽出クラスタに属する予測点 $\boldsymbol{x}_*^{(1)}$ と抽出クラスタに属する入力点 $\boldsymbol{x}_k^{(1)}$ を入力ベクトルとするカーネル関数の値を並べたベクトルである．また，$\boldsymbol{K}^{(1)-1}$ は抽出クラスタに属する $\boldsymbol{x}_k^{(1)}$ のカーネル関数の値を並べた行列の逆行列である．ただし，$\boldsymbol{K}^{(1)-1}$ には正則化項を追加するため，$\boldsymbol{K}^{(1)-1} = \left(\boldsymbol{K}^{(1)} + \lambda \boldsymbol{I}_1\right)^{-1}$ より計算される．

従って，GPSCRM の目的関数は以下の式 (26) で表される．

$$J_{\mathrm{gpscrm}}(\boldsymbol{U}, \boldsymbol{K}) = \sum_{k=1}^{n} u_{k1} \left(y_k^{(1)} - \boldsymbol{k}_*^{(1)T} \boldsymbol{K}^{(1)-1} \boldsymbol{Y}^{(1)}\right)^2 + \sum_{k=1}^{n} u_{k_0} D \tag{26}$$

また，帰属度の制約条件は以下の式 (27) で表される．

$$\mathcal{U}_{\mathrm{gpscrm}} = \left\{ (u_{ki}) : u_{ki} \in \{0, 1\}, \sum_{i=0}^{1} u_{ki} = 1, {}^{\forall}k \right\} \tag{27}$$

GPSCRM は非類似度 $\left(y_k^{(1)} - \boldsymbol{k}_*^{(1)T} \boldsymbol{K}^{(1)-1} \boldsymbol{Y}^{(1)}\right)^2$ とノイズパラメータ D を比較する．入力データが抽出クラスタ，ノイズクラスタに属するか判定して分類することでクラスタリングを行う．

また，SCRM と同様に，ノイズパラメータの初期値として $D = 0$ として，抽出のたびに更新を行う．その際のハイパーパラメータを θ とする．

Algorithm3 に GPSCRM のアルゴリズムを記す．

Algorithm 3 GPSCRM

GPSCRM 1 ノイズパラメータを θ，カーネルパラメータ α，
正則化パラメータ λ を設定する．

GPSCRM 2 ランダムに帰属度 u_{ki} を設定する．

GPSCRM 3 収束するまで **GPSCRM2** を繰り返す

GPSCRM 3.1 カーネル行列 $\boldsymbol{K}^{(1)}$ を更新する．

GPSCRM 3.2 帰属度 u_{ki} を更新する．

$$u_{ki} = \begin{cases} i & (d_{k1} \leq D) \\ 1 - i & (\text{ otherwise }) \end{cases} \qquad (i = 0, 1)$$

GPSCRM 4 $\{x_k \mid u_{ki} = 1\}$ の要素を $\boldsymbol{X}$ から抽出する．

GPSCRM 5 $\boldsymbol{X} = \emptyset$ の場合アルゴリズムを終了する．
$\boldsymbol{X} \neq \emptyset$ の場合 **GPSCRM2** に戻る．

END GPSCRM.

4　実験

本実験は，提案手法であるガウス過程に基づく逐次抽出型 c-回帰の有用性，ハイパーパラメータを変化させたときのクラスタリング結果の検証を目的とする．以下に実験に用いるデータセット，実験条件および実験結果を示す．

4.1　データセット

今回使用したデータセットは 4 種類であり，各データセットの特徴は以下の通りである．

- データセット 1: データ数 $n = 150$，クラスタ数 $c = 2$ のデータである．ばらつきのある線形なクラスタ構造が 2 つ並んだデータである (図 3).
- データセット 2: データ数 $n = 120$，クラスタ数 $c = 2$ のデータである．クラスタ同士のデータ数に偏りがある．o のデータ数が 100，+ のデータ数が 20 である (図 4).
- データセット 3: データ数 $n = 300$，クラスタ数 $c = 3$ のデータである．クラスタ数を 3 つとし，ばらつきの小さい線形なクラスタ構造が 3 つ並んだデータである (図 5).
- データセット 4: データ数 $n = 100$，クラスタ数 $c = 2$ のデータである．クラスタ数が 2 つであり，1 つのクラスタ構造は線形構造，他方のクラスタ構造は非線形構造を持つデータである (図 6).

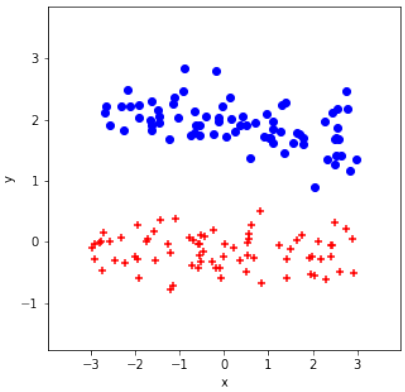

図 3: データセット 1

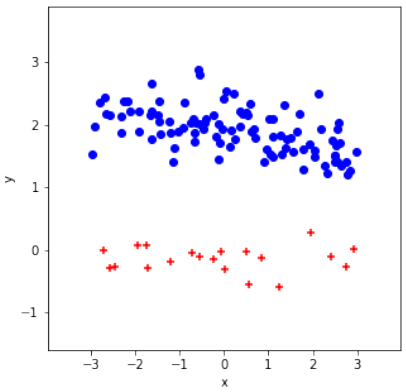

図 4: データセット 2

4.2　実験条件

本節では，SCRM，および提案手法である GPSCRM に関する実験条件について説明する．

4.2.1　逐次抽出型 c-回帰モデルの実験条件

SCRM によって得られる分割，回帰線は帰属度を 30 回ランダムに初期化し，目的関数が最小となる出力を実験結果として用いる．分割された結果に関しては，Adjusted Rand Index(ARI)[15] を用いる．ARI とは，

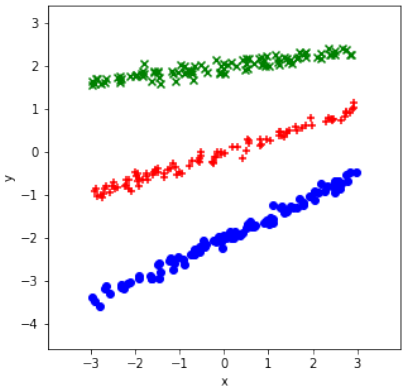

図 5: データセット 3

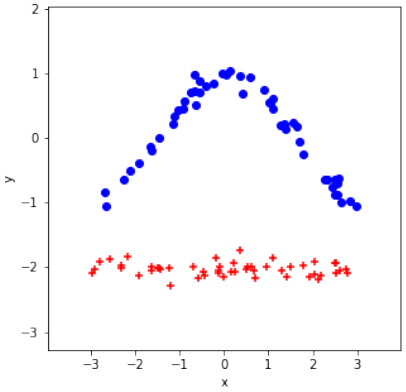

図 6: データセット 4

2 つの異なる有限集合の一致度を測る評価指標である．ARI は 0 以上 1 以下の値をとり，値が 1 に近づくほど一致度が高い分割であることを表し，良いクラスタリングが行えていると評価する．また，ノイズパラメータ θ は，$(1.4, 1.5, 2.3, 5)$ の中から選んで実験を行う．逐次抽出される順番に関しては，○，+，×，▽，*の順番でプロットする．

4.2.2　ガウス過程に基づく逐次抽出型 c-回帰モデルの実験条件

提案手法である GPSCRM によって得られる分割，回帰線についても，SCRM，GPCRM の場合と同様に 30 回ランダムに初期化し，目的関数が最小となる出力を実験結果として用いる．分割結果の評価には ARI を用いる．クラスタリング結果の実践部分は入力点 x_i に対する期待値を表し，半透明の帯のような箇所は，標準偏差を表す．ノイズパラメータ θ は $(1.5, 3)$ の中から選んで実験を行う．使用するカーネル関数はガウスカーネルとし，ガウスカーネルに用いるハイパーパラメータ α は，$(0.0001, 0.001, 0.04, 0.05, 0.09, 5)$ の中から用いる．ノイズパラメータは $\theta = 1.5$ とする．また，正則化パラメータ λ は，$(0.045, 0.01, 0.1, 0.5, 4)$ の中から用いる．逐次抽出される順番に関しては，○，+，×，▽，*の順番でプロットする．

4.3　実験結果

本節では，データ 1，2，3，4 に対する SCRM，GPSCRM を用いた実験結果を示す．

4.3.1　データ 1 に対する実験結果

データ 1 に対して，SCRM，GPSCRM を用いた結果をそれぞれ図 7，8 に示す．また，図 9，10，11，12 の実験結果は，GPSCRM に対して α，λ を変えた場合の出力結果である．本実験の目的は，線形なクラスタが並んだデータに対して，GPSCRM がクラスタリングできるのかを示すことである．図 7 より，SCRM では ARI が 0.65 と低く，良好な結果とは言えない．一方，GPSCRM では ARI=0.96 と良好な結果が得られている (図 8)．

図 8 の時のパラメータと比べて，α，λ の値を小さくした場合の出力結果が図 9，10 である．図 8 の場合よりも，標準偏差を表す帯についても，全体的に小さい値となっている．回帰線については大きい違いはないと言える．また，ARI=0.97 と良好な結果が得られた．

図 8 の時のパラメータと比べて，α の値を極端に大きくした場合の出力結果が図 11 である．標準偏差を表す帯が，大きい値を取るため，目視で確認することができない結果となった．回帰線についても，過学習が発生しており，良好な結果でないことが確認された． ARI=0.027 と良好でない結果が得られている．

図 8 の時のパラメータと比べて，α の値を極端に小さくした場合の出力結果が図 12 である．ほとんど水平な直線の回帰線が求まり，ARI=0.97 と良好な結果が得られた．

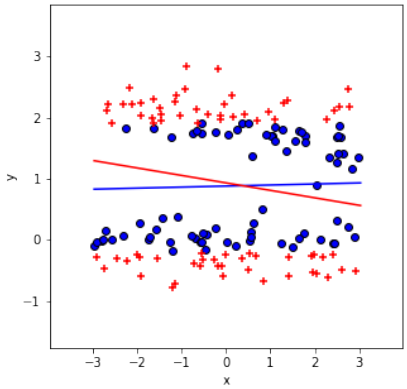

図 7: データ 1，SCRM による出力，$\theta = 2.3$，ARI=0.65

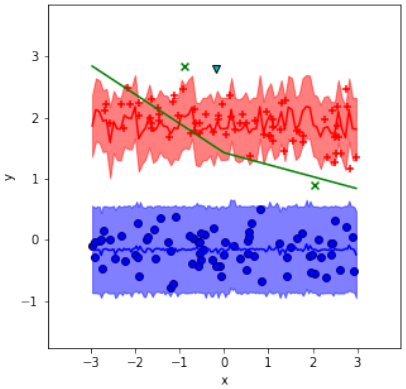

図 8: データ 1，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.05$，$\lambda = 0.5$，ARI=0.96

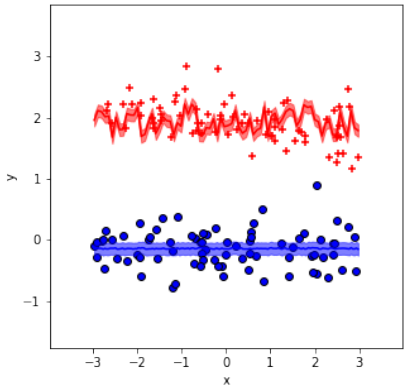

図 9: データ 1，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.001$，$\lambda = 0.01$，ARI=0.97

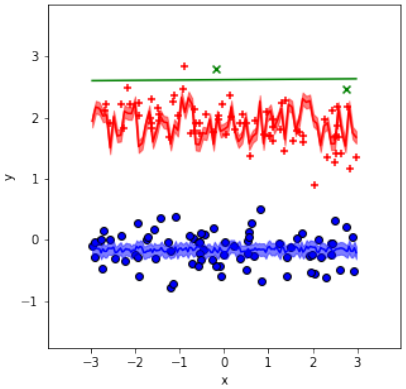

図 10: データ 1，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.0001$，$\lambda = 0.05$，ARI=0.97

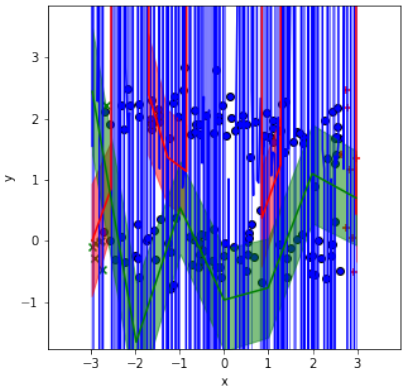

図 11: データ 1，GPSCRM による出力，$\theta = 1.5$，$\alpha = 5$，$\lambda = 0.5$，ARI=0.027

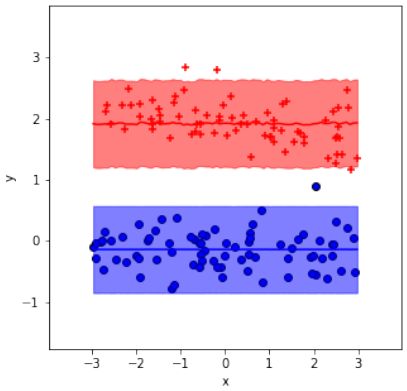

図 12: データ 1，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.0001$，$\lambda = 0.5$，ARI=0.97

4.3.2　データ 1 の実験結果からの考察

本節では，SCRM で良好な結果が得られなかった理由について説明する．逐次抽出法では，クラスタ内のデータのサイズが均衡で，クラスタ内のデータにばらつきのあるデータに対しては良好な結果を得られないからである．逐次抽出法はノイズクラスタリングを基にした手法であり，抽出クラスタかそれ以外のクラスタかの 2 つに分類する．そのため，抽出クラスタに属するデータのみで回帰線を求めることになり，データ数が多いクラスタに偏る傾向がある．従って，SCRM では良好な結果が得られなかったと考える．

4.3.3　データ 2 に対する実験結果

データ 2 に対して，SCRM，GPSCRM を用いた結果をそれぞれ図 13，14 に示す．また，図 15，16 の実験結果は，GPSCRM に対して α，λ を変えた場合の出力結果である．データ 2 は不均衡なデータであり，クラスタ同士でデータのサイズが異なる．本実験の目的は，不均衡データに対して GPSCRM がクラスタリングできるかどうかを示すことである．SCRM は ARI=0.95 である．SCRM の回帰線に関して，上側のクラスタが下側のクラスタの影響を受けていることが確認できる．

GPSCRM では ARI=0.91 となっている．GPSCRM の回帰線に関して，上側のクラスタでは，尖った回帰線を持つのに対して，下側のクラスタでは滑らかな構造を持つ．また，GPSCRM では，非線形な回帰結果を得られていることも図 14 から確認できる．

次に，図 14 に対して，α と λ の値を小さくした結果として，図 15 に示す．○の回帰線については，図 14 と同様の結果になっている．しかし，＋ の回帰線に関しては，水平な直線が得られた．また，図 16 は，正則化項を大きくした場合の出力結果である．分散の値が大きく表示されている．また，全ての回帰線に対して，水平な直線が得られた．

4.3.4　データ 2 の実験結果からの考察

本節では，逐次抽出法が不均衡なデータに関して良好な結果を得られた原因について述べる．また，GPSCRM の ARI が SCRM に比べて低下した理由についても述べる．逐次抽出法は，元々ノイズクラスタリ

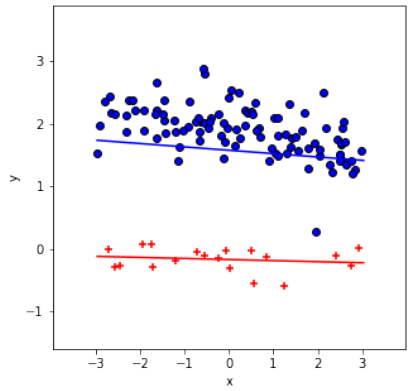

図 13: データ 2，SCRM による出力，$\theta = 5.0$，ARI=0.95

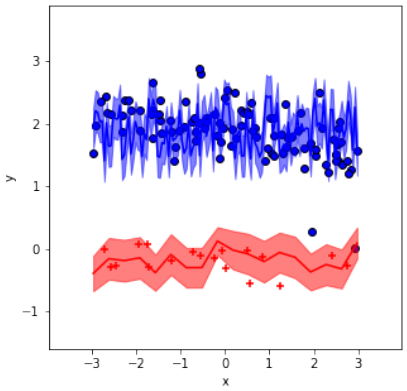

図 14: データ 2，GPSCRM による出力，$\theta = 3.0$，$\alpha = 0.04$，$\lambda = 0.1$，ARI=0.91

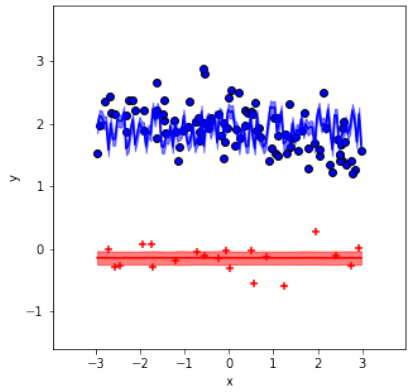

図 15: データ 2，GPSCRM による出力，$\theta = 3.0$，$\alpha = 0.001$，$\lambda = 0.01$，ARI=1.0

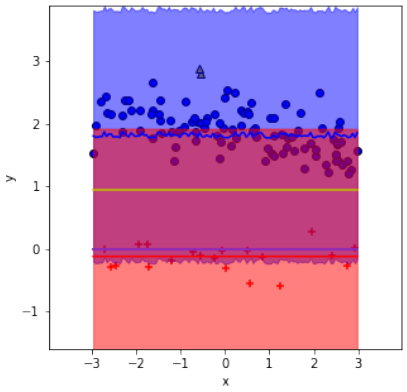

図 16: データ 2，GPSCRM による出力，$\theta = 3.0$，$\alpha = 0.001$，$\lambda = 4$，ARI=0.90

ングのために用いられた手法であるため，不均衡なデータに対して有効にクラスタリングを行うことができると考えられる．抽出クラスタかノイズクラスタかに分類する際，抽出クラスタに属するデータのみで回帰線を考える．そのため，得られる回帰線は，抽出クラスタ以外の構造を考慮しない．次に，GPSCRM の ARI 低下理由について述べる．これは，GPSCRM の期待値の値に強く正則化がかかり，尖った回帰線になっていることが起因している．そのため，上側の期待値のプロットが Y=1 付近に下がっているところで誤分類が発生している (図 14)．

4.3.5　データ 3 に対する実験結果

データ 3 に対して，SCRM，GPSCRM を用いた結果をそれぞれ図 17，18 に示す．また，GPSCRM について，パラメータを変えた結果を図 19，図 20 に示す．データ 3 は，クラスタ数が 3 つの場合のデータである．本実験の目的は，クラスタ数が 3 つの場合に GPSCRM がクラスタリングできるのかを示すことである．SCRM に対するクラスタリングは，ARI=0.74 であり，良好な結果を得られなかった (図 17)．次に，

GPSCRM に対するクラスタリングに対しても，ARI=0.67 であり，良好な結果は得られなかった (図 18)．GPSCRM の回帰線に関して，1 番上のクラスタ×には正則化パラメータ λ が働いていないのに対し，真ん中のクラスタ○には正則化が強く働き，尖った回帰線となっている．1 番下のクラスタ + は，その中間をとったような回帰線である．

ここで，α の値を大きくした場合の出力結果として，図 19 に示す．○のクラスタに関する回帰線が過学習を起こしていることが確認された．また，α を小さくした場合の出力結果として，図 20 に示す．全ての回帰線に対して，水平な直線が出力された．

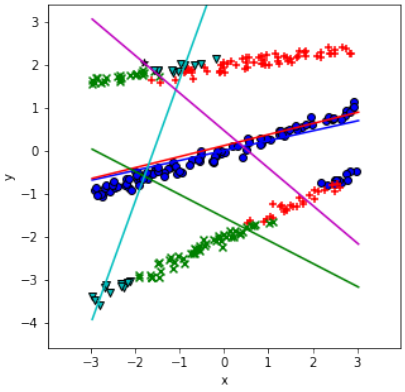

図 17: データ 3，SCRM による出力，$\theta = 1.5$，ARI=0.74

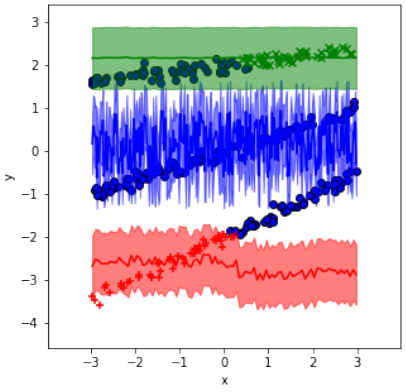

図 18: データ 3，GPSCRM による出力，ARI=0.67，$\theta = 1.5$，$\alpha = 0.001$，$\lambda = 0.5$

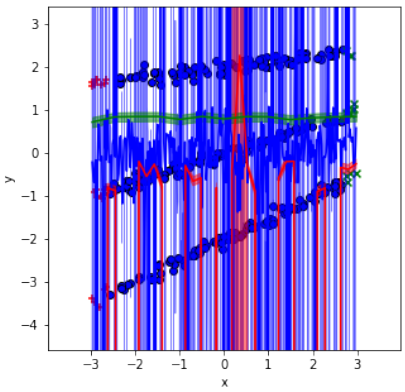

図 19: データ 3，GPSCRM による出力，$\theta = 1.5$，$\alpha = 1.0$，$\lambda = 0.01$，ARI=0.045

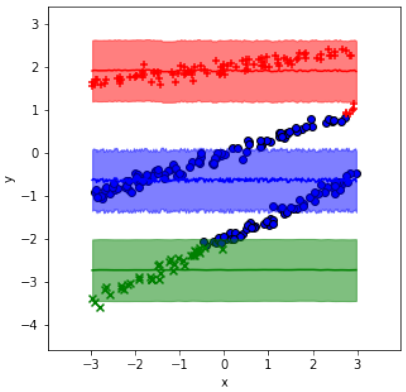

図 20: データ 3，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.0001$，$\lambda = 0.5$，ARI=0.37

4.3.6　データ 3 の実験結果からの考察

本節では，データ 3 に対して，GPSCRM が良好なクラスタリング結果が得られなかった理由を 2 つ述べる．理由の 1 つとして，GPSCRM に対して適切にハイパーパラメータを与えることが困難なことにある．逐次

抽出法では，抽出クラスタとノイズクラスタの２つに分類した後に，データが空でなければ，抽出クラスタに属するデータを取り除く．そのため，抽出ごとに適合するハイパーパラメータが変化するため，抽出するたびにハイパーパラメータを更新する必要がある．しかし，現状ではノイズパラメータ θ 以外のパラメータについて更新手法を確立できていない．実際に図 18 において，正則化パラメータがうまく機能していない．正則化パラメータとは，データの過学習を調整するパラメータであり，データのサイズに対して正則化パラメータが大きいと，期待値が横に真っ直ぐプロットされる．図 18 において，3 回目に抽出を行った×点に対して，同様の現象が起こっている．そのため，今後の課題として，正則化パラメータ λ を適切に更新する式についての研究を進める必要がある．これは，カーネルパラメータ α にも言えることである．カーネルパラメータ α は，最尤推定で計算する方法 [8] が研究されており，今後の研究でどのように実装するのかを決める必要がある．

もう 1 つの理由として，逐次抽出法ではクラスタ内のデータのサイズが均等の場合に，有効にクラスタリングが行えないことにある．逐次抽出法は，基本的に抽出クラスタかノイズクラスタかの２つに分類する．その上，抽出クラスタに属するデータに関する回帰線から，ノイズクラスタを判定するため，抽出クラスタ以外のデータ構造に関しては考慮できない．実際に，データ 3 に対して SCRM を用いた場合，良好な結果を得られていない (図 18)．従って，クラスタ数が 3 つである本データに対して，GPSCRM では良好な結果を得ることはできないと考察する．

4.3.7　データ 4 に対する実験結果

データ 4 に対して，SCRM を用いた結果をそれぞれ図 21 に示す．また，GPSCRM について，パラメータを変えた結果を図 22，図 23，図 24，図 25，図 26 に示す．データ 4 は非線形な構造を持つクラスタと線形な構造を持つクラスタの２つで構成されているデータである．本実験の目的は，非線形なデータに対して，GPSCRM が有効にクラスタリングができるかを確かめることである．

まず，SCRM に対する実験結果は ARI=0.50 であり，良好なクラスタリング結果は得られなかった．また，回帰線についても，非線形な構造を得ることはできなかった．

次に，GPSCRM に対する実験結果として，パラメータが $\theta = 1.5, \alpha = 0.09, \lambda = 0.80$ の場合，ARI=0.80 となり，SCRM に比べて良好なクラスタリング結果が得られた (図 22)．しかし，上側のクラスタ○は過学習が起こっており，上下に尖った回帰線が得られた．下側のクラスタ + は，比較的滑らかな回帰線が得られたことが図 21 から確認できる．上側の回帰線については，非線形な構造を捉えているとは言えない結果となった．

ここで，GPSCRM のパラメータを変えた場合の出力結果を図 23，図 24，図 25，図 26 に示す．図 23，24 は，図 22 のパラメータに対して，λ の数値のみを変更した場合の出力結果である．図 23 では，λ の数値を大きくした場合の出力結果である．図 22 に比べて，分散を表すプロットが大きく出力されている．また，回帰線については，図 22 と似た結果になっており，ARI も同様の値となった．次に，図 24 は，λ の数値を小さくした場合の出力結果である．クラスタ数が 3 つに分類されており，ARI が 0.37 と低い結果になった．分散を表すプロットが，図 23 と比べて，小さくなっている．

次に，図 25，26 は，図 22 のパラメータに対して，α の数値のみを変更した場合の出力結果である．図 25 は，α の数値を大きくした場合の出力結果である．○のクラスタの抽出過程において，分散が大きい値を持つため，発散していることが確認された．また，ARI が 0.021 と非常に低い結果になった．次に，図 26 は，α の値を小さくした場合の出力結果である．横に水平な回帰線が得られた．また ARI が 0.37 と低い結果になった．

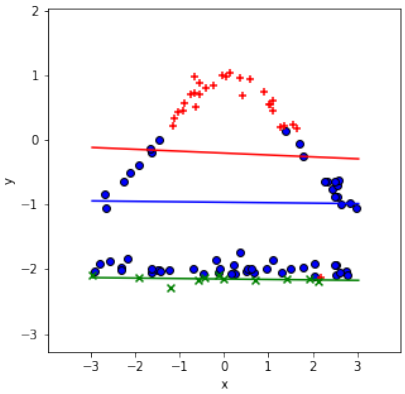

図 21: データ 4，SCRM による出力，$\theta = 2.0$，ARI=0.50

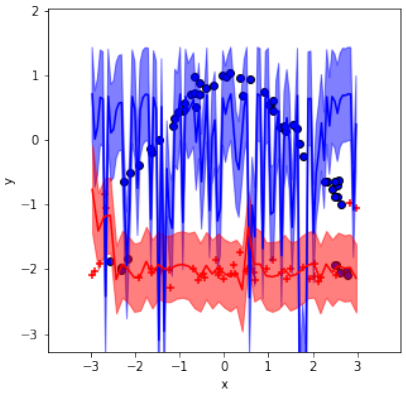

図 22: データ 4，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.09$，$\lambda = 0.5$，ARI=0.80

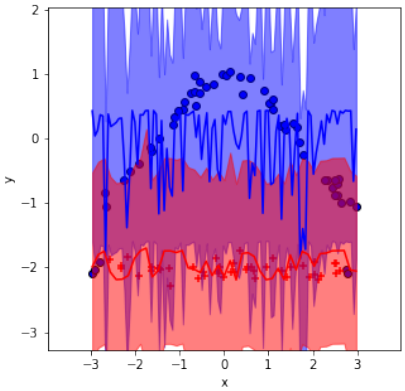

図 23: データ 4，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.09$，$\lambda = 4.0$，ARI=0.80

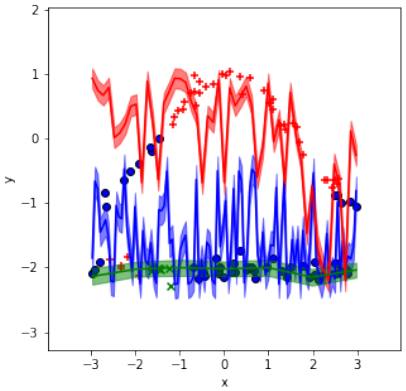

図 24: データ 4，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.09$，$\lambda = 0.045$，ARI=0.75

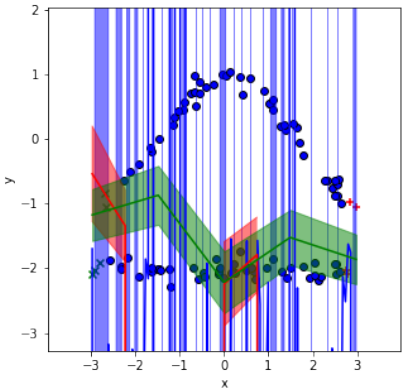

図 25: データ 4，GPSCRM による出力，$\theta = 1.5$，$\alpha = 5.0$，$\lambda = 0.5$，ARI=0.021

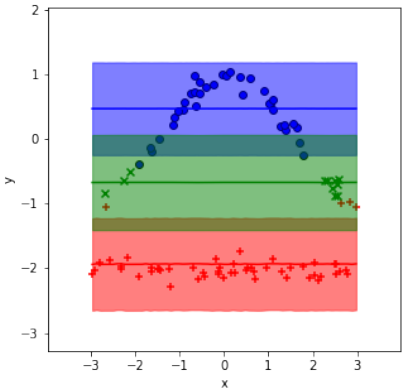

図 26: データ 4，GPSCRM による出力，$\theta = 1.5$，$\alpha = 0.0001$，$\lambda = 0.5$，ARI=0.37

4.3.8　データ 4 の実験結果からの考察

本節では，GPSCRM を用いて，非線形な回帰線が得られなかった原因について考察を行う．非線形な回帰線が得られなかった理由として，カーネルパラメータ α が有効に働いていないことが考えられる．α は，回帰線の形状に影響を与えるパラメータである．図 25 によると，α を大きくした場合，回帰線がジグザグな構造をとる．これは，カーネル関数の値が大きいことが原因であると考えられる．反対に，図 26 では，α を大きくした場合，横に水平な直線を回帰線をとる．これは，ガウスカーネルの式 (1) より，カーネル関数が 0 に近い値を持つことが原因であると考えられる．今回用いたガウスカーネルは，ハイパーパラメータが 1 つである．他にも，ハイパーパラメータを複数持つカーネル関数が提唱されており，得られる非線形構造はカーネル関数によって異なる．そのため，今後の課題として，様々なカーネル関数を用いて実験を行う必要がある．カーネル関数に関するハイパーパラメータをいくら増やしても，最尤推定によって求まることが分かっている [21]．そのため，最適なカーネルパラメータ α を最尤推定を用いて計算するができる．

正則化パラメータ λ についても実験を行った．図 23 は，図 22 の場合と比べて，λ の値を大きくした場合の出力結果である．分散の値が極端に大きくなっており，これは正則化により，カーネル行列の各要素の値が大きくなっていることが原因である．しかし，回帰線についてはあまり変わっていない．また，図 24 は，図 22 の場合と比べて，λ の値を小さくした場合の出力結果である．分散の値が小さくなっており，こちらでは，回帰線が大きく変わっていることがわかる．

現在の研究では，3 つのハイパーパラメータを総当たりで求めている．逐次抽出法では，抽出ごとにデータのサイズが変化するため，抽出ごとのハイパーパラメータも必要となる．そこで．抽出ごとに最尤推定でカーネルパラメータを求めることで，パラメータ決定を簡略化することができる．ただし，局所解に陥りやすいという問題があり，実装に至っていない．今後の研究で改善できる点であると考える．

4.4　実験全体を通しての考察

提案手法であるガウス過程に基づく逐次抽出型 c-回帰モデル (GPSCRM) は，既存手法である逐次抽出法 (SCRM) の性能に加え，ガウス過程回帰の特徴を持つ．ガウス過程回帰の特徴として，非線形な回帰線やガウス過程の予測分布から得られる分散がある．しかし，データ 4 による実験では，良好な非線形な回帰線を得ることができなかった．これは，クラスタを抽出するごとに，データのサイズが減少し，適切なカーネルパラメータ α の推定が難しい点ことが原因にある．クラスタを抽出するたびに，カーネルパラメータを最尤推定によって求めることが必要と考えられる．しかし，最尤推定によって生じる，カーネルパラメータに対する最適化問題を解く際に，局所解に陥りやすいという問題がある．そのため共役勾配法，MCMC 法 [8] などの最適化手法を比較実験する必要があると考えられる．

また，データ 1，2 において，不均衡なデータに対して逐次抽出法が有効である可能性が示唆された．これは，ノイズクラスタリング [10][18] に基づいていることが原因として考えられる．そのため，不均衡データを含め，小さなノイズを多く持つデータに対して有効に分類できる可能性が示唆された．今後の研究で，小さなサイズのノイズを多く持つデータに対してのクラスタリングの有効性を検証する必要がある．

ここで，パラメータの値を変えた実験により，α と λ の値がどういった影響を与えるかどうかについて考察を行う．まず，α の値を変えることにより，回帰線の形状が変化することが示された．α の値を極端に大きくした場合，回帰線が過学習を起こすことが確認された．これは，ガウスカーネルの値が極端に大きくなることにより，期待値 $\boldsymbol{k}_*^{(1)T}\boldsymbol{K}^{(1)-1}\boldsymbol{Y}^{(1)}$ の値が大きく増減することが原因であると考えられる．逆に小さくした

場合，水平な直線構造を持つ回帰線が得られることが確認された．これに関しても，カーネル関数の値が小さくなることにより，期待値 $\boldsymbol{k}_*^{(1)T}\boldsymbol{K}^{(1)-1}\boldsymbol{Y}^{(1)}$ の値が同じ値に収束することが原因であると考えられる．次に，λ の値を変えることにより，標準偏差の値を表す帯のプロットに対して影響が出ることが示唆された．また，α と λ は密接な関係を持つ．カーネル行列 $\boldsymbol{K}$ は，両者の影響を受けるため，α と λ の微妙な組み合わせによって回帰線が求まる．そのため，手動で適切なパラメータを求めることが困難であることが本実験より分かった．

5 結論

本論文では，逐次的にクラスタリングと回帰を同時に行う逐次抽出型 c-回帰モデル (SCRM) にガウス過程を導入し，ガウス過程に基づく逐次抽出型 c-回帰モデル (GPSCRM) を提案した．また，数値実験より，提案手法である GPSCRM と既存手法である SCRM との比較を行うことで提案手法の有用性を確認した．

提案手法は，既存手法では得られない，非線形な回帰線を得られることを目的とした手法であるが，数値実験では良好な非線形な回帰線を得ることはできなかった．原因として，カーネルパラメータ α および正則化パラメータ λ の最適な値を求めることが困難なことにある．そのため，それぞれのパラメータについて最適化手法を用いて，パラメータ推定を行う手法の構築が今後の課題として考えられる．

謝辞

本研究を進めるにあたり，ガウス過程に関する書籍の輪講，研究の相談，および発表練習など多くの熱心なご指導をしていただいた本近畿大学理工学部情報学科の知能情報基礎研究室の濵砂幸裕講師講師には，深く感謝いたします．

また，1 年の間，ガウス過程に関する研究を共に切磋琢磨してきた横山裕哉君をはじめ，知能情報基礎研究室の皆様に，心より感謝申し上げます．

最後に，大学 4 年間，経済面，精神面，生活面で支えていただいた両親に深い感謝を申し上げます．

参考文献

[1] A. K. Jain, “Data clustering: 50 years beyond K-means”, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651–666, 2010.

[2] 久保拓弥, ‘データ解析のための統計モデリング入門’, 岩波書店, 2012.

[3] R. J. Hathaway and J. C. Bezdek, “Switching Regression Models and Fuzzy Clustering”, *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 3, pp. 195–204, 1993.

[4] 中森義輝, “ファジィクラスタリングと回帰分析”, 日本ファジィ学会誌, Vol. 8, No. 3, pp. 431–439, 1996.

[5] 金月優斗, “ガウス過程に基づく c-回帰モデル”, 近畿大学大学院総合理工学研究科エレクトロニクス系工学専攻令和 2 年度修士論文, 2020.

[6] C. C. Kung, H. C. Ku, and J. Y. Su, “Possibilistic c-Regression Models Clustering Algorithm”, *IEEE International Conference on System Science and Engineering*, pp. 297–302, 2013.

[7] 大井祐介, 遠藤靖典, “カーネルファジィ c-回帰法について”, 第 33 回ファジィシステムシンポジウム講演論文集 (FSS2017 山形大学), pp. 325–330, 2017.

[8] 持橋大地, 大羽成征, ‘ガウス過程と機械学習’, 講談社, 2019.

[9] 赤穂昭太郎, “ガウス過程回帰の基礎”, システム/制御/情報, Vol. 62, No. 10, pp. 390–395, 2018.

[10] R. N. Davé, and R. Krishnapuram, “Robust Clustring Methods: A Unified View”, *IEEE Transactions on fuzzy system*, Vol. 5, No. 2, pp. 270–293, 1997.

[11] S. Miyamoto, Y. Kuroda, K. Arai, “Algorithms for Sequential Extraction of Clusters by Possibilistic Method and Comparison with Mountain Clustering”, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 12, No. 5, pp. 448–453, 2008.

[12] S. Miyamoto, K. Arai, “Different Sequential Clustering Algorithms and Sequential Regression Models”, *IEEE International Conference on Fuzzy Systems*, pp. 1107–1112, 2009.

[13] Y. Hamasuna, and Y. Endo, “On Kernelized Sequential Hard Clustering”, *International Symposium on Advanced Intelligent Systems*, pp. 416–419, 2016.

[14] R. R. Yager, and D. P. Filev, “Approximate Clustering Via the Mountain Method”, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 8, pp. 1279–1284, 1994.

[15] L. Hubert, “Comparing Partitions”, *Journal of Classification*, Vol. 2, No. 1, pp. 193–218, 1985.

[16] 春山秀幸, 遠藤靖典, 大久保貴義, “カーネル関数を用いた階層的クラスタリング”, 知能と情報 (日本知能情報ファジィ学会誌), Vol. 17, No. 4, pp. 459–467, 2005.

[17] 赤穂昭太郎, ‘カーネル多変量解析’, 岩波書店, 2008.

[18] R. N. Davé, “Characterization and detection of noise in clustering”, *Pattern Recognition Letters*, Vol. 12, No. 11, pp. 657–664, 1991.

[19] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, “Gaussian Processand Kernel Methods”, *A Review on Connections and Equivalences*, 2018. arXiv:1807.02582[Stats.ML].

[20] M. Sander, “Market timing over the business cycle”, *Journal of Empirical Finance*, Vol. 46, pp. 130–145, 2018.

[21] C. K. I. Williams, and C. E. Rasmussen, “Gaussian Processes for Regression”, *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, pp. 514–520, 1995