

博士学位論文

生成的モデリングによる 集合データの Visual Analytics

令和 4 年 3 月

九州工業大学大学院生命体工学研究科

渡辺 龍二

論文要旨

本論文は集合データの Visual Analytics (VA) を実現する汎用的方法論を提案するものである。VA とは、人間とデータ分析システムが視覚的インターラクションを介してデータ分析・仮説検証・意思決定を行う過程を指す。すなわち VA はデータ駆動型意思決定における人間主導的アプローチであり、全自動化型の AI 主導的アプローチの対極としてその重要性が高まっている。本論文は VA の中でも未開拓な領域である集合データの VA システムを対象とする。集合データの典型例はビジネスやスポーツにおけるチームデータであり、既存チームの分析と新規チームのメンバー選択支援が VA システムの主な用途となる。本研究の目的は、このような集合データ VA システムを実現する汎用的な方法論を確立することである。

本論文の主な貢献は次の通りである。(1) 集合データ VA システムが満たすべき要件を明確化し、要件を満たす上で克服すべき困難点を明らかにした。(2) 多様体上の確率分布表現を用いた集合データのモデル化法を実現した。これにより既存チームの分析と新規チームの生成・予測が可能となった。(3) さまざまなデータ構造に適応可能な多様体ネットワークモデルを提案した。(4) ユーザの関心標的の指示により駆動されるインターラクティブな視覚的インターフェースを実現した。これにより複数要因が絡ったデータに対する対話的分析が可能になった。(1)(2) により集合データ VA の方法論が確立され、(3)(4) により汎用的 VA システムの構築法が確立された。これらは VA の分野において今まで達成されなかった成果である。

本論文の構成は次の通りである。第一章では序論として人間主導的データ駆動アプローチとしての VA について概説し、さらに集合データ VA の必要性と課題、提案手法のキー アイディアと本論文の貢献について述べる。

第二章では本研究の背景と関連研究について三つの観点から述べる。第一は VA の観点であり、VA の枠組みや現在までの研究動向について詳しく解説する。第二は機械学習の観点であり、集合データを扱う機械学習の困難点や既存のアプローチ、および最近の研究動向について解説する。とりわけ本研究と関わりの深い集合データの生成モデルについて詳しく述べる。第三はデータ駆動型のチーム編成支援の観点であり、既存の研究群を自動化型・情報提示型・VA 型の三種に分類して概説する。

第三章では集合データ VA に求められる要件を定義する。集合データ VA という概念自体が本研究独自のものであるため、その概念と要件を定義し、さらに実現上の技術的課題を明らかにする。具体的には集合データ VA のシステムが満たすべき四つの要件と、システムの汎用的構築手法が満たすべき一つの要件を定義する。

第四章では集合データ VA システムの構築法を提案する。まず多様体上の確率分布表現を用いた集合データのモデリング法を述べ、次に複雑なデータ構造に適応可能な多様体ネットワークモデリングへの拡張を述べる。さらに対話的な可視化を実現する視覚的インターフェースの構築法についても述べる。

第五章は提案手法のデモンストレーションである。提案手法を用いてバスケットボールチームの VA システムを構築し、過去のメンバー構成とゲーム成績のデータ分析や、新規ゲームにおけるメンバー選択支援などをデモンストレーションする。また比較手法となる VA システムを構築し、提案システムが十分な能力を持つことを実証する。

第六章は議論である。第三章で定義した要件の妥当性の検証や、提案手法のさらなる拡張についての検討を行う。また提案手法のデータモデリングの枠組みと、既存の機械学習のパラダイムとの関連についても述べている。

第七章は総括として本論文をまとめる。

以上、本論文では集合データ VA システムの実現を提案するとともに、汎用的な VA システムの構築方法を提案した。本研究の意義は単一用途の VA システムを開発したことではなく、集合データを含むさまざまなデータに適応可能な汎用的 VA システムの構築法を実現したことにある。これは用途特化型の VA システム開発が多い VA 研究領域においては稀有な試みといえる。すなわち本研究は人間主導型データ駆動アプローチの新たな基盤構築をめざしたものである。

目次

論文要旨	i
第 1 章 序論	1
1.1 本研究の背景と目的	1
1.1.1 データ駆動型意思決定	1
1.1.2 Visual Analytics (VA): 「人間主導」なデータ駆動型意思決定 .	2
1.1.3 集合データ VA とメンバー選択問題	4
1.2 本研究のゴールとキーアイディア	6
1.3 本研究の貢献と本論文の構成	7
第 2 章 背景と関連研究	9
2.1 VA	9
2.1.1 VA の枠組み	9
2.1.2 VA の定義	11
2.1.3 VA システムが提供する可視化とインタラクション	13
2.1.4 Predictive VA (PVA)	15
2.2 集合データを扱う機械学習	17
2.2.1 集合データの定義	17
2.2.2 集合データを扱う困難	17
2.2.3 代表的なタスク	18
2.2.4 代表的なアプローチ	19
2.2.5 集合データの生成モデル	20
2.3 チームマネジメントにおけるメンバー選択支援	25
第 3 章 集合データ VA が満たすべき要件	27
3.1 必要要件の定義	27
3.1.1 チーム編成においてどのような VA システムが必要とされるか .	27

3.1.2 必要要件の一般化	29
3.2 必要要件充足のための困難とチャレンジ	30
第 4 章 生成的多様体ネットワークモデリング	33
4.1 多様体モデル	33
4.2 提案手法の枠組み	35
4.3 問題設定	38
4.4 実装	39
4.4.1 生成的多様体モデリング (GMM)	39
4.4.2 GMM を用いた多様体ネットワークモデルの推定	40
4.5 可視化法	42
4.5.1 視覚的インターフェースの構築	42
4.5.2 ユーザの TOI の指定に対する可視化	43
4.5.3 新規チームのシミュレーション	46
4.6 提案手法が要件を満たすことの確認	46
第 5 章 デモンストレーション	49
5.1 データセットと前処理	49
5.2 モデルの学習	51
5.3 構築したシステム	53
5.4 デモンストレーション	55
5.5 システムが要件を満たすことの確認	58
5.6 予測タスクと生成タスクの定量評価	60
第 6 章 議論	63
6.1 提案手法の評価と 5 つの要件	63
6.2 提案手法と他の学習パラダイムとの関連	64
6.3 提案手法の拡張	65
第 7 章 総括	67
参考文献	69
研究業績リスト	79
謝辞	83

第1章

序論

1.1 本研究の背景と目的

1.1.1 データ駆動型意思決定

さまざまな分野において、データを活用することで有効な意思決定を下す、いわゆるデータ駆動型意思決定が重要視され始めている。ビジネスの領域はその傾向が特に顕著であり [1,2]、データ駆動型意思決定のための組織やシステムの改革 (Digital Transformation: DX) に関する調査結果 [3] によると、データ利活用について何らかの計画を立てている、もしくは既に取り組んでいる企業の割合は 97% に達している。また DX について成熟した計画の立案や多額の投資を行っていると評価された企業 (Digital Adopters) の割合は 2016 年に 14% であったのに対して 2020 年には 39% と 4 年間で 25% 増加している。このことから各企業はここ数年でデータ活用のための内部改革を加速的に進めているのが分かる。その他にも政治 [4]、スポーツ [5]、医療 [6] といった領域でデータ駆動型意思決定への注目は高まり続けている。

そのようなデータ駆動型意思決定として代表的なアプローチの 1 つは、意思決定の自動化、つまりコンピュータシステムによる意思決定の代理である。自動化が盛んに推進されている場面の具体例としては、金融における融資の審査 [7] や Web サービスにおける消費者へのレコメンドや広告配信 [8] などが挙げられる。つまり意思決定を大量に行う必要があり、個々の意思決定の重要性が比較的低い場面において自動化が推進される傾向がある。このような場面における意思決定の自動化は、効率化や属人性の解消といった効果が期待できる。自動化を実現するためには、システム上に意思決定の機構を実装する必要がある。従って意思決定を自動化するための最低限の条件として次の 2 つが挙げられる。1 つ目は定式化可能であることで、最適化したい定量的な基準が明確である必要がある。2 つ目は状況の判断とそれに伴う意思決定のルールが分かっている、もしくは大量に存在す

るデータに暗黙的な判断のルールが内在していると見なせる状況である^{*1}.

一方で実社会の意思決定の場面では、自動化することができない、もしくは自動化することが不適切な場合もある。前者の例としては、定式化できるほど問題を具体的に定義することが難しい場合である。例えば最適化したい基準が複数あり、それらはトレードオフの関係にあるような状況である。このような場合は、人間がトレードオフの妥協点を探る、もしくはトレードオフの打開策を検討する必要がある。一方で後者の例としては、定量化できない情報が重要である場合も挙げられる。具体的には衣服の販売戦略の策定のような状況では、衣服についての定量化できる情報よりも衣服から受ける印象といった定性的な情報が重要なことは大いに考えられる。従ってこのような状況では人間の目で確認するというプロセスがデータ分析の中に入っていることが望ましいと考えられる。また別の自動化が不適切な場面の例としては、一度の意思決定の重要度が極めて高い場合も挙げられる。具体的な場面としては企業の経営上の判断や政府の政策決定などである。このような問題の担当者は2種類の「責任」が問われる[10]。1つ目は意思決定前に問われるその意思決定の結果に対する責任(Responsibility)である。2つ目は結果が出た後でそれを説明する責任(Accountability)である。よってこの場面においても、人間を介さない完全な自動化は適していないと考えられる。

1.1.2 Visual Analytics (VA): 「人間主導」なデータ駆動型意思決定

ここまで議論を踏まえて、複雑かつ重大な問題に対するデータ駆動型の意思決定の形として考えられるのは、人間自身の意思決定をデータ分析によって支援するシステムの構築である。より具体的には、システムがデータ分析結果を提示することで、推論・予測といった問題解決に有効な認知能力を人間から引き出し、人間自身がデータについて理解を深めながら最終的な意思決定を下せる、そのような意思決定のプロセスを実現することである。

このような動機のもとで発展を続けてきたのがVisual Analytics (VA)[11]である。本論文においてVAとは「コンピュータと人間との間に生じる視覚的なインタラクションを通して、探索的なデータ分析・仮説検証・意思決定を人間が主導して行うための方法論や技術」と定義する。ここで、データを入力として、意思決定の結果を出力する「データ駆動型意思決定システム」を考えてみる。先程の完全自動化型とVA、それぞれのアーキテクチャを比較しているのが図1.1である。完全自動化型の場合、意思決定システムはコンピュータシステムそのものである。データはまず統計手法や機械学習手法のような自動分

^{*1} 「大量に存在するデータに暗黙的な判断のルールが内在していると見なせる状況」において意思決定を自動化する際は、機械学習やデータマイニングの技術によってデータから帰納的に判断機構を構築し、数理最適化技術を利用して意思決定機構を構築することになる[9]

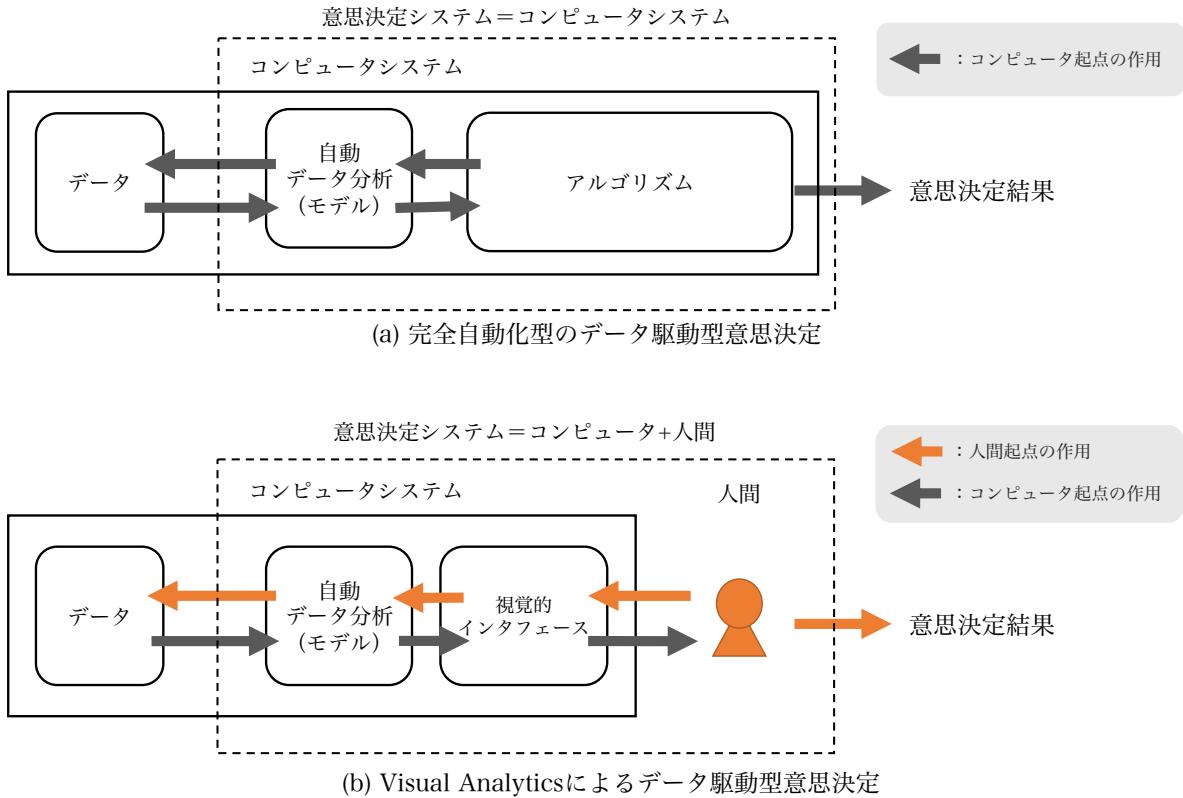


図 1.1 完全自動化型のデータ駆動型意思決定 (a) と VA によるデータ駆動型意思決定 (b) の比較。前者は人が介在することなく、データを元に自動分析技術（モデル）やアルゴリズムを利用して自動的に意思決定を行う。これに対して後者はデータ分析や意思決定を人が起点となって行う「人間主導」なプロセスとなっている。

析技術によって処理され、意思決定に必要な情報が取り出される。その情報を元に、事前に定義されたルールもしくは数理最適化技術のようなアルゴリズムが意思決定を行う。これに対して VA では、コンピュータシステムのみならず、人も含めた二者の総体を意思決定システムであると捉えている。コンピュータシステムは単純なフィルタリング機能や統計手法、機械学習手法を備えたデータの自動分析機構（モデルと呼ぶ）と、それと人間との間を取り持つ視覚的インターフェースから構成される。VA ではコンピュータシステムがデータを分析し、その結果を視覚的インターフェース上において可視化する。その結果を知覚したユーザは、可視化結果から連想されるさらに分析したい関心の対象や観点 (Target of Interest: TOI) をインターフェース上での操作によってコンピュータシステムにフィードバックする。このコンピュータシステムとユーザのインタラクションを繰り返しながら人はデータを探索し、仮説を検証し、知識を得た上で最終的な意思決定を行う。即ち VA とは、人が起点となってデータ分析のループと最終的な意思決定を稼働させる、「人間主導」な意思決定システムを目指す枠組みである。

このような VA の枠組みは、前述した複雑かつ重大な意思決定問題に有効であると考え

られる。データ活用による複雑な問題の支援においては、そもそもどのデータにどのような分析を行えば良いのかを自動的に決めるることは難しい。従って複雑な問題に対しては、VA のようにデータ分析の観点を人間が動的に選択し、それを分析技術にフィードバックするという人間の認知能力を活かした枠組みが有効であると考えられる。また、最終的な意思決定も「人間主導」であるという点は、前述した重大な問題につきまとう 2 つの責任の観点からも重要である。

1.1.3 集合データ VA とメンバー選択問題

本論文は VA の枠組みにおいて「集合データ」と呼ばれるデータ構造を扱うためのコンピュータシステム（以降 VA システムと呼ぶ）を構築する方法を議論する。この集合データを取り扱う VA のことを本論文では集合データ VA と呼ぶことにする。ここで集合データとは、要素として数ベクトルで表現された特徴量を取る集合のことであり、この集合という単位が分析の対象となる。この集合データが複数得られた場合、データセット全体は特徴量集合の集合となる。集合データはさまざまな場面で見られるデータ構造でありながら扱いが難しい対象であることが知られており、近年ではその取り扱い方が機械学習の領域を中心に活発に議論されている [12, 13]。本論文はこの集合データの集合を含むデータセットに対して、VA の枠組みによる人間主導なデータ分析・仮説検証を実現する意思決定支援システムの構築方法を検討する。

本論文が集合データ VA の実現を目指す動機の 1 つは、メンバー選択という集合データが登場する典型的な問題に対して、人間主導型の支援を実現することである。メンバー選択とは選択可能なメンバーの集合が与えられたときに、その出力が最大となるように部分集合を選択する問題である。このような場面の具体例としてチーム編成が挙げられる。チーム編成では選択可能なメンバー^{*2}の集合から、その成果が最大となるようにチームを構成するメンバーを決める必要がある。このとき個々のメンバーの能力が数ベクトルとして表現されている場合、メンバーの集合であるチームは集合データとして表現される。また別のメンバー選択の具体例としてはファッショナウットフィット構成 [14] が挙げられる。この問題では、選択可能なファッショナアイテムの集合が与えられたときに、受け取り手が満足するようにアウトフィット（同時に着用しているアイテムの一式）を構成するアイテム集合を選択する。このとき個々のアイテムの特徴が数ベクトルとして表現されれば、アウトフィットは集合データとして表現される。

メンバー選択は、VA で取り組むべき問題として前述した重大かつ複雑な問題に該当する。チーム編成を具体例としてメンバー選択の重大性について述べると、組織というのは

^{*2} メンバー選択におけるメンバーは数学における集合の要素という意味での member を指している。これに対してチーム編成におけるメンバーは、チームを構成する人員のことを指している。

分野を問わずチーム単位で活動するため、取り組む活動によっては編成したチームの成果が組織の命運を左右する。従って活動前の時点でいかに良いチームを編成するかということが重要になる。当然のことながら、編成を行うチームのマネージャには Responsibility と Accountability という 2 つの責任が生じてくる。従って、VA のように最終的な判断がマネージャに委ねられることが必須となる。

一方でメンバー選択の定式化し難いほどの複雑さには次の 3 つの側面がある。1 つ目はメンバーの組合せと出力の関係が非常に複雑であること、2 つ目はメンバーの組合せ以外の要因を考慮しなければならない場合があること、3 つ目はメンバーの組合せに対する評価が主観的かつ定性的に与えられる場合があることである。チーム編成を具体例として詳しく述べると、1 つ目のメンバーの組合せと出力との間の関係が複雑である理由は、シナジー効果 [15–17] にある。シナジー効果とは、メンバーの組合せによって、個々人が出す成果の総量をチームが出す成果が上回る、もしくは下回る現象のことである。そのため、単純に個々人の能力を表現する特徴量の総和や平均を取る処理を行うのではなく、メンバーの組合せの情報を保った表現、つまり集合データを扱う必要がある。しかしながら、集合データを入力とした関数のモデル化は決して容易ではない [12,13]。2 つ目のメンバーの組合せ以外の要因とは、チームが取り組むタスクの種類や外的要因（チームスポーツであれば敵チーム、ビジネスであればマーケットの状況など、マネージャやメンバーが制御できない要因）に相当する。現実的なチーム編成の状況を定式化する場合、これらの要因を考慮することが求められる。3 つ目のメンバーの組合せに対する評価が定性的・主観的に与えられる状況の典型例はファッショナウトフィット構成である。この問題ではアイテム同士の相互的な相性の良さを考慮しながらアイテムの集合を選択する必要があるが、この場合「相性が良い」という判断はアウトフィットの受け取り手によって主観的かつ定性的に行われるものである。このような場合に問題を定式化し計算機上で実装することは困難である。

このようにメンバー選択とは複雑かつ重大な問題であり、このような問題にこそ VA のアプローチは有効である。チーム編成を例に VA アプローチの利点を挙げると、チームマネージャは VA システムとのインタラクションを通して、過去のデータの分析結果からチームの強みや弱みについて仮説を立てることが可能になる。またマネージャはその仮説についてデータを探索し検証を行うことができる。その結果として、マネージャはシステムから得た知識とユーザ自身が持つ経験的な知識を統合し、最終的な決定を下すことができる。チーム編成におけるデータ駆動型意思決定支援の既存研究のほとんどは完全自動化型のアプローチ [18–20] であるため、VA アプローチの実現の必要性は高い。そしてチーム編成を含むメンバー選択では集合データを扱う必要があるため、本研究で焦点を当てる集合データ VA の確立は重要な課題であると言える。

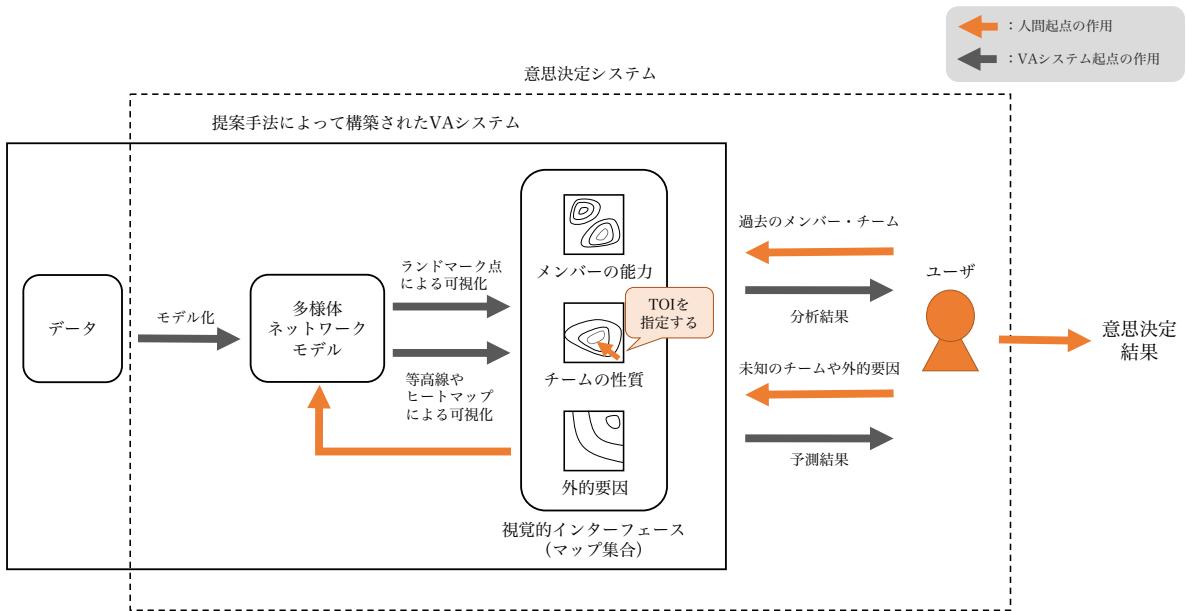


図 1.2 提案手法のキーコンセプト。地形図の集合の上に可視化される MNM としてデータがモデル化される。地形図の上では分析対象 (e.g. チーム, 構成メンバー, 外的要因) がランドマークのように可視化される。また成果のようなほかの情報は等高線のよう色付けによって可視化される。

1.2 本研究のゴールとキーアイディア

ここまで述べたように、集合データ VA の実現は、メンバー選択のような集合データを取り扱う必要のある問題を支援する観点から重要な課題である。しかしながら、集合データは機械学習の領域において取扱いの難しい対象であり [12, 13]、集合データの VA は挑戦的な課題である。

よって本研究のゴールは、特にメンバー選択支援システムに代表されるような実際の VA システムを構築できる、集合データに対する汎用的な手法を確立することである。より具体的には次の 5 つの課題^{*3}に取り組む。(1) VA システムが集合データをどのように扱うべきか (2) 構成要素、集合の性質、といった集合の出力を決定づける要因の間の関係をシステムがどのように可視化するべきか (3) システムはどのようにしてユーザ自身による探索的な分析と予測を可能とするのか (4) システムがどのようにして組合せ探索問題を

^{*3} これらの課題は本論文で定義する集合データ VA システムおよびその汎用構築手法が満たすべき 5 つの要件に対応している。本研究の実施にあたって VA および集合データを扱う機械学習の領域について広範な調査を行なったが、集合データを扱う汎用的な VA の方法論の研究は見つけることができなかった。従って、「集合データ VA とはどのようなものか?」を具体的に定義するため、また提案する手法がゴールである集合データ VA を構築する汎用的な手法となっているかを評価するために、5 つの要件を定義した。この要件については第 3 章にて説明する。

避けるのか (5) 個々の応用例に対して適応できるような可塑性をいかにして獲得するか.

本論文はこれらの課題を解決する VA システムの汎用的な構築手法として、生成的多様体ネットワークモデリングを提案する。提案手法のキーアイディアは、低次元な潜在変数から集合データとその出力が生成される過程をモデル化することである。このモデルを多様体ネットワークモデル (Manifold Network Model: MNM) と呼ぶ。集合データの生成過程を表現するために、メンバーに対する生成モデルとメンバー構成に対する生成モデルとを組み合わせる。また集合データの出力の生成過程を表現するために、出力に関連するそれぞれの要因についての生成モデルを導入し、それらを結合する。MNM を利用して、集合データ、その要素およびその他の関連する要因の関係性が地形図のような散布図 (マップ) の上で可視化される。これらのマップはユーザが TOI を特定し、関連する情報を可視化するための対話的な視覚的インターフェースとして利用される。このスキームを図 1.2 に示している。提案手法によって構築したシステムの最も重要な役割は、システムと人間のインタラクションの循環を通して、人間によるデータの探索、仮説検証、知識発見、チーム候補の検討を支援することである。

1.3 本研究の貢献と本論文の構成

本研究の貢献は次の通りである。

1. 集合データに対する VA システムおよびそれを構築するための方法論が満たすべき一般的な要件を定義した。また要件を満たす上で克服すべき困難点を明らかにした。
2. 多様体上の確率分布表現を用いた集合データのモデル化を実現した。これにより既存チームの分析と新規チームの生成・予測が可能となった。
3. さまざまなデータ構造に適応可能な多様体ネットワークモデルを提案した。
4. ユーザの TOI の指示により駆動されるインタラクティブな視覚的インターフェースを実現した。これにより複数要因が絡まったデータに対する対話的分析が可能になった。
5. バスケットボールチームのデータに対して提案手法を適用し VA システムのプロトタイプを構築しデモンストレーションを行った

貢献 1 により集合データ VA システムを構築する提案手法が解決すべき課題が明確化され、貢献 2, 3, 4 によってその課題を解決する提案手法が実現される。

以降、本論文は次のように構成される。第 2 章では本研究の背景と関連研究について三つの観点から述べる。第一は VA の観点であり、VA の枠組みや現在までの研究動向について詳しく解説する。第二は機械学習の観点であり、集合データを扱う機械学習の困難点

や既存のアプローチ、および最近の研究動向について解説する。とりわけ本研究と関わりの深い集合データの生成モデルについて詳しく述べる。第三はデータ駆動型のチーム編成支援の観点であり、既存の研究群を自動化型・情報提示型・VA型の三種に分類して概説する。第3章では集合データVAに求められる要件を定義する。集合データVAという概念自体が本研究独自のものであるため、その概念と要件を定義し、さらに実現上の技術的課題を明らかにする。具体的には集合データVAのシステムが満たすべき四つの要件と、システムの汎用的構築手法が満たすべき一つの要件を定義する。第4章では集合データVAシステムの構築法を提案する。まず多様体上の確率分布表現を用いた集合データのモデリング法を述べ、次に複雑なデータ構造に適応可能な多様体ネットワークモデリングへの拡張を述べる。さらに対話的な可視化を実現する視覚的インターフェースの構築法についても述べる。第5章は提案手法のデモンストレーションである。提案手法を用いてバスケットボールチームのVAシステムを構築し、過去のメンバー構成とゲーム成績のデータ分析や、新規ゲームにおけるメンバー選択支援などをデモンストレーションする。また比較手法となるVAシステムを構築し、提案システムが十分な能力を持つことを実証する。第6章は議論である。第三章で定義した要件の妥当性の検証や、提案手法のさらなる拡張についての検討を行う。また提案手法のデータモデリングの枠組みと、既存の機械学習のパラダイムとの関連についても述べている。第7章は総括として本論文をまとめる。

第2章

背景と関連研究

本章では本研究の背景や関連研究について3つの観点から述べる。1つ目の観点はVAである。VAの意思決定支援の枠組みについて解説した上で、先行研究や本論文におけるVAの定義を整理する。2つ目の観点は集合データを取り扱う機械学習である。集合データを取り扱う困難について解説した上で、近年のアプローチについてまとめる。また提案手法と関わりの深い集合データの生成モデルについても述べる。3つ目の観点はデータ駆動なチーム編成支援研究である。既存研究について自動化型・情報提示型・VA型の3つに分類して解説する。

2.1 VA

2.1.1 VAの枠組み

図1.1(b)においてVAが考える意思決定支援システムの枠組みについて述べたが、この詳細を示したものが図2.1である。本節ではこの図を通して、VAではどのような枠組みのもとで意思決定を支援しようとしているのか解説する。

いま入力をデータ、出力を意思決定結果とするような「意思決定システム」を考える。VAにおいて、このシステムはコンピュータシステム（以降VAシステムと呼ぶ）と人間という二者から構成されると考える。この二者が視覚的なインタラクションを繰り返すことで、人間が主導してデータ探索・仮説検証・知識発見を行い最終的な意思決定を行うのがVAにおける「意思決定システム」の仕組みである。この視覚的インタラクションにおいて、VAシステム側から人間への働きかけに相当するのがデータの分析結果の可視化であり、人間からVAシステム側への働きかけに相当するのが次に分析したい観点や関心の対象(TOI)の指定である。

ここからはVAシステムと人間が、それぞれどのような構造を持っているのかを詳しく

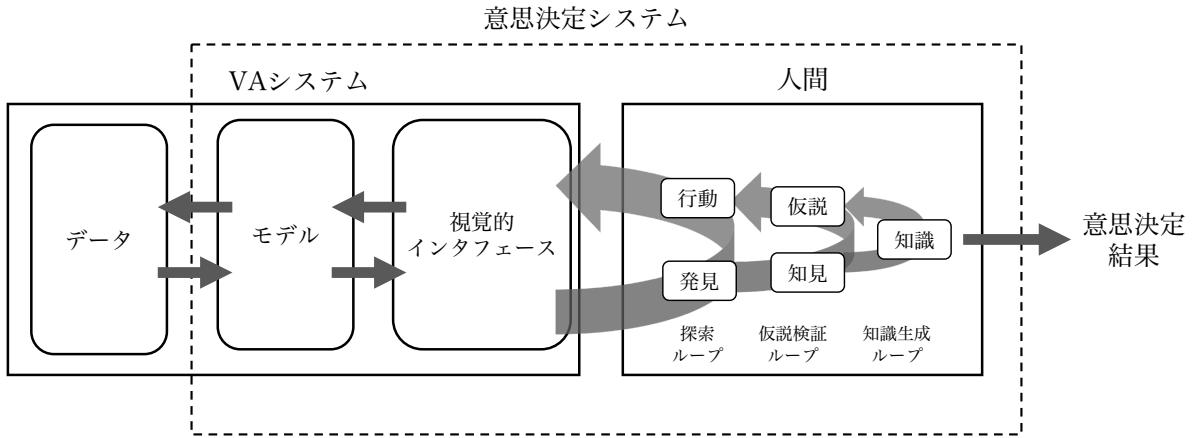


図 2.1 データから意思決定の結果を出力する「意思決定システム」を考えた際の VA の枠組み。意思決定システムは VA システムと人間から構成され、二者のインタラクションを繰り返すことでデータ分析・仮説検証・意思決定を行う。VA システムはデータおよびデータを抽出・分析するためのモデル、人間への情報の可視化や人間からの操作を受ける視覚的インターフェースから成る。人間は探索・仮説検証・知識生成という 3 種類のループが影響を及ぼし合いながら作用するような認知機構を持つと考えられている [21]。

見ていく。まず VA システムはデータを自動的に抽出・分析する「モデル」と、そのモデルから受け取った情報の可視化と人間の操作を受け取る「視覚的インターフェース」からなる。本論文で定義するモデルとは、視覚的インターフェース上でデータについての何かしらの情報を可視化するために、データを処理する機構のことである。この処理の例としては、人間から TOI として指定された具体的な特定の対象や特定の条件を満たす複数の対象を抽出するようなフィルタリングをはじめとして、複数の対象についての統計量を求めるような統計処理、機械学習手法の適用のような複雑な情報抽出が該当する。一方で視覚的インターフェースはモデルによって処理された情報を可視化する領域と人間が操作を行うためのボタンやチェックボックスのようなコンポーネントを備えた、いわゆるグラフィカル人間インターフェース (GUI) である。

一方で人間はこの視覚的インターフェースを通してデータについての情報を知覚し、それを元に TOI を VA システムに伝達するが、このインタラクションを生じさせる人間側の認知プロセスとして次の 3 種類のループ構造が並行して動作していると考えられている [21]。1 つ目は「発見 (Finding)」と「行動 (Action)」を生成する「探索 (Exploration)」のループである。「発見」は可視化結果から見つけた注目すべき情報のことである。「行動」は発見から促される次の分析のための具体的な行動を指していて、例えば可視化方法の切り替えやより詳しく知りたい対象の指定といったものである。2 つ目のループは「知見 (Insight)」や「仮説 (Hypothesis)」を形成する「検証 (Verification)」のループであ

る。ここでの「知見」とは、探索ループで生じた「発見」に対して、人間がドメイン知識を踏まえて解釈したり、推論したりすることで生じる情報の単位である [21]。可視化結果に含まれる情報そのものではなく、そこからさらに踏み込んだ情報であると言える。この知見は、分析の対象となる問題についての新たな仮定、つまり「仮説」を生み出すことも考えられる。この仮説を肯定・否定するための証拠を発見するために次なる行動を促す。つまり検証ループは探索ループを駆動させ、仮説の真偽を確認し新たな知見を生み出すものである。この検証のループを繰り返すことで知見に対する確信が深まり、また複数の知見の間の関係が体系化され、「知識（Knowledge）」として確立されていく。これが3つ目のループである「知識生成」のループである。このように3種類のループ構造が影響し合いながら並行して動作している。

この枠組みの元でのデータ分析過程の典型例を挙げると次のようになる。

0. システムが初期可視化（そのためのデータの前処理や分析を含む）を行う。
1. 人間は可視化結果を知覚し、注目すべき「発見」を見出した上で、自身の既存の知識を踏まえて「知見」を得る。
2. 人間は新しい仮説を作り、それを検証するためにさらなる情報を得ようと TOI を定め、インターフェース上での操作によって VA システムに対し TOI を伝える。
3. TOI を元にシステムは分析・可視化を行う

これらのステップを0から実行し1~3を繰り返す。その結果として「知識」が生成されると考えられる。ステップ0の初期可視化はVAシステムによって行われるが、その後の可視化やモデルの分析は基本的に人間から与えられるTOIが起点となって行われており、「人間主導」な分析過程となっている。

VAの枠組みの特徴は、VAシステム側のアーキテクチャだけでなく、人間側の認知過程をモデル化しているところである。これはVAの根底にある「人間主導」な意思決定を目指すという価値観に由来していると考えられる。

2.1.2 VA の定義

前節で紹介したVAの枠組みを踏まえて、本論文ではVAを「コンピュータと人間との間に生じる視覚的なインタラクションを通して、探索的なデータ分析・仮説検証・意思決定を人間自身が駆動させるための方法論や技術」と定義する。これは複雑で重大な問題に対してあくまで人間が意思決定を行うために、データ分析や意思決定の過程を人間が主導して進めていくというVAの「人間主導」な側面に着目した定義である。

ただしこれはあくまで本論文における定義であり、この他にもVAの形式的な定義は様々な研究者によって試みられている。複数の定義が存在する理由はVAが非常に抽象的

な概念・枠組みであり、一言で「VA とは何か？」を表現することが難しいためだと考えられる。

そこで本節では Cui の包括的なレビュー [22] を参考に、VA についての既存の定義について複数紹介する。これらの定義が VA のどの性質に「VA らしさ」を見出しているのかという点を紹介し、「VA とは何か」ということへの多面的な理解を試みる。また、既存の定義と本研究における定義を比較する。

最も古い VA の定義は、2005 年に Thomas ら [11] によって次のように与えられている。

Visual analytics is “the science of analytical reasoning facilitated by interactive visual interfaces”.

(Visual analytics とは「インタラクティブな視覚的インターフェースによって促進される分析的推論の科学」である。) (訳は引用者による)

Thomas は VA 分野の創設者であり、単純な情報の可視化方法にとどまらず、人間とのインタラクションとそれを通したデータ分析の過程に注目すべきであるという視点から VA という概念を提唱したと考えられる。この定義によって「VA システムと人間がインタラクションする」「そのための視覚的インターフェースを持つ」といういわば VA であるための最低限の条件が定められたと言える。

この Thomas らの定義を踏まえた上で、2008 年に Keim ら [23] は VA を次のように定義し、VA を具体化した。

Visual analytics “combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets”.

(Visual analytics は「非常に巨大かつ複雑なデータセットを元にした効果的な理解や推論、意思決定のために自動化された分析技術をインタラクティブな可視化と組み合わせたもの」である。) (訳は引用者による)

この「非常に巨大かつ複雑なデータセット」という対象において「効果的な理解や推論、意思決定」という目的をいかに達成するかということは VA 分野における極めて重要な関心事であり^{*1}、この Keim らの定義によってそれが明確化されたと言える。そもそも「非常に巨大かつ複雑なデータセット」から意思決定に有用な情報を引き出すことは難しいタスクであり [24]、一度きりの情報可視化ではなく情報を引き出す観点を逐次的に切り替え

^{*1} なお本論文では、この「非常に巨大かつ複雑なデータセット」のうち、「複雑さ」にどう対処するかということは主要な関心事の 1 つであるが、「巨大さ」をどう扱うかということについては言及しない。だがこの「巨大さ」への対処は VA 分野の重要な動機の 1 つであることを付け加えておく。巨大なデータに対する可視化の困難といった詳細については伊藤の著書 [24] が詳しい。

られるインタラクティブな可視化の必要性がこの定義によって強調されていると言える。

ここまで定義を踏まえた上で Cui は 2019 年に次のように定義している。

Visual analytics is a multidisciplinary research field mainly based on visualization, algorithmic data analysis and analytical reasoning, which takes advantage of visualization and interactions as suitable tools to integrate human judgment into the KDD process to visually discover explainable patterns (knowledge) and to gain insight into large and complex data sets.

(Visual Analytics とは人間の判断をデータベースからの知識発見 (Knowledge discovery in databases: KDD) 過程に統合して説明可能なパターン (知識) を視覚的に発見し大規模で複雑なデータセットへの洞察を得るための適切なツールとしての可視化とインタラクションを活用した主に可視化・アルゴリズムによるデータ分析・分析的推論に基づく学際的な研究分野である。) (訳は引用者による)

この定義で注目すべきなのは「人間の判断を統合する」という点である。Cui の定義は、Keim らの定義によって確立された VA の対象と目的を踏まえた上で、それらに対する「人間の判断を統合する」というアプローチをとるのが「VA らしさ」の一つであると定義しているのである。

そして本論文では VA を次のように定義する。

VA とは「コンピュータと人間との間に生じる視覚的なインタラクションを通して、探索的なデータ分析・仮説検証・意思決定を人間が主導して行うための方法論や技術」である。

Cui が示した「人間の判断を取り込む」という点からさらに踏み込んで、人間をシステムに取り込んでいるだけ（人間が能動的・受動的かは問わない）ではなく人間が能動的に分析過程を駆動しているという VA の「人間主導」的な側面を強調しているのが本定義である。

最後にこれらの定義はそれぞれ定義する観点が異なっていることも付け加えておく。Thomas らは科学としての VA を定義している。一方で Keim らはデータ分析のシステムという観点から定義している。また Cui は研究領域という観点から定義している。そして本論文はデータ分析の方法論・技術としての VA を定義している。

2.1.3 VA システムが提供する可視化とインタラクション

本節では、データ分析の過程における VA システムの機能について解説する。ここまで述べてきたように、VA の枠組みにおける VA システムは、人間が自らデータを探索する

ためにデータと人間とを橋渡しする役目を担っている。そのため次の2つの機能が求められる。1つ目は可視化であり、人間が知覚できる情報となるようにデータをモデルが分析をした上で可視化を行うことである。2つ目はインタラクションであり、人間のTOIをフィードバックとして受け取り、それをモデルの分析に反映させることである。それぞれの機能について、Cuiの分類[22]を元に詳しく述べる。

可視化

VAシステムの可視化について、「どのようなデータから」「どのように分析して」「どんな視覚表現を」生成しているかという観点から次の4つに分類できる。1つ目は2D-to-2D、2つ目はMulti-Dimensional-Reduction-2D、3つ目はMulti-Dimensional-transformation-2D、4つ目はMulti-Dimensional-to-3Dである。

1つ目の2D-to-2Dとは、2次元のデータから2次元座標系での視覚表現を生成する可視化であり（以降、2次元可視化と呼ぶ）、最も基本的な可視化のアプローチと言える。具体例としては、オブジェクトが2次元空間を移動しているような場面においてオブジェクトの各時刻の2次元座標が得られているような動きを表現するデータに対して、地図上での動きの軌跡を可視化するVAシステム[25]がある。

一方で2つ目のMulti-dimensional-to-2Dとは、多次元データを対象とした2次元可視化である。2D-to-2Dの場合は、元々のデータが2次元座標系に落とし込みやすいため、データをそのまま2次元可視化すること可能である。しかし、ここでの多次元データとは3つ以上の次元や属性を持つデータであり、そのままでは2次元の領域に可視化できない。そこでデータの本質的な情報を保ったまま、2次元にまで削減することを考える。このタスクは機械学習においては次元削減と呼ばれる。次元削減を行えば、2つの次元のそれぞれをグラフの軸にとって、散布図としてデータセットの全体像を可視化することが可能となる^{*2}。この散布図は類似したデータ同士が近くに配置されるような「近さ」が定義された地図のような二次元空間であるため、本論文ではしばしば「マップ」と呼ぶことにする。次元削減手法としては主成分分析(Principal Component Analysis: PCA)[29,30]、多次元尺度構成法(Multidimensional scaling: MDS)[31]、t分布型確立的近傍埋め込み(t-distributed Stochastic Neighbor Embedding: t-SNE)[32]が用いられることが一般的である。Cuiによって示されたVAの原則[22]に”Analyze/Overview first”とあるように、VAにおいてはまず与えられているデータセットについて概観することが重要であり、次元削減法はそれに適したアプローチであると言える。

^{*2} なお低次元表現を獲得しVAシステム上で可視化する対象は高次元データセットだけではない。例としてグラフ[26]、離散的な記号である単語[27]、その系列である文書[28]が挙げられる。なお、この場合に低次元表現を獲得するタスクは次元削減ではなく埋め込みと呼ばれることが多い。本研究では集合データに対してこのような低次元な表現を抽出することで可視化を行う。

3つ目の Multi-Dimensional-transformation-2D は、次元削減のような複雑な分析処理を必要としない方法で多変量データを 2 次元可視化するものである。この際は平行座標法 (Parallel Coordinate Plots: PCP) やヒートマップといった可視化手法が用いられる。PCP は多変量データの各次元を表す垂直な線分を左右方向に並べ、特定のデータの各次元の値を折れ線で結ぶ可視化法である [24]。次元削減の可視化の場合は元のデータの各次元の情報は埋もれてしまうが、こちらの描画では各次元の相関を視覚的に把握できるという利点がある。だが複数の折れ線が絡み合うためデータが多い場合は視認性に欠ける場合もあり、また各次元の線分同士の配置によっては本当に必要な相関を見逃してしまうことも起こりうる。

4つ目の Multi-Dimensional-to-3D は多次元データから 3 次元座標系上での視覚表現 (以降、3 次元可視化) を生成するものである。実際に Kurzhals らは眼球運動のデータに対して人間が自由に視点を制御できる 3 次元可視化インターフェースを構築している [33]。2 次元可視化と 3 次元可視化の優劣や利点・欠点にはさまざまな議論があり [24]、実現したいデータ分析過程や想定する利用者の分析リテラシーに応じて使い分けが必要だと考えられる。

インタラクション

インタラクションについては、何を指向しているかという点に注目した上で次の 2 つの分類がある。1つ目は探索指向であり、2つ目は表現指向である。1つ目の探索指向はデータや可視化の座標系を探索するためのインタラクションを指している。例えば異なるデータの抽出・変換方法の指定や表示領域のズームが挙げられる。一方で2つ目の表現指向は、主にモデルが次元削減のような複雑な分析処理を行う際に、人間が自身の知見をモデルにフィードバックし、分析に影響を与えるためのインタラクションを持つようなシステムである。例えば interactive PCA [34] では、インターフェース上有る複数のスライダを調整することで、PCA の次元削減結果を変更することができる。この他でも次元削減をはじめとする機械学習手法に対して、人間の知見をフィードバックし、機械学習モデルの学習を向上させようとする試みは複数行われており、詳しくは Endert らのサーベイ論文 [35] にて言及されている。

2.1.4 Predictive VA (PVA)

VA システムが提供する情報は、過去のデータに対する分析結果だけにとどまらず、未知のケースに対する予測結果も含まれる。予測分析を主目的に据えた VA を Predictive VA (PVA) [36] と呼ぶ。

PVA では予測モデルの解釈性の向上が主要な目的の 1 つとなる。予測分析はデータに

存在しない未知のケースに対する予測結果を示すため、より直接的に意思決定を支援することができるデータ分析のアプローチである。近年では機械学習に代表されるデータ駆動な予測モデルが急速に発達しあらゆるドメインで高い予測精度を達成しているが、これらの手法のブラックボックス性が問題視されており、実際に意思決定に予測結果を利用していくためにはモデルの解釈性の向上が重要であることが示唆されている [37]。より具体的にはモデルは意思決定の結果として「何が」起きるかを予測するだけではなく、それが「なぜ」起きると予測したのかについての知識を提供することが重要である。これにより人間は自身の経験的なドメイン知識と予測結果とを照らし合わせたり統合したりした上で、最終的な意思決定を行うことができる。このモデルの理解を VA によって支援するのが PVA であり、予測分析のモデリングや結果の探索、モデル選択といった各プロセスにおいて VA の技術を適用することで、モデルそのものやモデルの予測結果について人間が主体的に探し理解を深めることを支援する。

本論文の提案手法によって構築される VA システムには PVA の側面がある。まずこの VA システムでは、任意の集合データに対して、その出力を予測する機能を提供している。この所望の集合データの指定や、それに対する予測結果の確認は視覚的インターフェース上でインタラクティブに行うことが可能であり、人間はモデルによる予測結果を探索することでモデルに対する理解を深めることができる。また、このインターフェース上には過去のデータについての分析結果も同時に表示されるため、人間が所望する未知ケースと類似した過去のケースを発見し、そこからモデルの予測根拠について検討することができる。

このような機能を実現できる理由は、提案手法は予測分析を実現するアプローチとして潜在変数を仮定した生成過程のモデル化を行なっているためである。予測分析の最もシンプルなアプローチはデータとして得られている変数 x と y のペアに対して、 x が決まると y が定まるという関係を仮定した上で、その対応関係を表現する関数 $y = f(x)$ を推定することである。多くの PVA ではこのモデル化を前提としている。これに対して、提案手法の場合は x と y のペアは共通する低次元な潜在変数 z から生成されると仮定する。 z を 2 次元程度の可視化できる次元に設定することで、2 次元マップの上で予測分析を行うことが可能となる。また過去のケースとして得られている既存の x と y のペアについて z を推論しマップの上で可視化することで、予測分析のインターフェース上に過去の分析結果も同時に示すことができる。加えて生成的アプローチでは x から y への予測だけではなく、 y から x の予測も可能となるため、2 つの変数の間の双方向な探索が可能となる。つまり潜在変数を仮定した生成的なデータのモデル化は、インターフェース上のモデルと人間とのインタラクションを通して未知のケースの予測と過去のケースの分析をシームレスに実行でき、なおかつ変数間の因果の方向に依らない探索を実現することが可能であるため、PVA の目的の達成に有効なアプローチであると考えられる。

2.2 集合データを扱う機械学習

2.2.1 集合データの定義

集合データを扱う機械学習について詳しく述べる前に、まずは本論文で扱う集合データとは何か定義する。集合データとは特徴量の集合である。即ち、 D 次元の数ベクトル $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ が集まつた集合 $X = \{\mathbf{x}_j\}_{j=1}^n$ である。この集合データという単位が分析対象となる 1 サンプルであり、データセット全体としては集合データの集合 $\{X_i\}_{i=1}^N$ となる。

2.2.2 集合データを扱う困難

機械学習や数理最適化の問題において集合データを扱う際の困難として、代表的なものを 3 つ紹介する。1 つ目は、集合データに対して定義できる距離や類似度が非自明なことである。これは集合データが持つ次の 2 つの性質 [13] に由来するものである。

- 可変長であること。2 つの集合データ $X_i, X_{i'}$ はそれぞれ異なる要素数 $n_i, n_{i'}$ を持つ。
- 交換可能であること。要素の順序が変わっても同一の集合データと見なす^{*3}。

機械学習で扱う典型的なデータである数ベクトルは固定長かつ各次元が交換不可能であり、ユークリッド距離のような自明な距離を導入できる。しかしながら可変長・交換可能な集合データに定義できる類似度や距離は非自明である。機械学習においてデータとデータの間が距離や類似度が定義されていることは本質的に不可欠であり^{*4}、この 2 つの性質は機械学習で集合データを扱うための大きな障害となる。そのため、集合データをどう扱うか、より厳密に言えばどのようにして集合データをその内在的な性質を保ったまま距離や類似度を定義できる「表現」に変換するか、という問題が機械学習において大きな研究課題の 1 つとなっている [12, 13]。素朴な表現抽出としては特徴量の総和や平均を取る操作が考えられるが、タスクの精度の向上のために必要な集合データと集合データの間の差異が埋もれてしまう可能性がある。

2 つ目の困難は、場合によっては集合データ $X = \{\mathbf{x}_j\}_{j=1}^n$ が観測された過程において、

^{*3} この性質に由来して、一部の文献 [38, 39] では集合データのことを交換可能データ (Exchangeable data) と呼ぶ。

^{*4} 例えば機械学習の典型的な問題設定である有限の観測データ集合からの関数 f の学習の場合、この関数の入力に距離や類似度が導入できなければ（関数の定義域が連続性を持たない位相空間であれば）、観測データにない入力に対して出力を求めること（いわゆる汎化すること）はできない。

各要素が独立同分布 (i.i.d.) に従うと見做せないという点である。例えば集合データをあるチームのメンバーの特徴量集合だとすると、成果を最大化するために選出されたラインナップであればメンバー間の相互的な相性の良さを考慮して選択されているため、このような場合に i.i.d. を仮定することは非現実的である。この点は主に、次節で述べる集合データの生成タスクにおける困難となる。詳しくは次節で述べる。

3つ目の困難は、集合データ X についての関数 $y = f(X)$ についての最適化は離散最適化問題になるという点である。任意の X から y を推定する（順問題）を行う関数 f は、前述の通り類似度や距離が定義できる表現を集合データから抽出することで構築できる。しかしながらその逆問題、つまり任意の y に対する X を求める問題は、本質的に組合せの探索が必要となる。当然ながら、集合データ X の要素数に対して指数的に探索回数は増加する。

2.2.3 代表的なタスク

集合データを扱う問題としてはメンバー選択以外にも三次元点群のモデリング [40–42] やマルチタスク学習 [12, 43] などがある。また本研究で定義する集合データには含まれないが、要素が離散的なラベルを取る例としてはドキュメント解析 [44] がある。

集合データを扱う必要のある問題は様々だが、これらを定式化すると主要なタスクとして識別タスクと生成タスクの2種類が挙げられる [45]。

識別タスクは集合データ X とそれに付随する出力 y のペアがデータセット $\{(X_i, y_i)\}$ として与えられているとき、その関係を $y = f(X)$ と仮定した上で関数 f を推定することが目的となる。 y が連續値を取る場合は回帰問題、離散値を取る場合は分類問題となる。この f は入力の集合の要素数が変わることを許容し、なおかつ要素の順序に対する不变性 (permutation-invariant と呼ばれる性質) を満たさなければならない。また、 f はデータセットに含まれない未観測の X に対して出力を予測できなければならないが、前述の通り集合に対しての距離は自明ではないため、 f を獲得する確立された手法というのは存在しない。近年よく用いられるアプローチは集合を扱うことができるニューラルネットワークのモジュールを用いて（詳しくは次節で述べる）、識別の精度が高くなるような集合の表現の抽出を自動的に行うやり方である。

生成タスクでは、与えられた集合データの集合 $\{X_i\}$ から、集合データを生成したであろう確率分布 $p(X) = p(\mathbf{x}_1, \dots, \mathbf{x}_n)$ を推定することが目的となる^{*5}。本論文ではこの $p(X)$ を集合の生成モデルと呼ぶ。ここで注意すべきなのは 2.2.2 節で第 2 の困難として述べた非 i.i.d. という仮定を導入することである。つまり、 $p(X) = \prod_j p(\mathbf{x}_j)$ と仮定せず、

^{*5} なお、任意の X が与えられた時の確率密度まで計算できるような明示的な $p(X)$ を獲得することことは要求せず、 $p(X)$ からサンプルが生成できるような非明示的な表現でも可とする。

全ての要素の同時分布としての $p(X)$ を推定する必要がある。その上で $p(X)$ は要素の順序に対する不变性を満たす必要がある。つまり全部で $n!$ 通りあるいかなる順列 per に対しても $p(\mathbf{x}_1, \dots, \mathbf{x}_n) = p(\mathbf{x}_{\text{per}(1)}, \dots, \mathbf{x}_{\text{per}(n)})$ である必要がある。このような $p(X)$ の構築方法は自明ではない。交換可能性は満たさず可変長性のみを満たす状況は自然言語処理の領域で頻出し、このときよく用いられるアプローチとしては自己回帰型のニューラルネットワークで $p(X) = \prod_{j=1}^n p(\mathbf{x}_j | \mathbf{x}_1, \dots, \mathbf{x}_{j-1})$ とモデル化することである [46, 47]。しかしこれは交換可能であることを無視したモデル化であり、このモデルに対して集合データの順序をシャッフルしてデータ拡張を行い学習を行なっても、精度向上は見込めないことが知られている [48]。この生成タスクを解決するために様々な手法が提案されている。それらの詳細は 2.2.5 で述べる。

2.2.4 代表的なアプローチ

識別タスク・生成タスクを問わず、機械学習における集合データを取り扱う方法は次の 2 つのアプローチに分けることができる。

1 つ目のアプローチはその集合データの各要素を生成した確率分布 $p(\mathbf{x})$ を考え、その確率分布の性質を表す統計量を集合データの表現として利用するというものである。素朴な特徴量としては平均や分散が挙げられるが、この場合より高次なモーメントの情報が欠落してしまう。この対策としてカーネル平均を用いた方法 [49] や、ノンパラメトリックに確率分布を推定した上でその確率分布に対するカーネルを定義する方法 [50] が提案されている。

2 つ目は解きたいタスクに有効な特徴量を抽出するような機構を学習により獲得するアプローチである。集合に対応したニューラルネットワークのモジュールが提案されており、代表例として素朴なプーリング層をベースとした Deep sets [12]、注意機構をベースとした Set transformer [13]、最適輸送に基づいた Optimal Transport Kernel Embedding(OTKE) [51]、常微分方程式を元にした Neural Exchangeable Neural Ordinary Differential Equation(ExNODE) [52] などがある。これらのモジュールをネットワークに組み込めば、目的関数を最適化するように end-to-end に学習を行うことができる。学習がうまくいけば、タスクに適した表現を抽出する機構が自動的に獲得されることになる。

後者は予測や識別のように、明確な評価指標があり、それを最大化することが最優先となるようなタスクを解くような場合に適したアプローチである。一方で何らかの知識発見をしたいというように評価指標が明確でない場合、二つのアプローチは一概に優劣をつけられるものではない。

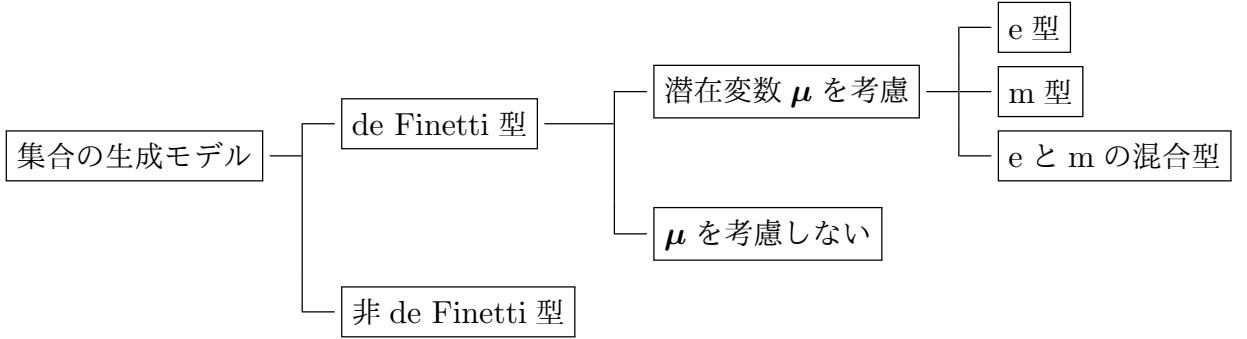


図 2.2 集合データの生成モデル $p(X)$ の表現方法の分類. まず大きく de Finetti の表現定理の等式 (2.1) を仮定するか否かによって分かれる. 前者は個々の特徴量 \mathbf{x} に対する潜在変数 μ を明示的に導入するか否かでさらに分かれる. また, 明示的に μ を導入する場合は m 型と e 型とその混合の 3 種類に分類できる.

2.2.5 集合データの生成モデル

提案手法では集合データの生成タスクを解くことになる. そこで, 本節では生成タスクを解く既存手法について概説する. 生成タスクの解く方法は大きく (1) 生成モデル $p(X)$ に対してどのような仮定を置くか, (2) 仮定した生成モデルのパラメータをデータからどのように推定するか, という 2 つの要素から成る. 本節においては主に (1) に着目し, 様々な手法における集合データの生成モデル $p(X)$ の表現方法について図 2.2 の分類に沿って説明する.

なお, 生成タスクの具体例として三次元点群のモデリングがあり, この場合に特化した手法が数多く提案されているが [53–59], 本節では一般化された特徴量 $\mathbf{x} \in \mathbb{R}^D$ に対する集合データを扱う手法について述べる.

de Finetti の表現定理を仮定するモデル化

困難な $p(X)$ のモデル化においてヒントを与えるのが, 次の de Finetti の表現定理 [60] である. 任意の確率分布に従う無限個の交換可能な集合 $\mathbf{x}_1, \mathbf{x}_2, \dots$ に対して, その中から取り出された有限個の部分交換可能集合 $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ の同時分布 $p(X)$ に対して, 潜在変数 $\tau \in \mathcal{L}^{(t)}$ を導入した

$$p(X) = \int \prod_{j=1}^n p(\mathbf{x}_j | \tau) p(\tau) d\tau \quad (2.1)$$

が成り立つような $p(\mathbf{x} | \tau)$ と $p(\tau)$ が存在する。つまりこの定理に基づけば、集合データ $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ の生成過程を次のように考えることができる。まず事前分布 $p(\tau)$ から潜在変数 τ がサンプルされる。次にその潜在変数による条件付き分布 $p(\mathbf{x} | \tau)$ に従い X の各要素が独立に生成される。その結果、集合データ X が得られる。

ただし、この de Finetti の定理が常に成り立つのはあくまで背後に無限個の交換可能な集合がある前提であり、有限個の集合に関しては必ずしも成り立つわけではない⁶。しかしながら、推定したい集合データの生成モデル $p(X)$ に対して (2.1) が成り立つことを仮定することで、潜在変数の事前分布 $p(\tau)$ と潜在変数 τ による条件付き分布 $p(\mathbf{x} | \tau)$ さえ決まれば、その 2 つによって $p(X)$ が一意に定義される。このように集合データの生成モデルの扱いが容易になることから、式 (2.1) が成り立つという仮定はしばしば導入される。

また τ は集合データ X の各要素に共通した潜在変数であるため、集合 X の性質を表現していると言える。この τ を連続潜在変数、つまり固定長の数ベクトルである仮定し、 $\{X_i\}$ に対する $\{\tau_i\}$ を推定すれば自明な類似度や距離を定義できる集合データの表現を得ることができる。また、本論文で目指す集合データに対する VA の観点では、この数ベクトルを可視化できる程度に低次元であると仮定することで、集合データの集合の全体像を概観できるようなマップを獲得することができる。

従って式 (2.1) を仮定した場合、 $p(X)$ をどう表現するかという問題は、 $p(\mathbf{x} | \tau)$ をどう表現するかという問題に置き換わる⁷。まず素朴な $p(\mathbf{x} | \tau)$ の表現方法として考えられるのは、ガウス分布のようなパラメトリックな確率密度関数、もしくはカーネル密度推定 [65] のようなノンパラメトリックな確率密度関数を仮定するやり方である。しかしながら、前者の場合は表現力に乏しいこと、後者の場合は \mathbf{x} が高次元になると必要なデータ数が莫大になってしまうという欠点がある。

そこでまた別のアプローチとして候補に挙がるのが個々の特徴量 \mathbf{x} それぞれに割り当てる潜在変数 $\mu \in \mathcal{L}^{(m)}$ を導入するやり方である。これは直感的に言えば、ある τ から生成された $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ の要素間の差異を表現する潜在変数であり、要素間で共通する τ とは対照的である。具体的な $p(\mathbf{x} | \tau)$ への μ の導入方法として、次の 2 種類が

⁶ 交換可能な有限個の集合に対して de Finetti の表現定理を拡張しようという試みは数多くある [61–64]。中でも [64] では確率測度 ((2.1) では $p(\tau)$ に相当) が負を取りうることを許容すれば de Finetti の表現定理が成り立つことを示している。

⁷ 厳密には $p(\tau)$ をどう表現するかという問題も含まれるが、 $p(\tau)$ には一様分布やガウス分布など適当な扱いやすい分布を定義することが多い。

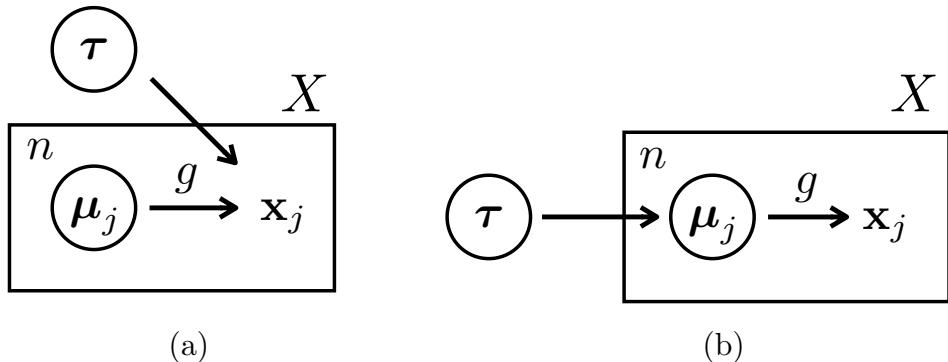


図 2.3 集合データ $X = \{\mathbf{x}_j\}_{j=1}^n$ の生成モデル 2 種. 円は潜在変数を表し, 矢印の元の変数が矢印の先の変数の生成に影響を与える. パネルは左上の回数だけ生成を繰り返すことを意味する. どちらも X に対応する潜在変数 τ を仮定する. (a) e 型. 集合内に共通する潜在変数 τ と集合内でも要素ごとに異なる μ_j がそれぞれ独立に生成され, 両者から特定の要素 \mathbf{x}_j が生成される. (b) m 型. まず τ から μ_j が生成され, そこから \mathbf{x}_j が生成される.

考えられる。

$$p(\mathbf{x} \mid \boldsymbol{\tau}) = \int p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\tau}) p(\boldsymbol{\mu}) d\boldsymbol{\mu} \quad (2.2)$$

$$p(\mathbf{x} \mid \boldsymbol{\tau}) = \int p(\mathbf{x} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid \boldsymbol{\tau}) d\boldsymbol{\mu} \quad (2.3)$$

この 2 種類の $p(\mathbf{x} | \boldsymbol{\tau})$ を元に集合データ X の各要素 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ の生成過程をベイズのグラフィカルモデルのように描いたのが図 2.3 になる。Ishibashi らは情報幾何学の観点から前者のモデルを e 型（図 2.3(a)），後者のモデルを m 型（図 2.3(b)）と呼んでいる [66]。e 型と m 型では $\boldsymbol{\tau}$ と $\boldsymbol{\mu}$ と \mathbf{x} の間の関係が異なる。e 型では $p(\boldsymbol{\mu}, \boldsymbol{\tau}) = p(\boldsymbol{\mu})p(\boldsymbol{\tau})$ と $\boldsymbol{\mu}$ と $\boldsymbol{\tau}$ が独立であるとした上で， $p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\tau})$ のように特徴量 \mathbf{x} が集合内の各要素に共通する潜在変数 $\boldsymbol{\tau}$ と共にしない潜在変数 $\boldsymbol{\mu}$ の 2 つから生成されると仮定している。一方で m 型では $\boldsymbol{\mu}$ と $\boldsymbol{\tau}$ が独立ではなく， $p(\boldsymbol{\mu}, \boldsymbol{\tau}) = p(\boldsymbol{\mu} | \boldsymbol{\tau})p(\boldsymbol{\tau})$ であることを仮定する。つまり $\boldsymbol{\tau}$ が定まった上で，チームを構成するメンバーの潜在変数集合 $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n\}$ が生成される。そして $p(\mathbf{x} | \boldsymbol{\mu})$ とあるように，個々のメンバー潜在変数から $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ が生成される。e 型と m 型は仮定する依存関係が全く異なるため，モデル化の目的やデータへの事前知識からどちらを用いるか選ぶ必要がある。

e型の場合は式(2.2)より $p(\mathbf{x} | \boldsymbol{\tau})$ を表現するためには $p(\mathbf{x} | \boldsymbol{\tau}, \boldsymbol{\mu})$ を表現することが目的となる^{*8}。2つの潜在空間から特徴量の空間への写像 $g : \mathcal{L}^{(t)} \times \mathcal{L}^{(m)} \rightarrow \mathcal{X}$ を導入して

$$p(\mathbf{x} \mid \boldsymbol{\tau}, \boldsymbol{\mu}) = \mathcal{N}(q(\boldsymbol{\tau}, \boldsymbol{\mu}), \beta^{-1} \mathbf{I}) \quad (2.4)$$

*8 潜在変数 μ の事前分布 $p(\mu)$ も定義する必要があるが、これには一様分布やガウス分布など扱いやすい適当な分布を仮定することが多い

と定義する。ただし β^{-1} は等方ガウス分布の分散パラメータである。このとき、 $\{X_i\}_{i=1}^N, X_i = \{\mathbf{x}_{ij}\}_{j=1}^{n_i}$ から $\{\boldsymbol{\tau}_i\}, \{\boldsymbol{\mu}_{ij}\}, g$ を推定することになる。この問題設定は content/style disentangled representation learning^{*9}とも呼ばれている [67]。

$\boldsymbol{\mu}$ を導入している大多数の既存手法はこれらの e 型に分類される。代表的なものとしては主成分分析 (Principal Component Analysis: PCA) を拡張した Multi-Level Simultaneous Component Analysis (MLSCA) [68] や Variational AutoEncoder (VAE) [69] を拡張した Multi-Level VAE (MLVAE) [67] や Group-based VAE(GVAE) [70], Nemeth らの手法 [71] がある。また $\boldsymbol{\tau}$ が離散潜在変数となっている手法には Set Distribution Networks [72] がある。

一方で m 型では $p(\mathbf{x} | \boldsymbol{\mu})$ と $p(\boldsymbol{\mu} | \boldsymbol{\tau})$ を構築する必要がある。前者の $p(\mathbf{x} | \boldsymbol{\mu})$ は

$$p(\mathbf{x} | \boldsymbol{\mu}) = \mathcal{N}(g(\boldsymbol{\mu}), \beta^{-1}\mathbf{I}) \quad (2.5)$$

のように、写像 $g : \mathcal{L}^{(m)} \rightarrow \mathcal{X}$ を用いて表現する。 β は等方ガウス分布の精度パラメータである。後者の $p(\boldsymbol{\mu} | \boldsymbol{\tau})$ の表現方法は各手法によって異なる。VAE をベースとした Neural Statistician(NS) [43] の basic model では、 $p(\boldsymbol{\mu} | \boldsymbol{\tau})$ を次のようなガウス分布であると仮定する。

$$p(\boldsymbol{\mu} | \boldsymbol{\tau}) = \mathcal{N}(\boldsymbol{\mu} | \bar{\boldsymbol{\mu}}(\boldsymbol{\tau}), \boldsymbol{\Sigma}(\boldsymbol{\tau})) \quad (2.6)$$

ただし共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ は対角行列であることを仮定する。平均 $\bar{\boldsymbol{\mu}}(\boldsymbol{\tau})$ と $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ は $\boldsymbol{\tau}$ の関数となっており、この関数はニューラルネットワークによって表現する。これに対して、Hierarchical Tensor SOM Network (HTSOMN) [73] では $p(\boldsymbol{\mu} | \boldsymbol{\tau})$ をヒストグラムとしてノンパラメトリックに表現する。具体的には潜在空間 $\mathcal{L}^{(m)}$ に O 個の代表点集合 $\tilde{M} = (\tilde{\boldsymbol{\mu}}_o)_{o=1}^O, \tilde{\boldsymbol{\mu}}_o \in \mathcal{L}^{(m)}$ を配置し $\boldsymbol{\mu} \in \tilde{M}$ とする。すると $p(\boldsymbol{\mu} | \boldsymbol{\tau})$ は O 次元のパラメータ \mathbf{o} を持つカategorical 分布 $Cat(x|\mathbf{o}), x \in \{1, \dots, O\}$ として表現できる。

また一部には e 型と m 型の混合のような表現をする手法も存在する。NS の full model [43] では、前述した basic model では表現力が乏しいとして、 $\boldsymbol{\tau}$ が $\boldsymbol{\mu}$ の生成だけではなく \mathbf{x} にも影響する e 型と m 型を混合させたようなモデルを提案している。

提案手法では、de Finetti の表現定理を仮定することで、集合データの潜在変数 $\boldsymbol{\tau}$ や各要素の潜在変数 $\boldsymbol{\mu}$ を導入する。これらの潜在変数は集合データや集合の要素についてのマップを作成するために利用される。またこれらのマップを元に VA システムの視覚的インターフェースが構築される。

^{*9} データに対して潜在変数を推定する際に、同じ集合に属している要素間で共通する次元 (content) と異なる次元 (style) に分解 (disentanglement) された潜在変数を推定しようという考え方による。

非 de Finetti 型のモデル化

提案手法のアプローチとは異なるが、ここで de Finetti の表現定理 (2.1) の形を取らないモデルについても簡単に触れておく。前述の通り、(2.1) による定式化は表現しうる生成モデルを制限するため、異なる形でモデルを表現することで、表現し得るモデルの集合を拡張しようという試みがなされている。

FlowScan [74] では交換可能でない系列に対しては自己回帰型モデル [46, 47] でうまく生成モデルが構築されていることに着目し、集合データの各要素に対して順序づけを行い、それに対して自己回帰型モデルを適用することで生成モデルを実現している。

Bayesian Recurrent Neural mOdel (BRUNO) [38] では、 X の同時分布 $p(X)$ のモデル化ではなく、 $\{\mathbf{x}_1, \dots, \mathbf{x}_{n-1}\}$ が与えられているときの \mathbf{x}_n の予測分布 $p(\mathbf{x}_n | \mathbf{x}_1, \dots, \mathbf{x}_{n-1})$ のモデル化を行なっている。このモデリングにおいても de Finetti の表現定理 (2.1) を

$$p(\mathbf{x}_n | \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) = \int p(x_n | \boldsymbol{\tau}) p(\boldsymbol{\tau} | \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) d\boldsymbol{\tau} \quad (2.7)$$

と導入し $p(\boldsymbol{\tau} | \mathbf{x}_1, \dots, \mathbf{x}_{n-1})$ と $p(\mathbf{x} | \boldsymbol{\tau})$ によってモデル化できる。ただしこの場合 $\boldsymbol{\tau}$ の積分の評価が難しく、そのために近似を導入すれば交換可能性が損なわれる可能性がある。BRUNO では Normalizing Flow の一種である Real NVP [75] と t 過程を組み合わせることで $p(\mathbf{x}_N | \mathbf{x}_1, \dots, \mathbf{x}_{N-1})$ を直接定義しこの問題を回避している。

BRUNO や FlowScan はモデルが (2.1) の形をしていないというだけで、(2.1) が成り立つような $p(X)$ であることは暗黙のうちに仮定している [76]。一方で Yang らの手法 [39] では BRUNO と同じように確率過程に着目しつつ、de Finetti を仮定しないモデル化に成功している。これは確率過程が潜在変数 $M = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n\}$ による条件付きの交換可能な分布 $p(X | M)$ を自然に表現することに着目し、まず $p(X)$ を

$$p(X) = \int p(X | M) p(M) dM \quad (2.8)$$

$$= \int p(X | M, \boldsymbol{\tau}) p(\boldsymbol{\tau}) p(M) d\boldsymbol{\tau} dM \quad (2.9)$$

と因子分解せず表現する。その上で $p(X | M, \boldsymbol{\tau})$ を表現する確率過程と、 M と $\boldsymbol{\tau}$ についての周辺化テクニックを提案している。また (2.1) の仮定を避けて他の手法として Set Flow [76] がある。

2.3 チームマネジメントにおけるメンバー選択支援

ビジネス [20], スポーツ [77], 医療 [78], 軍事 [79], 危機管理 [80], 宇宙開発 [81], アカデミア [82] などあらゆる分野において組織はチームという単位で成果の最大化に迫られる。そのためチームマネジメントの領域では、いかに成果を最大化するようなメンバーを選択するかというチーム編成の問題が重要視されている [83–85]。だが、チームの成果を最大化するには個々のメンバーの能力だけを考慮するだけでは不十分であり [86, 87]、個々人が出す成果の総量をチームが出す成果が上回るシナジー効果 [15–17] を生み出すようなメンバーの組合せを考慮する必要があるため、チーム編成は非常に困難であると言える。

このチーム編成を情報技術によって支援する試みが数多く行われている。基本的にこれらの研究はドメインごとの特定の問題に特化して支援するものがほとんどであり、細分化が進んでいる領域であるが、大きく自動化型・情報提示型・VA によるデータ探索型の 3 つに分類できる。

自動化型ではどのメンバーを選択するかという意思決定をコンピュータによって代理させることを目的としている。具体的には、メンバーの組合せの最適化 [19, 20, 88–92] や起用すべきメンバーの推薦 [93, 94] が行われている。前者では成果 y をラインナップ X の関数 $y = f(X)$ によって予測し、この f を用いて組合せ最適化問題を解くものである。関数 f が実際に解きたい問題に対して適切にモデリングできていることが重要となる。ほとんどの場合は専門家の知見を元に著者らが設計を行なっている。Liu ら [92] は機械学習を用いてデータ駆動的に f を獲得することを試みている。後者のメンバー推薦ではチーム全体ではなく個人またはごく一部の部分集合に焦点を当てている。例えばチームスポーツにおいて特定のポジションに向いたメンバーをデータから発見し推薦する [93]、ディープニューラルネットワークで二人のプレイヤーの成果の予測器を構築し特定のプレイヤーに相性の良い別のプレイヤーを推薦する [94] といったことが行われている。

自動化型とは対照的に、チーム編成の意思決定に有用な情報をデータから抽出しようという試みが情報提示型である。具体的にはチームスポーツにおいてメンバーに対し貢献度の算出やランク付けを行っている研究や [86, 95, 96]、規模の大きい研究者組織における研究者のネットワークのクラスタ構造を解析・可視化し、次に共同研究を行う研究者を発見しやすくする試み [97] がある。

自動化型と情報提示型は目指す支援の形が異なるため、どのようなチーム編成の場面かによって使い分けるのが適切であろう。自動化型はクラウドソーシングにおけるチーム編成 [90] やソーシャルゲームのプレイヤーのチームの編成 [94, 98] に利用されており、短時間に大量のチーム編成を行わなければいけない状況や、達成したい成果が明確で意思決定

を効率化したいという場合に向いていると言える。別の言い方をすれば、基本的にブラックボックスシステムとして最適な解を提供するだけにとどまっているため、重要な課題を取り組むようなチームの編成にそのまま適用するのは難しい場合がある。一方で情報提示型は、最終的な意思決定は人間に委ねられる。従って情報提示型が向いている場面としてあげられるのは、評価の項目が複数ありそのトレードオフを明確に定義できないとき、直感的に良し悪しを判定したいとき、データ化し難い編成者の知見や知識と統合した上で最終的な意思決定を行いたいときである。

情報提示型にはこのような利点がある一方で、VAで重要視されるような問題意識の観点では、システムから人間への一方的な情報の提示で意思決定に有用な知識を抽出できるのかという懸念がある。これを反映するようにチーム編成支援の領域でもVAシステムを提案する試みが存在する。例えばZhaoらは従業員のパフォーマンスを評価するためのVAシステムを提案している[99]。Ryooらはサッカーのクラブ間での選手移籍をサポートすることを目的にVAを適用している[100]。これらの研究は主にメンバーの組合せというよりも個々のメンバーの成果への貢献度に焦点を当てている。

ここまで既存研究を踏まえた上で、本論文が取り組む集合データVAの実現は、チーム編成支援の領域に貢献しうるものだと考えられる。その理由は大きく2つある。1つ目はVAという人間主導のアプローチが、特にチーム編成のような重大な意思決定を支援するのに適しているという点である。2つ目は効果的な支援のためにはメンバーの組合せと成果の非加算的な関係、即ちシナジーを考慮するのが重要であり、そのためには集合データを扱える必要があるという点である。既存研究、特に自動化型でメンバーの成果を予測する際には、メンバーの特徴量の集合に対して特徴量の総和や平均を取る[92]、与えられている役割を利用して特徴量を単純に連結する固定ベクトルとして扱う[93]^{*10}、といった処理が行われる。これらは非加算的なふるまいであるシナジーを考慮する上では適切ではない。即ち集合データを取り扱うということは、チーム編成支援の観点でも取り組む必要のある課題である。

^{*10} より詳しく言えばJayanthら[93]はクリケットのチームにおいて、ポジションごとに複数のメンバーの特徴ベクトルの平均を取った上で、ポジションごとの特徴ベクトルを連結することでチームの特徴ベクトルを得ている。これはポジション内で特徴量の平均を取っている点でシナジーを無視する処理であると言える。また一般に各メンバーに対してポジションのように明示的な役割が与えられているとは限らないため、チーム一般に適用可能な処理ではない。なお、各メンバーに役割が与えられている場合の本手法の適用方法については6章で検討している。

第 3 章

集合データ VA が満たすべき要件

第 2 章において、VA という意思決定支援の方法論の枠組みや、集合データを扱う際の困難について解説した。本章ではこれらを踏まえた上で、本論文で取り組む集合データを扱う VA（以降、集合データ VA と呼ぶ）とは具体的にどのようなものなのか、集合データ VA が満たすべき要件を定義することを通して述べる（3.1 節）。またこの要件を満たすにあたっての困難と、それを踏まえての本研究のチャレンジについて述べる（3.2 節）。

3.1 必要要件の定義

本節では集合データ VA が満たすべき要件を明確化する。一般的なケースについて考える前に、典型的な集合データ VA の活用先としてチーム編成の場面を想定し、具体例を交えながら要件を定義する。その上で要件を一般化し、特定の場面を限定しない要件を定義する。

3.1.1 チーム編成においてどのような VA システムが必要とされるか

集合データ VA の典型的な活用先であるチーム編成とは、例えば次のようない状況である（より実践的なケースについては [101] を参照のこと）。チームのマネージャは、選出可能なメンバーの集合が与えられている際に、チーム単位での成果が最大となるようなメンバーの部分集合を選択しなければならない。このときに、各メンバーについてその能力を表現するような特徴ベクトルや、過去に編成されたチームのラインナップやその時の成果といったデータを参照できるとする。このような状況において、マネージャは選出メンバーを決めるために、各メンバーがどのような特徴を持っているかをデータから把握する必要があるだろう。だが個々のメンバーの能力だけを考慮するのは不十分であり、メンバーの組合せも考慮する必要がある。なぜならシナジー効果 [15–17] と呼ばれる、メン

バーの組合せによっては個々人が出す成果の総量をチームが出す成果が上回るような現象が知られているためである。このメンバーの組合せの検討には、過去のデータからどのようなタイプのチームが結成され、どれだけの成果を達成したのか知ることは有効であろう。また多くの場合、マネージャはリスクやコスト、外的な条件などエクストラな要因についても考慮に入れておく必要がある。外的な条件の例としては、チームスポーツであれば敵チームのラインナップなどが挙げられる。これらの絡み合った要因を考慮しながら、マネージャは候補ラインナップを探査し、複数のラインナップを比較し、1つを選択するという意思決定を行う必要がある。またその後も、マネージャは意思決定についての説明も求められる。

このようなシナリオにおける意思決定を支援するための VA システムを考える。このシステムに求められる要件として次の 4 つを定義する。

1. チーム（メンバーの集合）を取り扱うことができる

第一の要件としては、システムはチーム、つまりメンバーの集合を取り扱えることである。より具体的には、マネージャがチーム同士を比較できる、メンバーの数や順序に依存しないチームの内在的な性質を表現した特徴量を獲得する必要がある。また逆にそのような特徴量がユーザから指定されたときには、システムはそれに対応するチームを生成し推薦できることが望ましい。

2. 複数の要因の間の複雑にもつれた関係性を解きほぐした分析を提供できる

第二にシステムはメンバーの能力、チームの性質、外的な要因、といったチームの成果を決定づける複数の要因が複雑に絡み合っているのをそれぞれの要因ごとに分解して分析結果を提示できる必要がある。チーム編成にまつわる過去のデータは、複数の要因が複雑に影響しあった結果として得られたものである。このデータから、メンバー同士の比較やチーム同士の比較といったそれぞれの要因の母集団（以降、ドメインと呼ぶ）の中で閉じた分析と、特定のチームと外的要因の組合せについてこのような成果が得られた、といった複数のドメインをまたぐ分析の両方を実現する必要がある。

3. マネージャが新しいラインナップを試行錯誤できる

第三に、システムはマネージャが新しいラインナップを探査できる「テーブルトップフィールド」を提供しなければならない。つまり、マネージャ自身が TOI として任意のラインナップ（とその他の条件との任意の組合せ）を与えることができ、システムはそれがどのような成果をもたらすのか予測しインタフェース上に提示するような機能が求められる。加えてそのような意思決定の結果「何が」起きるかということだけでなく、その成果が「なぜ」もたらされるか解釈するためのヒントとなる情報を示す事もまた望ましい機

能である。

4. 複数のラインナップの概観を提供できる

第四に、システムは考えうる複数のラインナップについて、各ラインナップがどのような傾向を持つかを一目で把握できる概観を提供しなければならない。これは、組合せ爆発を起こして無数に存在する候補ラインナップの中でマネージャが逐次的に候補を検討していかなければならない状況を避けるためである。

3.1.2 必要要件の一般化

前述した4つの要件を、チーム編成の場面に限らない集合データを扱う問題全般の要件として一般化すると次のようになる。

1. 集合データを扱える

第一に集合データを扱える必要がある。より厳密に言えば、与えられた集合データ X からその要素数や順序に依存しない集合データの内在的な性質を示す適切な「表現」を獲得する必要がある。なおかつ、逆にそのような性質が指定されたときに、対応する集合データ X を生成もしくは推薦できる必要がある。

2. ドメイン内分析とドメイン間分析の両方をユーザに提供できる

第二にドメイン内分析とドメイン間分析の両方を提供する必要がある。また、両者を別個のインターフェースとして提供するのではなく、一つのインターフェースの上で実現することで、ドメイン内分析で得られた仮説や知見を元に TOI を定めてドメイン間解析を行う、といったシームレスに分析を切り替えられることが望ましい。

3. 集合データとその出力の双方向な予測ができる

第三にシステムは集合データとその出力の双方向な予測ができる必要がある。つまりシステムは過去のデータから、任意の集合データに対する出力の予測モデルを獲得しなければならない（集合に対する識別タスク）。また与えられた出力から対応する集合データを生成しなければならない（集合の生成タスク）。

4. 組合せ探索問題を回避できる

第四に、システムはユーザを組合せ探索問題の煩わしさから解放しなければならない。すなわち、ユーザが一目で確認できる「集合データの集合の概観」を提供するために、集合データの集合の構造と、集合データの性質と出力の間の関係を可視化しなければなら

ない。

ここまで 4 つの要件は全て集合データ VA の「システム」が求められる要件である。本研究はこれらを満たす特定のシステムを開発するのではなく、これらの要件を満たす VA システムを構築する一般的な方法論を開発するものであり、ここでこの方法論に対する追加の要件として次の「可塑性」要件を追加する。

5. システム構築手法が個々の応用例に柔軟に適応できること（可塑性）

第五の要件として、システムの構築方法が個々の応用事例に適応できるように、VA システムを柔軟に設計できることを可能とすることを挙げる。入力（利用可能なデータセット）と出力（ユーザが見たい可視化された側面）は個々のケースによって異なるため、仮に応用領域が一緒であっても、手法に対してそのような汎用性は求められる。加えて、VA のプロセスはそれ自体がダイナミックなものであり、知見や仮説を得た後で異なる視点からの追加分析を行うことがしばしばある。その場合には新しいデータセットの追加や異なる側面での可視化が必要となる。したがって、VA システムは適応的で拡張的であることが望ましい。本研究では、この要件を可塑性（plasticity）要件と呼ぶことにする。この要件を満たすためには、VA システムを構築する方法論が特定のデータ構造に限定されず、一貫したアプローチで適応・拡張可能な汎用的な枠組みでなければならないと言える。

我々の知る限り、本研究はこの 5 要件に取り組む最初の試みである。その理由はおそらく、次節で詳しく述べる VA システムにおける集合データを取り扱うことの難しさにあると考えられる。

3.2 必要要件充足のための困難とチャレンジ

VA システムが前述の 4 つの要件を満たすためには (i) 与えられた集合データからその出力を推定すること（識別タスク）(ii) 特定の性質を満たすような集合データを生成すること（生成タスク）(iii) あらゆる集合データとその性質を可視化すること（可視化タスク），という集合データについての 3 つのタスクを同時に解く必要がある。

これらのタスクの難しさの根本的な要因は共通している。それは 2.2.2 節でも述べたように、集合データに対する距離や類似度の定義が非自明なことである。前述した識別・生成・可視化は機械学習における基本的なタスクであるが、これらはデータとデータの間に距離や類似度が定義できることが前提となっている。そのため、集合データから距離や類似度が定義できる表現を抽出しなければならないが、例えば集合データの要素の平均を取るような素朴な表現は、集合データがどのような要素で構成されているかという情報を過

度に削減してしまう可能性が高く、予測や知識発見の妨げになる。各タスクを解く上で重要なのは、集合内部の要素の多様性をできる限り損なわず、各タスクに適した表現をどう獲得するかということであり、これは非自明な問題である。

近年の研究で盛んに行われているのは 2.2.4 節で述べたように、あるタスクに対して有用な表現そのものの提案ではなく、有用な表現をデータから自動的に獲得する機構、即ち集合を取り扱えるニューラルネットワークのアーキテクチャの提案である。このアーキテクチャをニューラルネットワークに組み込むことで、定義された目的関数を最適化するような表現の抽出機構をデータから学習することができる。これを用いれば、識別のような特定のタスクについて目的関数を定義し、それに特化した表現を得られる可能性がある。

ただし、本研究では 3 つのタスクを同時に解く必要がある。この 3 つのタスクはしばしば互いに矛盾し、トレードオフの関係にある。例えば識別タスクの性能を上げるために獲得された表現はおそらく非常に高次元になり、可視化することが難しくなるだろう。この 3 つのタスク全てに最適な表現は存在せず、どれか 1 つのタスクに偏りすぎない折衷的な表現が求められる。このような状況において、どう目的関数を定義するかというのは非自明である。

また可視化特有の難しさとして、集合データをどのように可視化すれば VA に有用か自明ではないという点がある。要件の 1 つとしてドメイン内とドメイン間の可視化が求められるが、「集合」というドメイン（例：チーム）と、集合を構成する「要素」のドメイン（メンバー）をそれぞれ可視化し、なおかつこの 2 つのドメインの間の関係を可視化することが必要となる。特にドメイン間可視化については、集合と要素という非対称な関係について、何を可視化すればユーザの知識発見に繋がるのか特定しなければならない。この可視化の問題に取り組むには、前述した表現の抽出機構の自動獲得というアプローチよりも、可視化に適する表現を人手で設計するアプローチが適している。なぜなら可視化の良さを評価する目的関数を定義し、自動で表現を獲得することは困難だからである。なお、情報可視化の分野では集合の構造を持つデータに対して様々な可視化の試みがなされているが [102]、集合に対する仮定が本研究とは異なる。具体的には、集合の要素は距離が定義できない離散的なシンボルであり、集合と集合の間で要素が重複しているような集合族を仮定している。本研究で扱う集合データは、要素と要素の間に距離が定義できる（例：要素が数ベクトル）ことを仮定し、要素の重複に関しては問わない。従って、既存の集合データの可視化をそのまま本研究のケースに適用することはできない。

よって前述した 5 つの要件を満たす汎用的なシステム構築手法を実現するための中心的なチャレンジは、3 つのタスクの折衷的な集合データの表現を実現することである。より具体的に言えば識別や生成を実現するために、集合データに含まれる要素間の多様性を最大限に保ちながら、可視化に有用な表現を集合データから獲得することである。

第 4 章

生成的多様体ネットワークモデリング

本章では第 3 章で述べた要件を満たす VA システム構築の方法論である生成的多様体ネットワークモデリングを提案する。提案手法では集合データを含むデータセットの生成過程を包括的にモデル化し、そのモデルをインタラクティブに可視化する。まずはインタラクティブな可視化を可能とする生成過程のモデリング方法である多様体モデルについて説明する（4.1 節）。次に提案手法の全体的な枠組みについて述べ（4.2 節）、特定の状況に対する学習タスクとして与えられるデータセットと推定対象を示し（4.3 節）、実装レベルでどのようにモデルを推定するのか説明する（4.4 節）。加えて、獲得した生成モデルを元にどのようにインタラクティブな視覚的インターフェースを実現するか述べる（4.5 節）。最後に提案手法が要件を満たしていることを、最も素朴なデータセットの状況において確認する（4.6 節）。4.1 節から 4.5 節で述べるキーアイディアを全て活用し、集合データ VA システムを構築する手法のことを本論文では生成的多様体モデリングネットワークモデリングと呼ぶ。

4.1 多様体モデル

提案手法の枠組みでは複数種類のデータセットに対して生成過程のモデル化を行う。チーム編成の例で言えば、複数のチームについてのデータは集合データの集合になる。また選択可能な全てのメンバーの集合は特徴量の集合となる。

提案手法でデータの生成過程のモデル化をする際に、データの種類に依らず導入するのが、「意味のあるデータは低次元な部分空間（多様体）に分布している」という仮定^{*1}である

^{*1} 2.1.3 節で述べた次元削減も、本質的には「データが低次元な空間に分布している」という仮定を置いており、VA の領域では一般的な仮定であると言える。

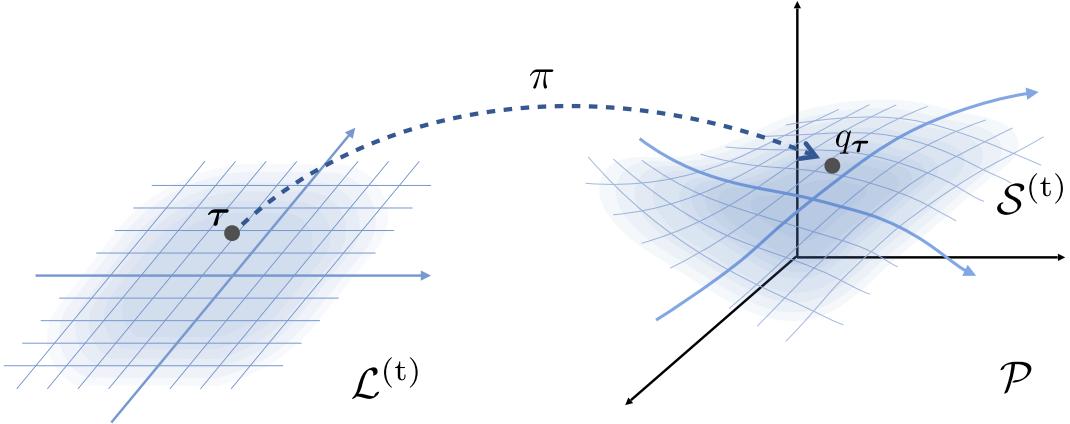


図 4.1 提案手法でデータの生成過程のモデル化に用いる多様体モデルの概念図. ここではチームの生成過程を考えている. 2つのチームの間に距離が定義されているようなチームの連続位相空間 \mathcal{P} を仮定する. その上で意味のあるラインナップを取るチームは非線形な部分空間 $\mathcal{S}^{(t)} \subseteq \mathcal{P}$ に分布すると仮定する. この $\mathcal{S}^{(t)}$ は全単射 $\pi: \mathcal{L}^{(t)} \leftrightarrow \mathcal{S}^{(t)}$ によって低次元な潜在空間 $\mathcal{L}^{(t)}$ と 1 対 1 の対応関係を持つ. その上で多様体 $\mathcal{S}^{(t)}$ 上の q_τ は, 潜在変数 τ が写像 π によって写されて生成されたとする.

る. チーム集合を例にして具体的に説明すると次のようになる. まず 2 つのチームについての距離が定義されているようなチームの連続位相空間 \mathcal{P} を仮定する. このとき, 考慮に値する現実的なラインナップを持つチームは（非線形な）部分空間である $\mathcal{S}^{(t)} \subseteq \mathcal{P}$ に分布していると仮定する. この部分空間 $\mathcal{S}^{(t)}$ とは, 低次元な潜在空間 $\mathcal{L}^{(t)}$ (典型的には $\mathcal{L}^{(t)} = \mathbb{R}^d, d \in \mathbb{N}$) との一対一の対応関係が全単射 $\pi: \mathcal{L}^{(t)} \leftrightarrow \mathcal{S}^{(t)}$ によって定義されている空間であり, 多様体と呼ばれる. この仮定の元で多様体上のチーム $q_\tau \in \mathcal{L}^{(t)}$ は次のように生成されるとする. まず潜在空間上でチームの性質を表現する潜在空間 τ が生成される. この τ が写像 π によって写され q_τ が生成される. このような生成モデルのことを本論文では多様体モデルと呼ぶことにする.

チーム集合に対する多様体モデルは, 第三の要件「集合データとその出力の双方向な予測ができる」と第四の要件「組合せ探索問題を回避できる」という 2 つの要件に対する解決策となっている. 前者の理由は, ラインナップを潜在空間上の表現に落とし込むことで予測モデルを学習できるためである. この多様体の仮定は過去に観測されたラインナップが多様体上に分布していることを意味している. そのため過去のデータから多様体を推定することができれば, ラインナップを距離の定義された潜在空間上の表現（潜在変数）に変換することができる. 潜在変数を利用すれば, チームの潜在変数を入力, チームの成果を出力とした関数を推定することも可能である. またその逆問題も解くことで第三の要件を満たすことができる. 後者の第四の要件については, チーム編成を解くためにメンバー構成を探索する際に探索空間をチームの空間 \mathcal{P} 全体ではなく多様体 \mathcal{S} に制限することで

解決する。潜在空間 $\mathcal{L}^{(t)}$ を次元を 2 としたうえで、全単射によって多様体上のチームを潜在空間に写すと、チームの集合の全体像を散布図（マップ）として可視化することができる。このマップの上で成果についての予測結果も併せて可視化することで、1 対 1 の比較を行わなければならない状況を避けることができる。

同様に、メンバーの性質も特徴量の空間上で低次元多様体（典型的には 2 次元）多様体に分布すると仮定する。即ちメンバーの特徴ベクトルは多様体 $\mathcal{S}^{(t)} \subseteq \mathcal{O}^{(m)}$ に分布すると仮定する。ただし $\mathcal{O}^{(m)}$ はメンバー混合分布のベクトル空間である。この時、多様体と位相同型な潜在空間 $\mathcal{L}^{(m)}$ と双方向写像 $g: \mathcal{L}^{(m)} \leftrightarrow \mathcal{S}^{(m)}$ を定義できる。この双方向写像 g は $\mu \in \mathcal{L}^{(m)}$ のメンバーを特徴空間に $x = g(\mu)$ として写像する。また逆も然りである。加えて他の考慮に入れなければならないデータセットについても多様体モデルによってモデル化する。例えば、外的要因について考慮しなければならないとき、多様体 $\mathcal{S}^{(c)} \subseteq \mathcal{O}^{(c)}$ と潜在空間 $\mathcal{L}^{(c)}$ を定義でき、このとき全単写は $h: \mathcal{L}^{(c)} \leftrightarrow \mathcal{S}^{(c)}$ となる。

メンバーの特徴量についてもチームと同様に多様体モデルでモデル化することが、第一の要件「集合データを扱える」を満たすためのキーアイディアである。なぜならこれによって、チームのラインナップはどのような性質を持つメンバーによって構成されているか、メンバー潜在空間 $\mathcal{L}^{(m)}$ 上での確率分布 $p(\mu)$ として表現することができるようになるためである。つまりチームの連続位相空間 \mathcal{P} はメンバー潜在空間 $\mathcal{L}^{(m)}$ 上の確率分布からなる関数空間として定義する。この $p(\mu)$ を本論文ではしばしばメンバー構成と呼ぶことにする。従って我々の手法においては、チームを離散的な集合の代わりに確率分布として取り扱う。このアプローチは 2 つのチームの間の距離の定義や新しいラインナップの生成を可能とする。

4.2 提案手法の枠組み

提案手法の枠組みについて定式化する。 Ω を全てのメンバーの集合とする。 n 人のメンバーで構成されるチームは集合 $T = \{\omega_1, \dots, \omega_n\}$ と表される。ただし $\omega_i \in \Omega$ である。本論文では、集合 T をチームのラインナップと呼ぶ。また \mathfrak{T} を全てのラインナップの集合とする。すなわち \mathfrak{T} は集合の集合である。なお、一般にメンバーの順序は固定されていない。また $\mathbf{x}(\omega) \in \mathcal{O}^{(m)}$ をメンバー ω の能力のような性質を表す特徴ベクトルとする。このときラインナップ T について特徴ベクトルの集合、即ち集合データ $X = \{\mathbf{x}(\omega_1), \dots, \mathbf{x}(\omega_n)\}$ が得られる。 T が決まると X も決まる事になるので、 X もまたラインナップと呼ぶことにする。

チーム編成支援の関連研究の典型的なモデル（図 4.2(a)）はチームの成果 $y \in \mathbb{R}$ が $y = f(X)$ というように X を直接入力とする関数 f によって決まると仮定する（図 4.2(a))。この f を用いて y を最大化するチームラインナップ $T_{\text{opt}} \in \mathfrak{T}$ を求めるのが典型

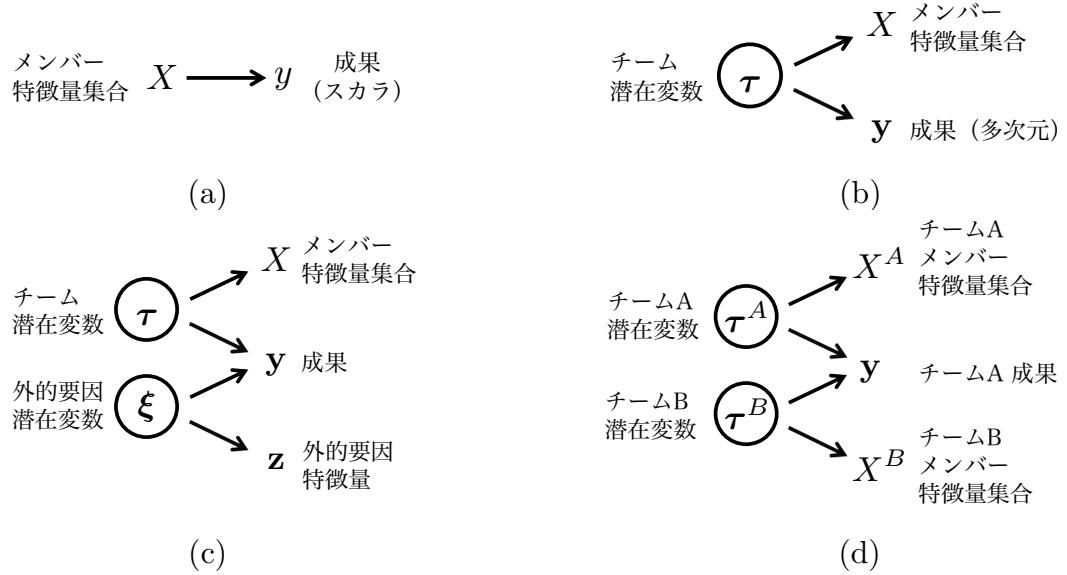


図 4.2 典型的な自動化型支援と提案手法それぞれでデータに対して仮定するモデル。前者を (a)、後者を (b, c, d) に示している。円は潜在変数を表し、矢印の元の変数が矢印の先の変数の生成に影響を与える。(a) 典型的な自動化型のアプローチ。集合データ X が決まると成果が定まる (b) チーム潜在変数 τ を仮定した生成モデル。(c) 生成モデル (b) に外的要因を考慮したモデル。(d) チームスポーツにおけるモデル。

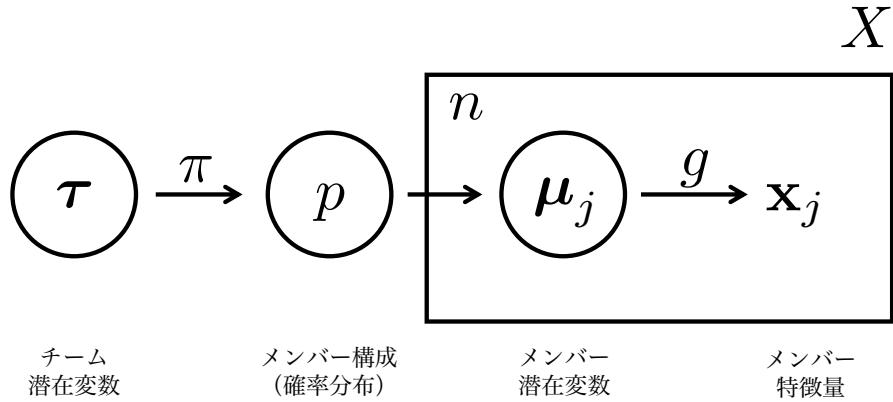


図 4.3 チームラインナップ X の生成モデル。まずチーム潜在変数 τ が生成され、写像 π によってメンバー構成を表現するメンバー潜在空間上の確率分布 p が定まる。この確率分布からラインナップを構成する個々のメンバーの潜在変数 μ_j が生成され、それが g によって写像されて最終的なメンバーの特徴量 \mathbf{x}_j となる。

的な問題設定である。多くの研究ではこの f が既知もしくは論文の著者が設計する。一部の研究 [92, 93] ではデータから帰納的に獲得する。

それに対して提案手法の枠組みではラインナップ X と成果 y (提案手法の場合はベクトルでも可) の両方が、チームの内在的な性質を表現する潜在変数 $\tau \in \mathcal{L}^{(t)}$ から生成されると仮定する。すると、成果は直接ラインナップ X からではなく $y = f(\tau) + \epsilon$ と生成さ

れる。ここで ϵ は観測の際に加わるノイズである。従来の枠組みとは異なり、 f は事前に与えられず、データから推定される。ここで成果 $\mathbf{y} \in \mathcal{O}^{(o)}$ についても多様体 $\mathcal{S}^{(o)} \subseteq \mathcal{O}^{(o)}$ 上に分布すると仮定する。成果 \mathbf{y} とは並行して、潜在変数 $\boldsymbol{\tau}$ から写像 π によって写像された $\pi(\boldsymbol{\tau})$ にノイズが加わることでチームのメンバー構成を表現する確率分布 p が生成される。ここで $q_{\boldsymbol{\tau}} \equiv \pi(\boldsymbol{\tau})$ とすると、この $q_{\boldsymbol{\tau}}$ を元に de Finetti の表現定理 (2.1) の形でチームのラインナップ X の生成モデル $q(X)$ を次のように定義する。

$$q(X) = \int q(X \mid \boldsymbol{\tau}) d\boldsymbol{\tau} \quad (4.1)$$

$$q(X \mid \boldsymbol{\tau}) = \prod_j \int \mathcal{N}(\mathbf{x}_j \mid g(\boldsymbol{\mu}), \beta^{-1} \mathbf{I}) q_{\boldsymbol{\tau}}(\boldsymbol{\mu}) d\boldsymbol{\mu} \quad (4.2)$$

ここで $q_{\boldsymbol{\tau}} \equiv q(\boldsymbol{\mu} \mid \boldsymbol{\tau})$ であり β^{-1} はガウスノイズの分散、 \mathbf{I} は単位行列である。これをグラフィカルモデルとして示したのが図 4.3 である^{*2}。このモデル化は 2.2.5 節で示した式 (2.3) の m 型の集合データの生成モデルとなっている。一方でラインナップ $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ が与えられているとき、対応する潜在変数 $\boldsymbol{\tau}(X)$ を

$$\boldsymbol{\tau}(X) = \arg \min_{\boldsymbol{\tau}} D_{KL}[p(\boldsymbol{\mu} \mid X) \parallel q_{\boldsymbol{\tau}}(\boldsymbol{\mu})] \quad (4.3)$$

と定める。ここで $p(\boldsymbol{\mu} \mid X)$ はカーネル密度推定により与えられ、 D_{KL} はカルバック・ライブラー (KL) ダイバージェンスである。この双方向写像がチーム潜在変数とラインナップとの間の双方向な探索を可能とする。これが本研究の利点である。

この生成モデルは、いくつかの追加の要因を考慮しなければならないような他のシナリオにも容易に適応できる。もし外的な条件を考慮する必要がある場合、条件の内在的な性質を表現するような潜在変数 $\boldsymbol{\xi} \in \mathcal{L}^{(c)}$ を追加することでモデルを拡張できる。このケースにおいては、期待できる成果は $\mathbf{y} = f(\boldsymbol{\tau}, \boldsymbol{\xi})$ と推定される（図 4.2(c))。また別の典型的な例としてはチームスポーツにおける対戦である。この場合、 $\mathbf{y} = f(\boldsymbol{\tau}_A, \boldsymbol{\tau}_B)$ のように対戦する 2 つのチームそれぞれの潜在変数から勝敗が決まると仮定する（図 4.2(d))。

我々のアプローチにおいては、生成モデル全体は多様体モデルのネットワーク構造、すなわち MNM となる。図 4.4 は外的要因を考慮した図 4.2(c) の場合の MNM のネットワーク構造を示したものである。MNM においてはラインナップの生成過程は $\mathcal{S}^{(t)}$ と $\mathcal{S}^{(m)}$ の 2 つの多様体モデルの潜在空間 $\mathcal{L}^{(m)}$ を介した直列な結合によって表現される。それに対して成果や重要な要因は多様体 $\mathcal{S}^{(t)}, \mathcal{S}^{(c)}, \mathcal{S}^{(o)}$ が潜在空間 $\mathcal{L}^{(t)}, \mathcal{L}^{(c)}$ を介した並列な結合によってモデル化される。このネットワーク構造は個々の応用例に適応できるため、提案する枠組みは第五の要件である可塑性要件を満たしていると言える。

^{*2} 図 2.3(b) の図に $q_{\boldsymbol{\tau}}$ を明示的に示しているのが図 4.3 となっている。提案手法においてはメンバー構成 $q_{\boldsymbol{\tau}}$ の存在が集合データの取り扱いと可視化の面で重要となる。

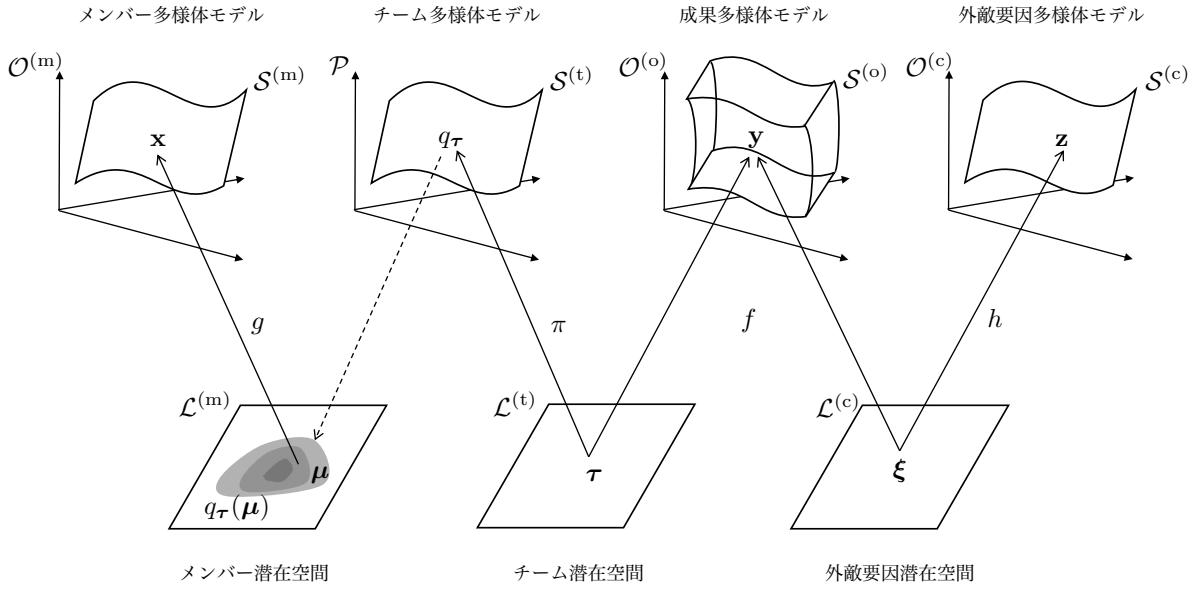


図 4.4 図 4.2(c) のシナリオにおける提案手法で構築した生成モデルの構造を示したもの。MNM は 4 つのアウトプットに対応する 4 つの多様体モデルと 3 つの潜在空間から構成される

4.3 問題設定

提案手法の枠組みにおいては MNM を与えられているデータセットから帰納的に推定する。ここでは外的要因を考慮する場面(図 4.2(c))を題材として、その MNM を図 4.4 を示し、具体的にどのようなデータが与えられていて、何を推定するのかを示していく。

まず与えられているデータは次の通りである。 $N^{(m)}$ 人のメンバーについての特徴量 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{N^{(m)}})^\top$ とそのメンバー集合から構成される $N^{(t)}$ 個のラインナップ $\mathcal{L} = \{T_i\}_{i=1}^{N^{(t)}}$ 、さらに $N^{(c)}$ 個の外的要因についての特徴量 $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{N^{(c)}})^\top$ が与えられる。ただし $T_i \in \mathfrak{T}$, $\mathbf{x}_j \in \mathcal{O}^{(m)}$, $\mathbf{z}_k \in \mathcal{O}^{(c)}$ である。さらに、 $N^{(o)}$ 回の試行に対する成果 $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{N^{(o)}})^\top$, $\mathbf{y}_l \in \mathcal{O}^{(o)}$ とその成果がもたらされた際のラインナップと外的要因の組合せ $\{(T_{i(l)}, \mathbf{z}_{k(l)})\}_{l=1}^{N^{(o)}}$ も与えられる。ただし、この $N^{(o)}$ 個の組合せは全ての取りうる組合せを網羅している必要はない。

このときに、提案手法のタスクは写像 f, g, h, π とチーム潜在変数 $\mathbf{T} = (\tau_1, \dots, \tau_{N^{(t)}})^\top$, $\tau_i \in \mathcal{L}^{(t)}$, メンバー潜在変数 $\mathbf{M} = (\mu_1, \dots, \mu_{N^{(m)}})^\top$, $\mu_j \in \mathcal{L}^{(m)}$, 外的要因潜在変数 $\mathbf{\Xi} = (\xi_1, \dots, \xi_{N^{(c)}})^\top$, $\xi_k \in \mathcal{L}^{(c)}$ を推定することである。

4.4 実装

4.4.1 生成的多様体モデリング (GMM)

ここまで述べたように、本論文で提案するアプローチのコアとなるアイデアは MNM を用いてデータセット全体の生成モデルを表現することである。これを達成するために、実装においては生成的多様体モデリング (Generative Manifold Modeling: GMM) をビルディングブロックとして適用する。すると MNM は GMM のネットワークとして推定される。これが本研究の実装面におけるキーアイデアである。

本論文における GMM とは、与えられたデータセットが多様体でモデル化する教師なし学習のパラダイムのことを指している。この目的のために、GMM はデータセットから低次元な潜在空間から高次元なデータ空間への写像と、各データに対応する低次元な潜在変数を推定する。学習を終えた後は、既存の観測データの分析と未観測データの生成が可能となる。これが提案手法で GMM を用いる理由である。代表的な GMM の手法としてはガウス過程潜在変数モデル (Gaussian Process Latent Variable Model: GPLVM) [103] や教師なしカーネル回帰 (Unsupervised Kernel Regression: UKR) [104] がある。

GMM の問題設定は次の通りである。 $\mathcal{O} = \mathbb{R}^D$ と $\mathcal{L} = \mathbb{R}^d$ をそれぞれ高次元な観測空間と低次元な潜在空間とする。高次元なデータセット $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)^\top \in \mathbb{R}^{N \times D}$ が与えられているときに、 $\forall j \mathbf{z}_j = f(\boldsymbol{\xi}_j \mid \boldsymbol{\Xi})$ となるような、データに対応する対応する潜在変数 $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N)^\top \in \mathbb{R}^{N \times d}$ と写像 $f: \mathcal{L} \rightarrow \mathcal{O}$ を推定することである。

UKR ではこれを次のように解く。まず、潜在変数 $\boldsymbol{\Xi}$ によって、写像 f が次のように表現されるとする。

$$f(\boldsymbol{\xi} \mid \boldsymbol{\Xi}) = \frac{1}{K(\boldsymbol{\xi})} \sum_{j=1}^N k(\boldsymbol{\xi}, \boldsymbol{\xi}_j) \mathbf{z}_j \quad (4.4)$$

$$K(\boldsymbol{\xi}) = \sum_{j=1}^N k(\boldsymbol{\xi}, \boldsymbol{\xi}_j) \quad (4.5)$$

$$(4.6)$$

ただし $k(\boldsymbol{\xi}, \boldsymbol{\xi}')$ は $\boldsymbol{\xi}$ と $\boldsymbol{\xi}'$ の類似度を定義するカーネル関数である。この関数の表現方法はカーネル回帰と同様である。このように写像を $\boldsymbol{\Xi}$ が決まれば一意に定まるように定義することで、アルゴリズム上の推定対象は $\boldsymbol{\Xi}$ のみとなる。 $\boldsymbol{\Xi}$ は次の目的関数を最大化するように推定する。

$$F_{\text{UKR}}(\boldsymbol{\Xi} \mid \mathbf{Z}) = -\frac{\beta}{2} \sum_i \|f(\boldsymbol{\xi}_i) - \mathbf{z}_i\|^2, \quad (4.7)$$

UKR の利点は確率分布の集合をデータセットとして扱うことができる点である。その場合は(4.7)のユークリッド距離は KL ダイバージェンスに置き換える。

一方で GPLVM の場合、潜在変数 Ξ によって定義される写像 f が確率的に表現される。具体的には、次のようなガウス過程^{*3}となる。

$$f_d(\xi) \sim \mathcal{GP}(m_d(\xi), \text{cov}(\xi, \xi')) \quad (4.8)$$

$$m_d(\xi) = \mathbf{k}(\xi) (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{z}_{:d} \quad (4.9)$$

$$\text{cov}(\xi, \xi') = k(\xi, \xi') - \mathbf{k}^\top(\xi) (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{k}(\xi') \quad (4.10)$$

ここで関数 f の出力はベクトルであるが、要素ごとに $f(\xi) = (f_d(\xi))_{d=1}^D$ と分解し、そのうち d 次元目のみの関数 f_d に注目している。 $\mathbf{z}_{:d} \in \mathbb{R}^N$ は \mathbf{Z} の d 番目の列のみを取り出した列ベクトルである。また $\mathbf{k}(\xi) = (k(\xi, \xi_1), \dots, k(\xi, \xi_N))^\top$, $\mathbf{K} = (k((\xi_n, \xi_{n'}))$ である。UKR と同様にアルゴリズム上の推定対象は潜在変数 Ξ のみとなり、次の目的関数（周辺尤度の対数）を最大化するように求める。

$$F_{\text{GPLVM}}(\Xi | \mathbf{Z}) = -\frac{ND}{2} \ln 2\pi - \frac{D}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \text{Tr} [\hat{\mathbf{K}}^{-1} \mathbf{S}] \quad (4.11)$$

ただし観測データについてのグラム行列を $\mathbf{S} = \mathbf{Z}\mathbf{Z}^\top$ としている。また $\hat{\mathbf{K}} = \mathbf{K} + \beta^{-1} \mathbf{I}$, $\mathbf{k}(\xi) = (k(\xi, \xi_i))_{i=1}^N$ であり、 β は正規分布を仮定した観測ノイズの精度パラメータである。UKR との大きな違いは写像を確率的に推定していることである。このとき任意の ξ に対する関数の平均 $\mathbf{m}(\xi)$ だけでなく、任意の ξ, ξ' に対する共分散 $\text{cov}(\xi, \xi')$ も推定することができる。自己共分散 $\text{cov}(\xi, \xi)$ は推定値 $f(\xi)$ の不確定性の大きさを表現していると見なせる。この不確定性は可視化にも活用することができる。Lawrence [103] は潜在空間のマップによる可視化において、分散が小さい領域は明るく、大きい領域は暗く表示している。これが VA の観点における GPLVM の利点である。

提案手法の実装では、それぞれの利点を考慮し、GPLVM を成果の予測関数 f , UKR をラインナップの生成関数 π のモデリングにそれぞれ利用する。

4.4.2 GMM を用いた多様体ネットワークモデルの推定

MNM の構造は 4 種類の接続からなる。

- (a) 単一の潜在空間から対応する単一の多様体への写像（例：図 4.4 における $\mathcal{L}^{(m)}$ から $\mathcal{O}^{(m)}$ への写像 g ）
- (b) メンバー構成の多様体とその要素の潜在空間との接続 ($\mathcal{L}^{(m)}$ と $\mathcal{S}^{(t)}$ の間の接続)

^{*3} ガウス過程の詳細な解説は文献 [105] や [106] を参照のこと。

- (c) 単一の潜在空間から 2 つ以上の多様体への写像（例：潜在空間 $\mathcal{L}^{(t)}$ から観測空間 $\mathcal{S}^{(o)}$ への写像 f と同じく $\mathcal{L}^{(t)}$ から $\mathcal{S}^{(t)}$ への写像 π ）
- (d) 2 つ以上の潜在空間から多様体への写像（例：チーム潜在空間 $\mathcal{L}^{(t)}$ と外的要因の潜在空間 $\mathcal{L}^{(c)}$ から成果の多様体 $\mathcal{S}^{(o)}$ への写像 f ）

これらの接続に対して、GMM の適用により次のように推定する。

(a) 単一の潜在空間から単一の多様体への写像：GMM による推定

これは通常の GMM の問題設定そのものであるため、GPLVM もしくは UKR によって推定する。

(b) メンバー構成の多様体とその要素の潜在空間との接続： m 型集合データモデリングやカーネル密度推定

まずメンバー構成の多様体 $\mathcal{S}^{(t)}$ からメンバー潜在空間 $\mathcal{L}^{(m)}$ への対応は、多様体上の座標系（潜在空間）上の τ からメンバー構成を表現する確率分布 $q_\tau = q(\mu | \tau)$ を推定することである。これは m 型の集合データモデリングによって実現できる。

それに対して、 $\mathcal{L}^{(m)}$ から $\mathcal{S}^{(t)}$ への対応は、チームラインナップ $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ をカーネル密度推定を用いてメンバー構成

$$p_X(\mu) = \frac{1}{n} \sum_j k(\mu | g^{-1}(\mathbf{x}_j)) \quad (4.12)$$

に変換することで実現する。もしラインナップが重み付きの集合として定義されるのであれば、重み付きのカーネル密度推定

$$p_X(\mu) = \frac{1}{W} \sum_j w_j k(\mu | g^{-1}(\mathbf{x}_j)) \quad (4.13)$$

と推定できる。ただし w_i は i 番目の要素の重みである。例えばスポーツの試合の場面において、重み w_i は i 番目の選手のプレイ時間として定義することができる。

(c) 単一の潜在空間から 2 つ以上の多様体への写像：複数の GMM の同時学習

この場合は複数のデータセットが与えられているときに、それぞれの位相空間における多様体を共通した潜在空間に対応させて学習する必要がある。これを実現するには、それぞれのデータセットに対して GMM を導入し、これらの GMM が推定する潜在変数を共通化させれば良い。これを本論文では次のように実装する。いま潜在変数を共通化したい GMM が 2 つあるとすると、 $\alpha_1 F_1 + \alpha_2 F_2$ のように 2 つの GMM の目的関数の線形和を取り、これを最適化することで共通の潜在変数を推定する。この係数 α_1, α_2 の比によっ

て、潜在変数に対してどのデータセットの影響を強くなるかが決まる。今回はそれぞれの目的関数の勾配ベクトルのノルムの平均が数値的に等しくなるように係数を適応的に定めることにする。

(d) 2つ以上の潜在空間から多様体への写像：結合カーネル関数の定義

複数の潜在空間から单一の多様体^{*4}への写像を推定するために、GMMで用いるカーネルを複数のカーネルの積として定義する。例えばこれらの各接続に対する推定方法を図4.4の f のカーネル関数は $k(\tau, \xi, \tau', \xi') \equiv k(\tau, \tau') \times k(\xi, \xi')$ と定義される。

4つの接続法を適用した例

これらの各接続に対する推定方法を図4.4の場合に適用すると、メンバー、チーム、成果、外的要因それぞれの多様体モデルに対応する GMM の目的関数 $F^{(m)}(\mathbf{M} | \mathbf{X})$, $F^{(t)}(\mathbf{T} | P)$, $F^{(c)}(\mathbf{\Xi} | \mathbf{Z})$, $F^{(o)}(\mathbf{T}, \mathbf{\Xi} | \mathbf{Y})$ が以下の2つの目的関数へと集約される。

$$F(\mathbf{M}) = F^{(m)}(\mathbf{M} | \mathbf{X}), \quad (4.14)$$

$$F(\mathbf{T}, \mathbf{\Xi}) = \alpha_1 F^{(t)}(\mathbf{T} | P) + \alpha_2 F^{(c)}(\mathbf{\Xi} | \mathbf{Z}) + \alpha_3 F^{(o)}(\mathbf{T}, \mathbf{\Xi} | \mathbf{Y}), \quad (4.15)$$

推定する際の順序としては (4.14) を最適化し得られた潜在変数 \mathbf{M} を用いてメンバー構成 $P = \{p_i(\mu)\}$ を推定し、(4.15) を最適化する。

このように、ネットワークの構造によらず、あらゆる MNM はこの4種類の接続を組み合わせることで推定することができる。したがって、実装のレベルでも可塑性要件は満たされる。

4.5 可視化法

本節ではデータから獲得した MNM をベースとして、どのように人間の TOI の指定に対して可視化を実現するか述べる。なお、これ以降は具体的な VA システムについて述べるため、VA システムとインタラクションする人間のことをユーザと呼ぶことにする。

4.5.1 視覚的インターフェースの構築

4.3節で述べたように、図4.4の状況ではデータからチームについての潜在変数 \mathbf{T} 、メンバーについての潜在変数 \mathbf{M} 、外的要因についての潜在変数 $\mathbf{\Xi}$ を推定する。これらの潜在変数から、メンバーマップ、チームマップ、外部要因マップを生成することができる（図4.5）。このマップはデータ間の類似関係を表現しており、似ている性質を持つデー

^{*4} この場合は積多様体とも呼ばれる

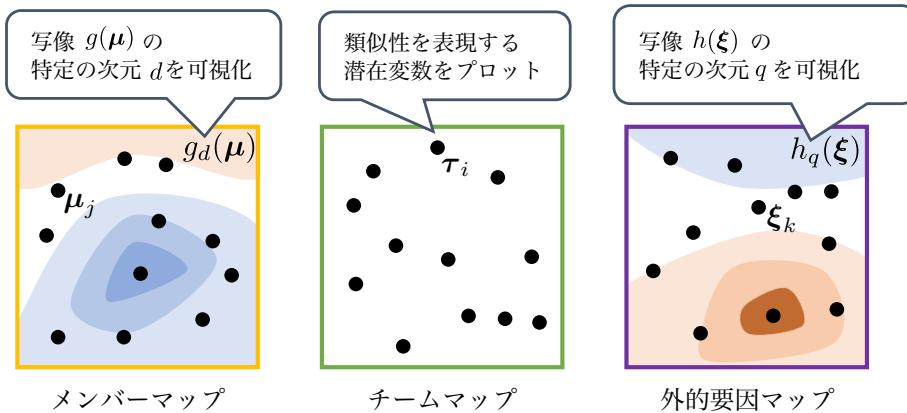


図 4.5 提案手法により生成されるメンバーマップ, チームマップ, 外的要因マップのイメージ図. 各マップにはデータの類似性を表現した潜在変数がプロットされる. またメンバーマップは写像 g を用いることで, 選択した特徴量次元 d についてその値をヒートマップに表示できる. 同様に外的要因マップでも写像 h を用いて, 選択した次元 q についてのヒートマップを表示できる

タはマップ上で近くに配置されている. 提案手法においてはこのマップを視覚的インターフェース上に配置し, データ探索のプラットフォームとする.

改めて, 本論文では潜在変数の可視化結果を「マップ」と呼んでいるが, 実際の地形図では等高線やヒートマップ表示を用いて追加の情報を可視化することがよく行われる. 提案手法においても, 潜在変数と同時に学習した非線型写像 g, h の値を等高線やヒートマップとして描画することで, より豊かな可視化を行うことができる. 写像がベクトル値関数である場合は, ユーザがその中の 1 つの成分を選んで可視化することができる. 例えば $g(\mu)$ の特定の成分 d の値を $g_d(\mu)$ とすると, この $g_d(\mu)$ でメンバーマップを着色することができる. これによりメンバーマップ上でどのような特徴を持つメンバーがどこに配置されているかを一目で把握することができる. 同様の着色が外的要因マップにおいても可能である.

4.5.2 ユーザの TOI の指定に対する可視化

これら複数のマップをインターフェースとして, ユーザからの TOI の指定に応じた可視化を行うことができる. 各マップを TOI として指定した際に, 次のような可視化が行われる.

メンバーマップ 図 4.6(a) に示すように, ユーザがメンバーマップ上の特定の座標 $\hat{\mu}$ にいるメンバーの特徴を知りたいとき, マップ上を TOI として指定すること

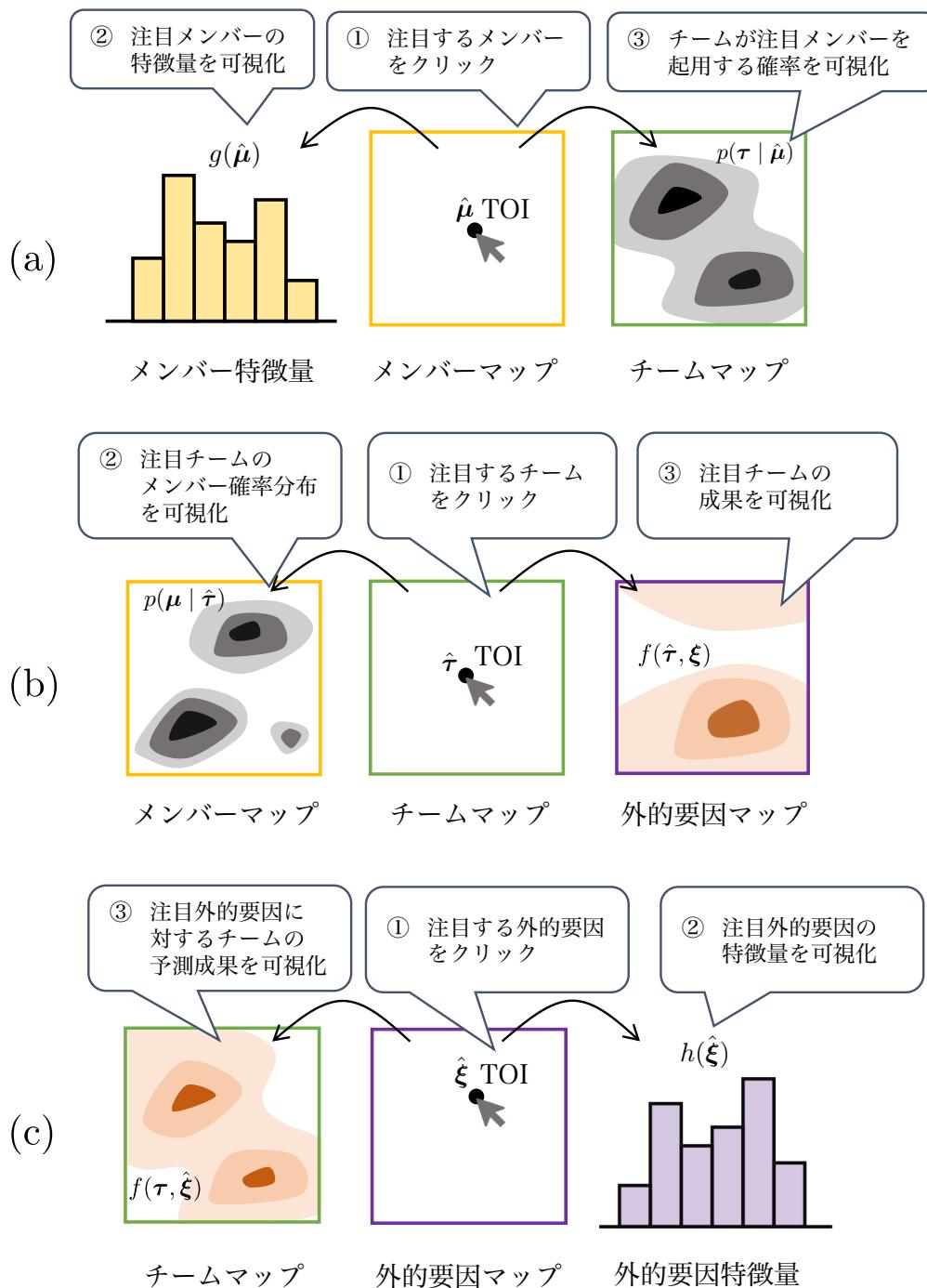


図 4.6 マップ上で可能な操作とそれに対する可視化の一覧. (a) メンバーマップ上の任意の座標 $\hat{\mu}$ を TOI として指定することで、そのメンバーはどのような特徴を持つか、またどのようなチームによく起用されるか可視化される. (b) チームマップ上の任意の座標 $\hat{\tau}$ を TOI として指定することで、そのチームはどのようなメンバーで構成されるか、あらゆる外的要因に対してどれだけの成果を出せるか可視化される. (c) 外的要因マップの任意の座標 $\hat{\xi}$ を選択することで、その外的要因の特徴やどのチームが成果を出しやすいかが可視化される.

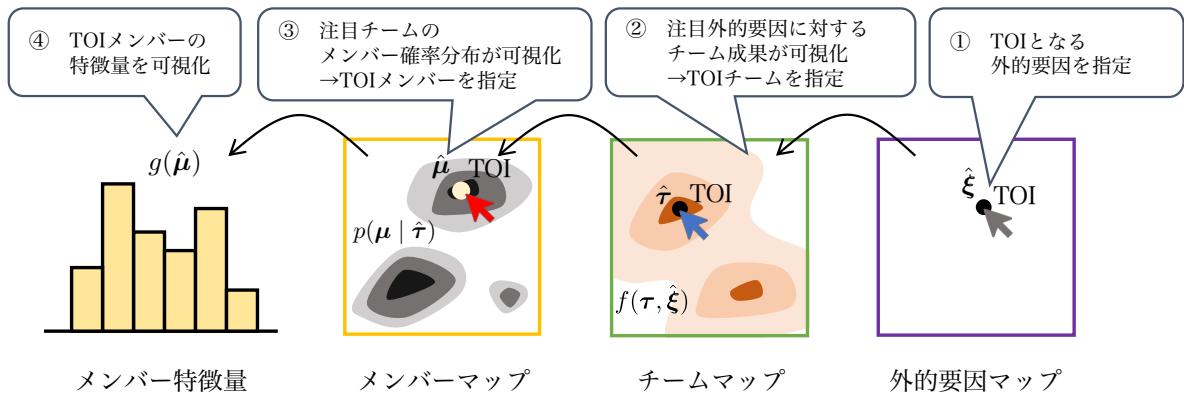


図 4.7 図 4.6 の操作と可視化を組み合わせることで実現できるデータ分析過程の一例. ①外的要因を指定する. ②その外的要因に対してあらゆるチームがどのような成果を出せるのか予測した結果がチームマップに可視化される. ユーザは成果が大きい領域に注目し $\hat{\tau}$ を TOI として指定したとする. ③メンバーマップ上に $\hat{\tau}$ のチームのメンバー構成を表現する確率分布 $p(\mu|\hat{\tau})$ が可視化される. ユーザが起用される確率が高い領域の $\hat{\mu}$ を TOI として指定したとする. ④その特徴量 $g(\hat{\mu})$ が可視化される.

で $\hat{\mu}$ を指定すると, $g(\hat{\mu})$ を棒グラフとして表示する^{*5}. 同時に $p(\tau|\mu)$ をチームマップ上に描画する. これは τ のチームが μ のメンバーを起用している確率である. これはベイズの定理を用いて $\pi(\mu|\tau)$ から計算できる. それに加えて, $\hat{\tau} = \arg \max_{\tau} p(\tau|\hat{\mu})$ による $f(\hat{\tau}|\xi)$ によって外的要因マップを着色することもできる.

チームマップ 図 4.6(b) に示すように, チームマップ上の任意の座標 $\hat{\tau}$ にあるチームを解析したい場合は, TOI として指定で $\hat{\tau}$ を指定することで, 確率分布 $p(\mu|\tau)$ でメンバーマップを着色する. 同時に, 期待される成果 $f(\hat{\mu}, \xi)$ も外的要因マップ上に可視化できる.

外的要因マップ 図 4.6(c) に示すように, 外的要因マップ上にある任意の座標 $\hat{\xi}$ にある外的要因について解析したい場合は, TOI として指定で $\hat{\xi}$ を指定することでその特徴量 $h(\hat{\xi})$ を棒グラフとして表示することができる. また, 期待される成果 $f(\mu, \hat{\xi})$ をチームマップ上で可視化することができる. それに加えて, 成果を最大化するようなチーム $\hat{\tau}$ に対応する確率分布 $p(\mu|\hat{\tau})$ をメンバーマップ上で可視化することができる.

これらの機能を活用することで, ユーザによる操作とインターフェース上の可視化を繰り返し行うことができる. その一例が図 4.7 である. ユーザは操作に対する可視化から新し

^{*5} データに対応する潜在変数 μ_n が指定された場合は, その写像 $g(\mu_n)$ ではなくデータ x_n を表示することも可能である. これは生のデータ行列を見るよりも効率的にデータの探索を行うことができる.

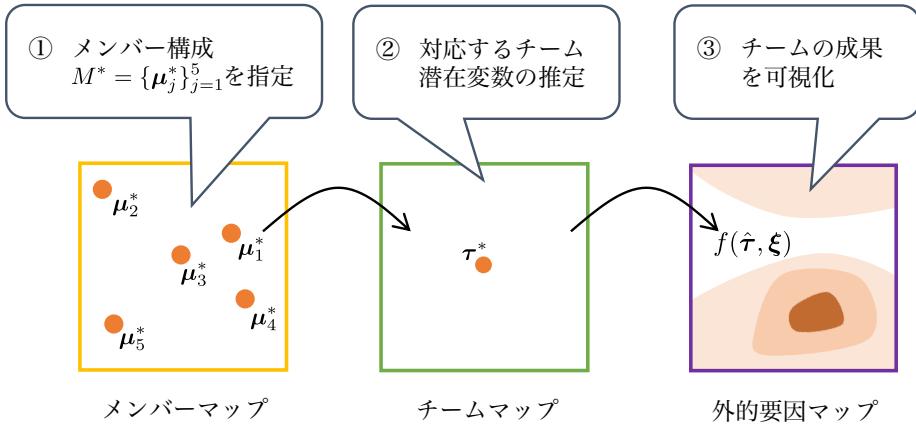


図 4.8 マップ上で実現できる新規チームのシミュレーション。起用するメンバーの潜在変数集合 $M^* = \{\mu_j^*\}_{j=1}^5$ をマップ上で指定したとする。対応する潜在変数 τ^* を推定することで成果を予測し可視化することができる。

い発見を得て、また次の可視化のための操作を行うことを繰り返すことでユーザ自身がデータとモデルを探査し、仮説検証のループと知識獲得のループを回すことができる。

このよう提案手法によって構築された VA システムの大きな特徴は、既存データについての情報の提示と未知データについての予測結果の提示の切り替えがシームレスに行われる点である。前者は既存データの潜在変数、後者はそれ以外のマップ上の座標への TOI として指定がトリガとなって行われる。

4.5.3 新規チームのシミュレーション

MNM を利用すると、マップ上で新規チームの試行錯誤を行うことができる「テーブルトップ」を提供できる。その手順を示したのが図 4.8 である。これは メンバーマップ上で起用するメンバーの潜在変数集合 $M^* = \{\mu_j^*\}_{j=1}^5$ を選択すれば、チーム潜在変数 τ^* をカーネル密度推定と確率分布集合に対する UKR を用いることで推定できる。これに伴って外的要因マップ上で $f(\tau^*, \xi)$ を可視化することができる。これによりユーザは新規チームを作り、それがどの外的要因に対して有効かを把握し、また別のチームを作りその結果を見るという試行錯誤を GUI 上で行うことが可能である。

4.6 提案手法が要件を満たすことの確認

提案手法についてここまで説明を踏まえて、本節では提案手法が第 3 章で示した要件を満たすことを確認する。ここでは最もシンプルな状況として、チームのラインナップを表現する集合データ X とその成果を表現する多変量データ y が与えられていることを

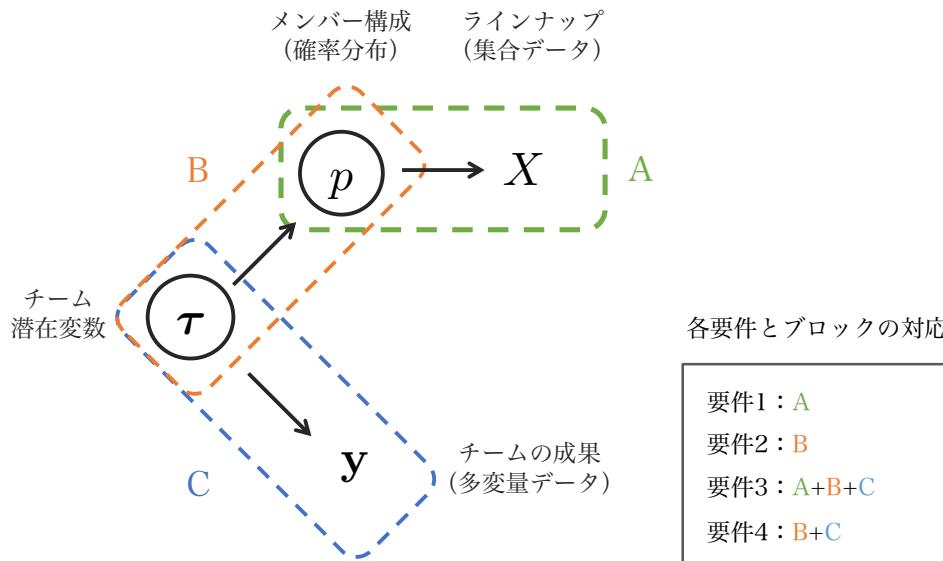


図 4.9 提案手法で構築した生成モデルと各要件の対応を示したもの。ここでは図 4.2(b) で示したラインナップ X と成果 y が与えられる状況における生成モデルを題材としている。ここで観測されない変数については円で囲うこととしている。A, B, C のブロックはそれぞれ 1 対の生成元の変数と生成先の変数のペアである。

想定する。このときに提案手法で仮定する生成モデルは図 4.2(b) で示した生成モデルである。この生成モデルを題材に、モデルのどの部分がどの要件を満たすことに対応しているのかを示したのが図 4.9 である。この図を参照しながら各要件について確認していく。

第一の要件である「集合データを扱うことができる」は、図 4.9 中の A で示すメンバー構成 p からラインナップ X が生成されるブロックによって満たされる。要件 1 をより具体的に言えば、類似度・距離の定義が自明でない集合データから、距離や類似度が定義できる表現を推定でき、なおかつその表現から集合データを生成することができる。提案手法ではラインナップ X の各要素が確率的に生成されるとして、その生成を制御するメンバー構成 p を X から推定する。このメンバー構成 p は確率分布^{*6}であり、KL ダイバージェンスのような距離^{*7}を定義することができる。また任意のメンバー構成 p からラインナップ X を生成することも可能である。

第二の要件である「ドメイン内可視化とドメイン間可視化ができる」については、B で示したチーム潜在変数 τ からメンバー構成 p が生成されるブロックにて実現される。ドメイン内可視化については、図 4.5 で示したように、チーム潜在変数をマップとして可視化することで、チームというドメイン内の構造の可視化を実現している。またドメイン間

^{*6} 図 4.3 などでも述べているが、メンバー構成 p は集合データのメンバー特徴量 x の空間上の確率分布ではなく、その特徴量 x の潜在変数 z の空間上の確率分布であることに注意。 p によって潜在変数 z が生成され、 z が特徴量の空間に写像されて x が生成される。

^{*7} 厳密には擬距離である

可視化については、図 4.6(b) に示したように、任意のチーム潜在変数を TOI として示すことで、そのチームがどのようなメンバー構成をとっているかメンバーMapper上で把握することができる。これが「チーム」ドメインと「メンバー」ドメインの関係性の可視化である。これは B ブロックでチーム潜在変数からメンバー構成が生成できることによって実現される可視化である。なお、図 4.6(a) のように任意のメンバーを TOI として指定して、どのようなチームに起用されやすいかチームMapper上で可視化するという逆方向のドメイン間可視化も可能である。

第三の要件である「集合データとその出力の間で双方向な予測ができる」については、生成モデル全体で満たされる。各ブロックで示す矢印はデータの「生成」を示しているが、生成先となるデータが与えられている場合はその生成元の変数を「推論」することができる。この「生成」と「推論」を連鎖させることで、集合データであるラインナップ X から成果 y を予測することも、その逆も可能である。

第四の要件である「ユーザの組合せ最適化を容易にする」については B と C で満たされる。B ブロックはユーザに対して、チームMapperを起点とする各チームがどのようなメンバー構成となっているかの探索を提供できる。一方で C ブロックはユーザに対して、チームMapperを起点とする各チームがどのような成果を挙げるかの探索を提供できる。つまり B と C を組み合わせることで、チームのメンバー構成と期待される成果を单一のチームMapperによって両方同時に探索することができる。これによって組合せ最適化を容易に行うことができる。例えば図 4.7 の②と③のように、ある特定の成果の項目について高いと予測されるチームがチームMapper上で色付けによって示され、特に高い成果をあげると思わしきチームをMapper上で指定することで、そのメンバー構成を見ることができる。そのメンバー構成を元に、ユーザはメンバーを選択することができる。より詳しい組合せ最適化の過程については、第 5 章のデモンストレーションでも詳しく示している。

最後に第五の要件である「システム構築手法が個々の応用例に柔軟に適応できること(可塑性)」についても満たしていると言える。なぜなら、図 4.2 に示したように、チーム編成のような具体的な個々の状況に対して生成モデルを構築でき、また図 4.4 で示したように、任意の生成モデルをデータから学習できるためである。

以上より提案手法である生成的多様体ネットワークモデリングは、集合データ VA の構築手法に求められる 5 つの要件すべてを満たしている。これは本研究の調査の範囲では他にない成果であり、本論文の主要な貢献である。

第 5 章

デモンストレーション

本章では、提案手法によって構築された VA システムのデモンストレーションを行う。まず 5.1 節にて、デモの題材として用いたバスケットボールチームの対戦結果のデータの詳細と適用した前処理について説明する。次に 5.3 節にて、データをモデル化し構築した VA システムについて概要を説明する。また 5.4 節では、そのシステム上で対話的な分析の過程をデモンストレーションする。そして 5.5 節で、デモの結果から第 3 章で示した要件を満たしていることを確認する。なお 5.6 節では、システムから得られた結果の妥当性を検討するために予測タスクと生成タスクの定量評価の結果も示している。

5.1 データセットと前処理

今回デモンストレーションのために用いたデータは National Basketball Associations(NBA) の試合についてのデータで、2018-2019 シーズンに行われた 1228 試合の結果である。これはウェブサイト Basketball Reference^{*1} で公開されている。試合データのうちおよそ 9 割に当たる 1105 試合を訓練データとして、うち 123 試合はハイパーパラメータの調整用のデータとしている。

選手に関するデータの前処理について述べる。まず出場試合数が 5 未満である選手を除外した。その結果、選手数は $N^{(m)} = 492$ 人となった。試合データには試合ごとの各選手のスタッツ（項目は表 5.1 参照）が与えられるが、これを次のように前処理することで選手特徴量とした。まず各試合ごとに各選手のスタッツをその選手の出場時間で除算し単位出場時間当たりのスタッツに変換した（ただし 3P%， 2P%， FT% のような割合を表現するスタッツは除く）。その上で各選手ごとに出場している試合全てでスタッツの平均を取った。その上で全ての選手のスタッツ集合の平均が 0、分散が 1 となるように標準化を

^{*1} <https://www.basketball-reference.com/>

表 5.1 NBA データセットのスタッツ一覧

選手のスタッツ		チームのスタッツ	
2PA	2 ポイントシュート試投数	2P	2 ポイントシュート成功数
2P%	2 ポイントシュート成功割合	3P	3 ポイントシュート成功数
3PA	3 ポイントシュート試投数	FT	フリースロー成功数
3P%	3 ポイントシュート成功割合	ORB	オフェンスリバウンド成功数
FPA	フリースロー試投数	DRB	ディフェンスリバウンド成功数
FT%	フリースロー成功割合	AST	アシスト成功回数
AST	アシスト回数	STL	スティール成功回数
ORB	オフェンスリバウンド数	BLK	ブロック成功回数
DRB	ディフェンスリバウンド数	TOV	ターンオーバー回数
STL	スティール成功数	PF	ファール回数
BLK	ブロック数	PTS	総得点
NTOV	ターンオーバー数 (符号反転)	PTD	得失点差
NPF	ファール数 (符号反転)		

行っている。その後、外れ値（ごく短い時間しか起用されていない選手が偶然にゴールを決めた時などに極端な値になりやすい）の影響を排除するために、出場時間が短いほど全選手の平均に近づくような正則化を加えている。これは事前分布として全アスリートの平均を中心とするようなガウス分布を与えた最大事後確率推定によって実現している。これにより全選手の特徴量 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{N^{(m)}})^\top$ が得られる。

今回の実験では状況を単純化し、492人のNBA選手全体を1つの起用可能な選手の集合であると見なし、その中から選手が選択されチーム編成が行われるとしている。従って、クラブ（ボストンセルティックスやブルックリンネッツのような一般に「チーム」と呼ばれるもの）の情報は学習には用いない。あくまでここで「チーム」として注目しているのは試合に出場した選手の集合（ラインナップ）のことである。チーム単位でのデータセットとして得られているのは、まず訓練データ 1105 試合分、 $N^{(t)} = 2210$ 個のラインナップ $\{t_i\}_{i=1}^{N^{(t)}}, t_i = \{\omega_{ij}\}_{j=1}^{n_i}, \omega_{ij} \in \{1, \dots, N^{(m)}\}$ である。このラインナップを元に特徴量集合の集合 $\{X_i\}, X_i = \{\mathbf{x}_{\omega_{ij}}\}_{j=1}^{n_i}$ が得られる。また単純な起用の有無だけではなく、 i 番目のラインナップにおける j 番目の選手の起用時間 w_{ij} も得られている。これをまとめて $\mathbf{W} = (w_{ij})$ とする。また各チームが獲得した成果を表現するデータとして、チーム単位でのスタッツ $Y = \{\mathbf{y}_i\}_{i=1}^{N^{(t)}}, \mathbf{y}_i \in \mathcal{O}^{(o)}$ が与えられている（項目は表 5.1 参照）。これについては前処理として標準化（データセット全体で平均 0 分散 1 になる規格化）を行

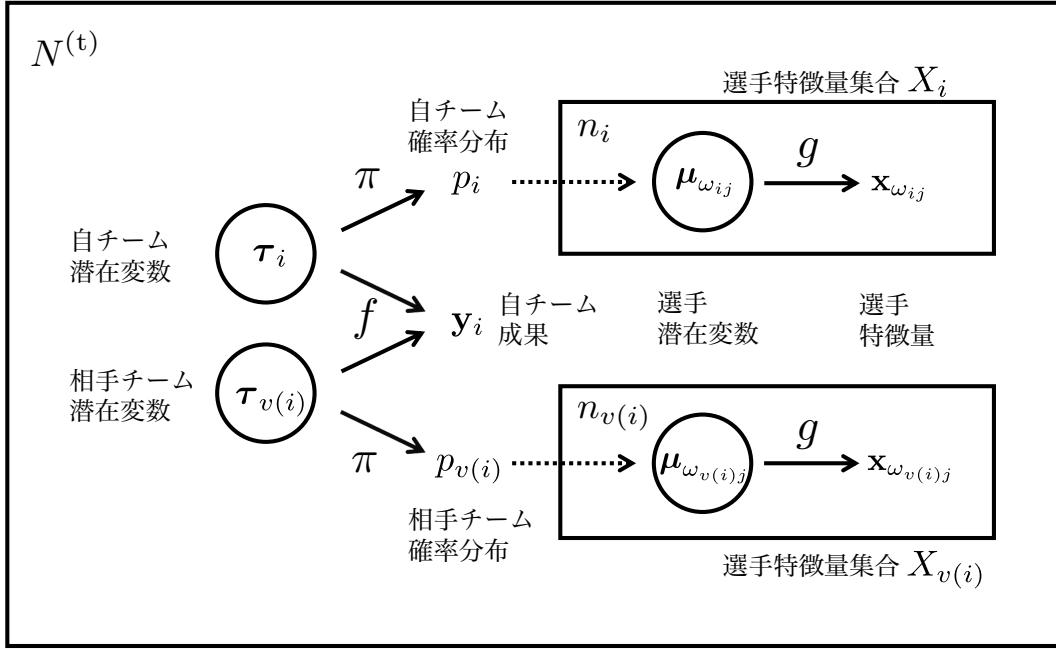


図 5.1 NBA データセットに対して仮定する生成モデル。

なっている。またこの成果が得られたときに対戦した相手チームはどのラインナップなのか $\{v(i)\}_{i=1}^{N^{(t)}}, v(i) \in \{1, \dots, N^{(t)}\}$ として与えられている。

5.2 モデルの学習

今回の NBA データをモデル化するにあたって仮定した生成モデルはチーム同士の対戦の状況における図 4.2(d) である。さらに詳細にデータセット全体の生成モデルを示したものが図 5.1 である。推定対象は選手潜在変数集合 $\mathbf{M} = \{\boldsymbol{\mu}_j\}_{j=1}^{N^{(m)}}$ 、チーム潜在変数集合 $\mathbf{T} = \{\boldsymbol{\tau}_i\}_{i=1}^{N^{(t)}}$ 、写像 $g : \mathcal{L}^{(m)} \rightarrow \mathcal{O}^{(m)}, \pi : \mathcal{L}^{(t)} \rightarrow \mathcal{P}, f : \mathcal{L}^{(t)} \times \mathcal{L}^{(t)} \rightarrow \mathcal{O}^{(o)}$ である。以下、学習に適用する手法やそのハイパーパラメータといった学習のステップの詳細について述べる。なお、これらの実装は GitHub^{*2}で公開している。

UKR による選手潜在変数と写像の推定

選手潜在変数集合 $\mathbf{M} = (\boldsymbol{\mu}_j)_{j=1}^{N^{(m)}}$ と写像 g の推定を UKR にて行う。4.4.1 節で述べたように、写像 g は式 (4.4) によって潜在変数 \mathbf{M} でパラメタライズされ、 \mathbf{M} は目的関数である (4.7) を最適化することで推定される。ここで $\forall j \boldsymbol{\mu}_j \in \mathcal{L}^{(m)} = [-1, 1]^2$ とする。今回は可視化の利便性のために潜在空間をこのような製法領域としている。また、写像の滑らかさを定めるハイパーパラメータとしてガウスカーネルの近傍半径 σ を指定する必要

^{*2} <https://github.com/furukawa-laboratory/demo-visual-analytics-set-data>

があり、次の計算式で求める。

$$\sigma = \sqrt{\frac{CS}{\pi N}} \quad (5.1)$$

これは面積 S の有界な潜在空間に N 個の潜在変数が一様に分布しているときに、半径 σ の円に入るデータ数が C となるような σ を求める計算式になっている。今回は $C = 4$ とすることで、データ数に対して適応的に近傍半径を計算する。また、潜在変数の最適化は最急降下法にて行った。学習係数 η は 0.03 としている。潜在変数の初期値はメンバー特徴量を主成分分析で 2 次元まで次元削減し、それを σ よりも十分小さくなるようにスケーリングしている。潜在変数の初期値を小さい範囲で与えることで、最終的に病的な局所解に陥ることを避ける効果があることが経験的に知られている。

カーネル密度推定による選手潜在空間上の確率分布の推定

式 (4.13) で示した重み付きのカーネル密度推定を用いて、各チームのメンバー構成を表現する確率分布の集合 $P = \{p_i\}_{i=1}^{N^{(t)}}$ を推定する。ガウスカーネルの近傍半径は 0.2 としている。

GPLVM と UKR の同時学習によるチーム潜在変数の推定

まず、確率分布集合 P に対して UKR を適用することを考える。すると P に対する写像 π がチーム潜在変数 \mathbf{T} によって表現され、 \mathbf{T} の目的関数が

$$F^{(t)}(\mathbf{T} | P) = - \sum_i D_{KL} [p_i \| q_{\tau_i}] \quad (5.2)$$

$$= - \sum_i \int p_i(\boldsymbol{\mu}) \ln \frac{p_i(\boldsymbol{\mu})}{q(\boldsymbol{\mu} | \tau_i)} d\boldsymbol{\mu} \quad (5.3)$$

と得られる。この KL divergence の積分は区分求積法によって数値計算する。

また、成果 \mathbf{Y} に対するモデルの学習には GPLVM を適用すると、写像 f がチーム潜在変数 \mathbf{T} によって定まるガウス過程として与えられる。またこのとき \mathbf{T} の目的関数 $F^{(o)}(\mathbf{T} | \mathbf{Y})$ が式 (4.11) により与えられるが、4.4.2 節の (d) で示したように複数の潜在変数から成果という单一のデータが生成されるモデルを学習する必要があるため、グラム行列 $\mathbf{K} = (k_{ii'})$ の要素は $k_{ii'} = k(\tau_i, \tau_{i'}) k(\tau_{v(i)}, \tau_{v(i')}) + \beta^{-1} \delta_{ii'} \delta_{v(i)v(i')}$ とカーネルの積の形で定義する。ガウスカーネル k の近傍半径は $C = 3$ として (5.1) で指定している。ただしホワイトカーネル δ の係数 β^{-1} は小さくしすぎると数値計算上ゼロ除算を起こすため、ゼロ除算を起こさない範囲で最小の値として $1/0.72$ に指定している。

このようにチーム潜在変数 \mathbf{T} に対する 2 種の目的関数を導入した上で、4.4.2 節の (c) で示したように、2 つの目的関数の線形和を取ることで以下の統合された目的関数を定義

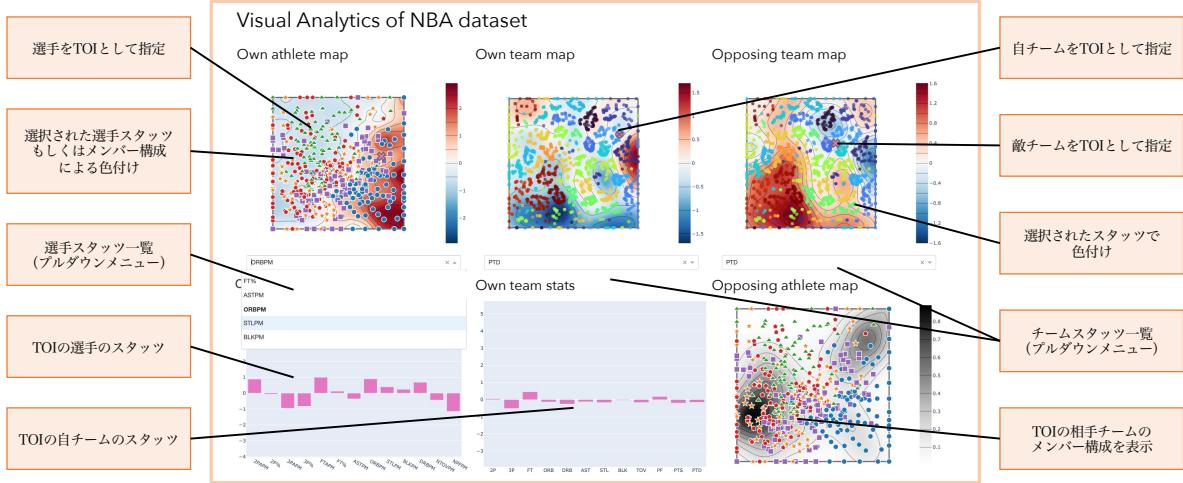


図 5.2 NBA データセットに対する VA システムのインターフェース。選手マップとチームマップが自チームと相手チームそれぞれに用意された計 4 つのマップと自チームの選手のスタッフと自チームそのもののスタッフを示す計 2 つの棒グラフによって構成されている。マップはそれぞれプルダウンメニューから指定したスタッフについて色付けされる。またマップ上では TOI を指定することができ、その TOI についての情報が他のマップの上で色付けにより可視化される。

する。

$$F(\mathbf{T}) = \alpha^{(t)} F^{(t)}(\mathbf{T}|P) + \alpha^{(o)} F^{(o)}(\mathbf{T}|\mathbf{Y}) \quad (5.4)$$

この目的関数を最適化する潜在変数 \mathbf{T} を最急降下法により推定する。今回はこの $\alpha^{(t)}, \alpha^{(o)}$ を直接与えるのではなく、それぞれの目的関数の勾配行列 $\frac{\partial F^{(t)}(\mathbf{T}|P)}{\partial \mathbf{T}}, \frac{\partial F^{(o)}(\mathbf{T}|\mathbf{Y})}{\partial \mathbf{T}}$ のフロベニウスノルムが等しくなるようにそれぞれ係数 $\lambda^{(t)}, \lambda^{(o)}$ をかけてスケーリングした上で、さらにそれぞれに $\alpha, 1 - \alpha$ かけることで更新のための勾配としている。なお、この α については、テストデータを用いてラインナップから成果の予測を行った結果、推定精度が最も良かった $\alpha = 0.65$ を採用している。

5.3 構築したシステム

NBA の試合の状況において、チームの成果は図 4.2 の生成モデルによって決まると仮定し、前述のデータに対してモデルの学習を行なった。これをベースとして開発したシステムのインターフェースを図 5.2 に示している。システムにおいて、潜在空間はマップのような正方領域として可視化される。NBA データセットの場合、システムは自チームと相手チームそれぞれについて選手マップとチームマップ、計 4 つのマップを表示する。自チームと相手チーム、それぞれ選手マップとチームマップの潜在変数の配置は同一であるが、システムはユーザの TOI に合わせてそれぞれのマップに対して色付けすることで情

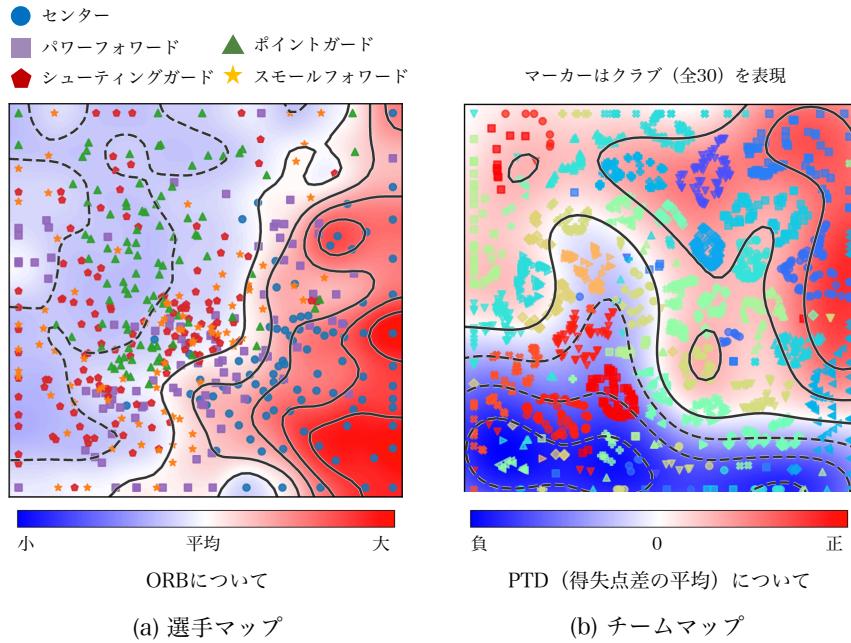


図 5.3 NBA データセットから得られた選手マップ (a) とチームマップ (b). (a) 選手マップ：点は選手を表す。また点のマーカーはポジションを表し、マップの着色はオフェンスリバウンド成功数 (ORB) の値の大きさを表す。全選手の平均の値であれば白、それより大きいほど赤く、小さいほど青い。(b) チームマップ：マーカーはクラブの情報を表している。マップの着色はそのチームが得られる得失点差 (PTD) の平均を表している。PTD の値が 0 であれば白く、正であれば赤く、負であれば青く表示している

報を可視化する。これらのマップに加えて、自チームの選手や自チーム自身のスタッフを表示する棒グラフが備わっている。またどのスタッフを選択するかはプルダウンメニューから選択することができる。

図 5.3 は得られた選手マップとチームマップを示したものである。選手マップにおいては、選手は固定されたランドマーク点として示されており、近くに配置された選手同士はスタッフが類似している。このマップから、大まかにポジションごとに選手が分かれていることが確認できる。同様にチームマップ上ではチームがランドマーク点として示されており、近くに配置されたチーム同士はスタッフとメンバー構成が類似するように配置されている。このチームマップを見ると、大まかにクラブごとにチームが分かれていることが確認できる。

一般的な地形図においては、標高や人口密度のような補足的な情報を等高線やヒートマップを用いて可視化することができる。これと同じように、システムのマップも特定のスタッフの値によって色付けすることができる。図 5.3 の (a) では、選手マップはオフェンスリバウンド成功数 (ORB) の値で色付けされている。もしユーザが他のスタッフ（例

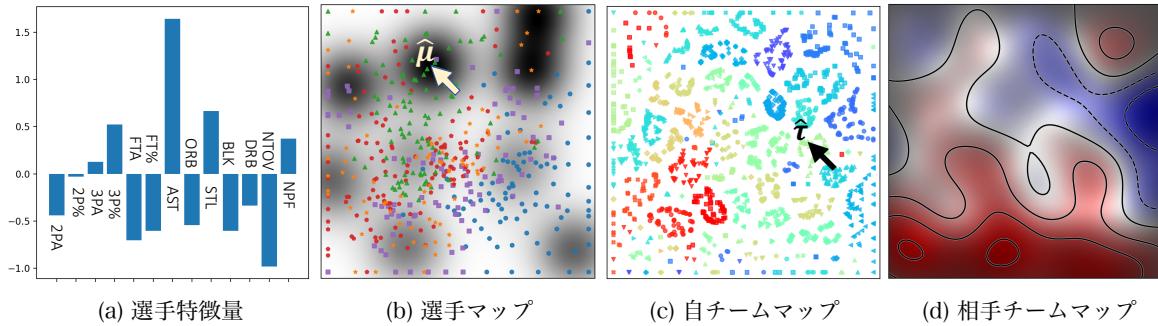


図 5.4 対話的な可視化によって TOI の伝達とそれに応じた可視化が連鎖していく一例. (c) 自チームマップ上の $\hat{\tau}$ を TOI として指定した時, (b) 選手マップ上でそのチームの確率分布 $p(\mu | \hat{\tau})$ が表示される. その確率密度が大きい点に着目し $\hat{\mu}$ を TOI として指定すると, (a) 棒グラフにその選手の特徴が可視化される. また $\hat{\tau}$ のチームが獲得する PTD の予測値が (d) 相手チームマップに表示される. 赤いほど $\hat{\tau}$ がその領域の相手チームに勝ちやすい. 色付けの明暗は予測の不確定性を表現している.

えば 2 ポイントシュート試投数など)について知りたい場合は、プルダウンメニューから所望のスタッツを選択すればよい。同様に、チームマップ(図 5.3(b))は平均的に得られると予測される得失点差(PTD)^{*3}によって色付けされており、すなわちこの色付けは勝つ見込みの大きさを示していると言える(赤いほど PTD が大きい、つまり大きな得点差をつけて勝利すると予測していることになる)。また、ユーザがプルダウンメニューから別のスタッツを選択した場合、システムはあらゆるチームのそのスタッツの傾向をマップ上で色付けによって示す。これによってユーザは各チームがどのような強みと弱みを持っているかを分析することができる。

5.4 デモンストレーション

図 5.4 は対話的な分析の例を示したものである。ここではユーザの TOI が自チームマップ(c)上の $\hat{\tau}$ であるとする。 $\hat{\tau}$ をユーザが TOI として指定すると、このとき $\hat{\tau}$ に対応するメンバー構成 $q(\mu|\hat{\tau})$ が人口密度のようにグレイスケールのヒートマップとして選手マップ(b)に表示される。これにより $\hat{\tau}$ というチームはどのようなメンバーで構成されているかを確認することができる。また選手マップの中で特定の $\hat{\mu}$ に注目したとすると、 $\hat{\mu}$ を TOI として指定すれば $\hat{\mu}$ に配置されている選手のスタッツ $g(\mu)$ を(a)の棒グラフ

^{*3} 厳密には次のように推定している。いま、潜在変数がそれぞれ $\tau_{\text{own}}, \tau_{\text{opp}}$ と表現されるような自チームと相手チームが対戦した時の自チームが得る得失点差 y をモデル化した関数 $\mathbf{y} = f(\tau_{\text{own}}, \tau_{\text{opp}})$ が学習によって得られている。マップ上の座標 τ とすると、その点の色付けに用いる $\bar{y}(\tau)$ を $\bar{y}(\tau) = \int f(\tau, \tau_{\text{opp}}) p(\tau_{\text{opp}}) d\tau_{\text{opp}}$ と推定する。これは即ち、 τ で表現されるチームが平均的にどの程度の PTD を得られるかを表している。

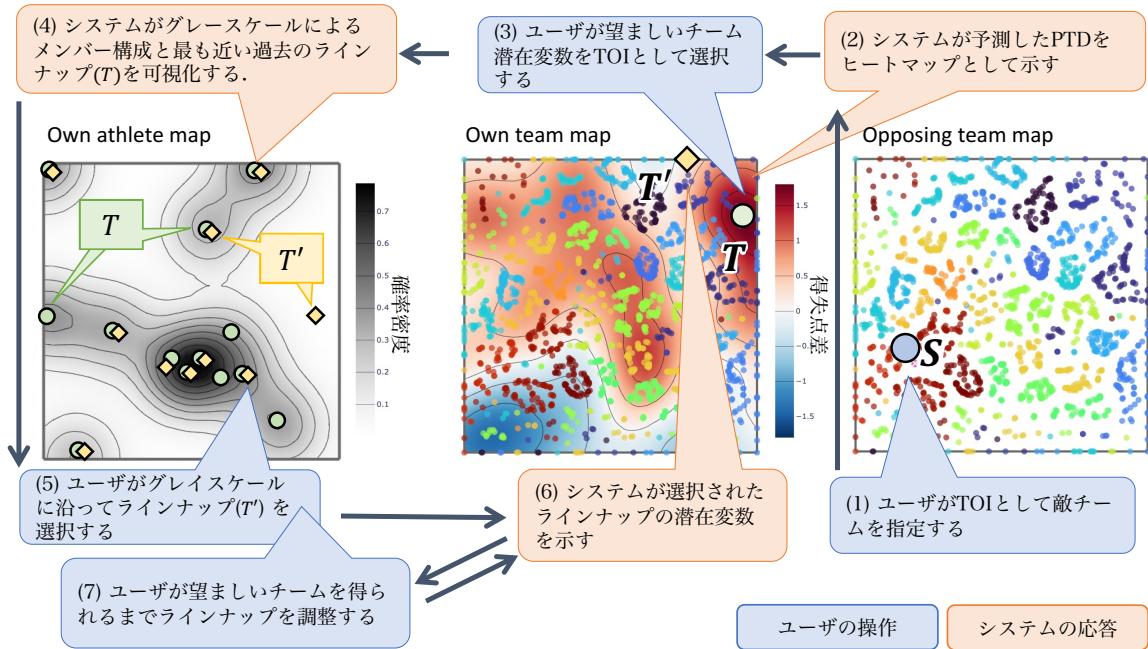


図 5.5 提案手法で構築した VA システムが支援するメンバー選択過程の例. 相手チームや理想的な自チームについて TOI として指定しつつ, 自チームの選手マップ上で実際に選手を配置する, その潜在変数を自チームマップ上で確認するといった操作を通してラインナップの試行錯誤を行う

ラフで確認することができる. その一方で相手チームマップ (d) では, 指定された自チーム $\hat{\tau}$ の PTD の予測関数 $f(\hat{\tau}, \tau_{\text{opp}})$ を τ_{opp} の関数と見なし, ヒートマップで可視化している. これはすなわち, チーム $\hat{\tau}$ が試合した時に勝てると予測される相手チームの領域は赤く, 負けると予測される相手チームの領域は青く表示される. 加えて, この関数は GPLVM でモデル化されているため, この予測値の不確定性も得られる. ここでは不確定性が大きいほど暗く, 小さいほど明るくなるように可視化している.

図 5.5 は VA システムによる支援の元でメンバー選択を行う過程の例を示したものである. このシナリオでは, 対戦する相手チームは相手チームマップにおける S に位置していると仮定している. ユーザが S を TOI として指定すると, システムが自チームマップ上で予測した PTD の値を可視化することで, どのチームであれば S に勝つことが期待できるかを示す. この可視化によってユーザはチームマップ上のすべてのチームについて予測結果を一眼で把握することができる. TOI として望ましいチーム (ここでは最も PTD の予測値が高かった T) を選択すると, システムは自チームの選手マップ上でメンバー構成をグレースケールで可視化する. さらに選手マップ上で高い密度の領域にある選手たちを選択すると, そのメンバー構成は大まかに望ましいメンバー構成と一致する. システムはまた T に近い過去のラインナップも可視化する. ラインナップをユーザが決定した後

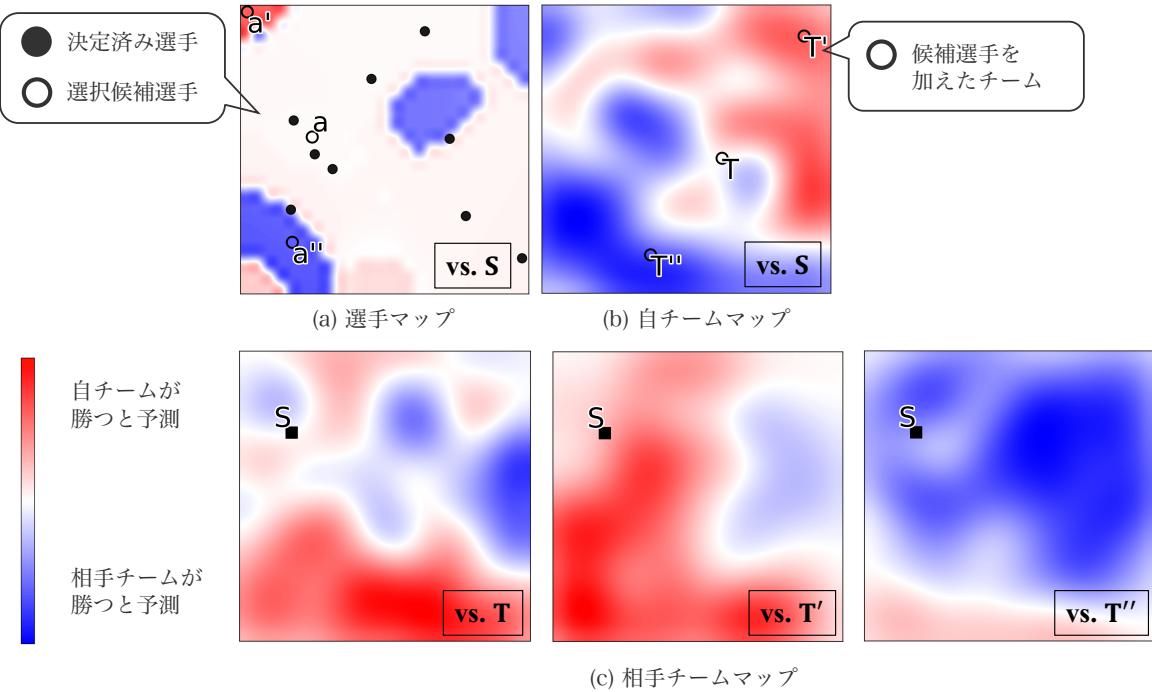


図 5.6 新規ラインナップのシミュレーションの例. 既に 9 人のメンバーは選択済みとして (●で示している), 10 人目のメンバーを選択することを考える. (a) 選手マップ. 10 人目に起用することでチームが (c) 相手チームマップ上の S に対して得られる PTD の予測値を色付けで示している. (b) 自チームマップ. 選手マップ上で示している候補メンバーを a, a', a'' (○で示している) を選択したときのチーム潜在変数をそれぞれ T, T', T'' で示している. 色付けは相手チーム S に対して獲得できる PTD の予測値である. (c) 相手チームマップ. T, T', T'' をそれぞれ自チームとして選択したときの獲得できる PTD の予測値を可視化している.

は, システムはそのラインナップが自チームマップ上でどこに位置するか示す (T'). もし必要であれば, ユーザは自チームマップ上でラインナップの位置を観察しながら, 試行錯誤しながらラインナップを調整することができる. またこのような過程の中で, ユーザは自身の経験的な知識といったデータには含まれない他の要因を考慮に入れることも可能である. この例が示すように, システムはチームマップ上の任意の点を対応する選手マップ上のメンバー構成に変換することができ, また逆も然りである. このようなチーム潜在変数とメンバー構成の間の双方向な対応づけは, メンバー選択の支援において本質的な役割を担っている. この点が提案手法の強みである.

図 5.6 はまた別のシナリオとして, 相手チームは相手チームマップ (c) 上の S であることが確定しているときに, すでに 9 人のメンバーを選択し (選手マップ (a) 上に黒いマークで示している), 新たに 10 人目のメンバーを選択しようとしている状況を示している. この状況において, (a) のヒートマップは, μ に位置するメンバーを起用した時に決

まる PTD の予測値を示したものである。つまり起用すると勝利するメンバーの領域は赤く、逆に負けてしまうメンバーの領域は青く示される。ここで代表的な座標として a, a', a'' を選手マップ上で挙げた。これにそれぞれ対応するチームが自チームマップ (b) 上の T, T', T'' である。自チームマップはまた、予測した PTD の値をヒートマップとして示している。これによると T' は勝利し、 T'' は負けると予測している。また図 5.6(c) は T, T', T'' がそれぞれ様々な相手チームと戦った時の PTD の予測値を示している。これによると T' は全体的に赤い領域が広がっているため、あらゆるタイプのチームに勝つことができるが、 T'' は青い領域が広がっているためあらゆるタイプのチームに負けてしまうと予測していることがわかる。このシナリオにおいて、ユーザのタスクは与えられたチームに対して新しい候補者を選ぶことである。VA システムはまた、これとは逆のケース、つまり新しく加わったメンバーに対して適したチームを決定するタスクにも適用することができる。すなわち VA システムは既存のチームに対して、新しいメンバーを選択しなければならないようなマネージャをサポートすることも可能である。

5.5 システムが要件を満たすことの確認

ここまでデモンストレーションを通して対話的な分析過程の例をいくつか挙げたが^{*4}、これらを踏まえてシステムが第3章で定義した要件を満たしていることを確認する。

第一の要件である「集合データを扱える」という点に関しては、図 5.5 のように、システムが集合データであるラインナップからメンバー構成を獲得できていることから満たしていると言える。ここでのメンバー構成とはメンバー潜在空間上の確率分布であり、これは集合データの要素の数や順序に依存しない表現になっている。また必要であればメンバー構成からラインナップを生成することも可能である。図 5.5 で示しているように、基本的にはユーザ自身がメンバー・マップ上のメンバー構成を見ながら潜在変数を自ら指定することで想定しているが、必要であればメンバー選択を自動的に行うことも可能である。これは選択可能なメンバーの潜在変数集合の部分集合によって構築される経験分布が、メンバー確率分布に近くなるように部分集合を選択するという組合せ最適化問題を解くことで自動的にメンバー選択が可能である。

第二の要件である「ドメイン内可視化とドメイン間可視化の実現できる」という点に関しては、図 5.4 や図 5.5 のように、ユーザがそれぞれのマップに対して自身の TOI を示

^{*4} ここまで示しているシステムとのインタラクションの例は、システムが提供できるインタラクションのごく一部である。その他の例については、<https://github.com/furukawa-laboratory/demo-visual-analytics-set-data> で公開しているシステムのソースコードをローカル環境で実行することで参照可能である。

すことで、注目した要因についての可視化結果をシステムが提示していることから満たしていると言える。これが実現できる理由は、モデルが絡み合った複数の要因を独立した潜在変数としてそれぞれ推定し、なおかつその潜在変数を元にデータを生成するための関数も推定しているためである。

第三の要件である「集合データとその出力の間で双方向な予測ができる」という点も満たせている。「集合データから出力の予測」「出力から集合データの予測」の2つに分けて確認してみると、まず「集合データから出力の予測」については、図5.5に示したように、与えられた任意のラインナップに対してそのスタッツが予測できていることから満たせていることが分かる。次に「出力から集合データの予測」については、今回のデモの例では「所望のスタッツを得られるようなメンバーのラインナップを得ること」に相当するが、それは次の3つのステップで実現できる。(1) 所望のスタッツを満たすチームを自チームマップ上で発見する。これはユーザが特定のスタッツを TOI として指定すると、システムが自チームマップ上をその値の大小で着色する機能(例: 図5.2)を利用することで可能となる。(2) 自チームマップ上のチームのメンバー構成を把握する。発見したチームを TOI として指定することでメンバー構成をメンバー マップ上での密度表示によって把握することができる(例: 図5.4)。(3) メンバー構成からラインナップを決定する。つまりメンバー マップ上での確率分布を元に、実際に起用するメンバーを選択することができる。手順については第一の要件の確認の際に示した通りである。

もしユーザが既存のチームやメンバーを TOI として選択した場合はシステムは過去のデータからの分析結果を示し、過去に存在しないチームやユーザを TOI とした場合は予測した結果を示す。

第四の要件である「ユーザが組合せ最適化を容易に行える」という点に関しては、システムがチームの集合についてのマップを示しており、ユーザがあらゆるチームを概観できるという点を以って満たしていると判断できる。ユーザが特定のスタッツおよび特定の相手チームを TOI として指定することで、あらゆるチームがどのような傾向を持つかを一眼で把握することができる。このマップ上でのインタラクションを用いて具体的にどうメンバー選択を行うかは図5.5に示した通りである。

なお、本論文の目的はあくまで今回デモをしたVAシステムを開発することではなく、集合データVAシステムを構築するための汎用的な手法を開発することである。第3章では、提案手法がシステムの開発者に対し柔軟な拡張や個々のケースに対する適応を可能としなければならないという「可塑性」要件を定義し、実際に提案手法がそれを満たすことを第4章で示したが、このような可塑性が提案手法の利点である。

そして、集合データを扱う意思決定問題に対して、提案手法はVAの枠組みによるデータ分析・仮説検証・意思決定を行うことが、即ち、ユーザは複雑で巨大なデータを対話的に探索することができる。また最終的な判断はユーザに委ねられているため、ユーザ自身

の経験的な知識も考慮に入れて決定をすることができる。これは同時に決定後の説明責任 (Accountability) を満たすことに寄与するであろう。

5.6 予測タスクと生成タスクの定量評価

提案システムが提示する情報、特に任意のチームに対する成果の予測結果（集合データに対する識別タスクの結果）や特定の成果を満たすチームの予測結果（生成タスクの結果）について、どれだけ妥当なものかを検証することは実際の意思決定支援の観点から重要である。システムが提示する予測結果ができるだけ精度の高いものであることが望ましいが、ここまでデモンストレーションから分かるように、システムは集合データに対するトレードオフの関係にある識別・生成・可視化の3つのタスクを同時に解いているため、特に可視化をするという制約によって他の識別や生成タスクの性能が損なわれていることが考えられる。

そこで本節ではシステムが提示する情報の妥当性を検討するための材料として、識別タスクと生成タスクそれぞれの定量的な評価の結果について示す。またベースラインやベンチマークとしての他のアプローチの評価結果についても併記する。

まず識別タスクについて定量評価を行うために、試合結果（勝敗）の予測において精度 (Accuracy) の評価を次のように行った。まず、ベースラインを 50% (チャンスレート) に設定する。加えて、比較用のベンチマークシステムを次のように設計した。各ラインナップを特徴ベクトルに変換するため、メンバーの特徴量の平均を用いた。次にこの特徴ベクトルを2次元に変換するために PCA を用いた。この2次元ベクトルは我々の手法におけるチーム潜在変数 τ に対応する。最後に自チームと相手チームの2次元ベクトルのペアから自チームを予測する関数をガウス過程回帰により推定した。なおここで注意すべき点として、このベンチマークシステムは VA システムの要件をすべて満たしているわけではないということを挙げておく。このシステムでは、集合データに対して特徴量について平均を取るという素朴な処理を行なっているため、生成タスクを解くことができない。あくまでこのシステムを設計する上で念頭に置かれていることは、モデルに対する可視化に由来する制約（チーム特徴量を2次元の数ベクトルにすること）を守った上で最も素朴なアプローチで予測タスクを解くということである。

精度評価の結果は、ベースラインが 50%，ベンチマークシステムが 56% であるのに対して、提案手法で構築したシステムは 63% であった。このことから可視化のための制約を守った素朴なシステムより提案手法で構築されたシステムの方が精度の面で優れていることが分かる。また参考までに、NBA の試合結果の予測のみを目的とした研究群 [107] の予測精度は 60% から 80% の間に分布しており、提案手法の予測精度はこの範囲の中に収まっている。ただし、これらの異なるデータセット、異なるタスク制約のもとで評価さ

れているため、公平な比較は難しい。

また生成タスクについても定量評価を行った。ここではモデルがどれだけ正確に元のメンバー構成を再構成できたかを測る指標として、ハイパーパラメータ選択に用いたデータセットに含まれる全てのラインナップをテストメンバー構成 $p_i^{\text{test}}(\boldsymbol{\mu})$ として、モデルによって再構成されたメンバー構成 $q(\boldsymbol{\mu}|\tau_i^{\text{test}})$ の間の イエンセン・シャノン (Jensen-Shannon: JS) ダイバージェンスを評価し、その平均を求めた。その結果、ベースライン（一様分布）が 0.24 であったのに対して、提案手法は 0.012 であった。また比較のために、テストメンバー構成とその平均のメンバー構成 $\bar{p}(\boldsymbol{\mu}), \bar{p}(\boldsymbol{\mu}) = \frac{1}{N^{(\text{t})}} \sum_i p_i^{\text{training}}(\boldsymbol{\mu})$ の間の JS ダイバージェンスを測定したところ、その平均は 0.14 であった。これは直感的には、提案システムが様々なラインナップの多様性を保持した潜在変数を推定し、その潜在変数を元にメンバー構成が再構成されていることを示している。なお、ベンチマークシステムを含めて集合データの集合を考慮しない手法においては、今回のように JS ダイバージェンスなどによってメンバー構成の再構成の度合いを評価することは不可能である。

第 6 章

議論

6.1 提案手法の評価と 5 つの要件

一般に、VA システムの評価というのは困難であり、確立された手続きというのは存在しないと言える [22]. ある応用に特化した VA システムの場合、その多くはケーススタディやユーザスタディおよび専門家によるレビューによって評価される [99, 108]. 視覚的インターフェースを開発が目的である場合は、被験者実験による操作性を評価することが多い [109, 110]. また分析のパフォーマンスを向上させることが目的である場合は、システムは他のシステムと比較されるべきである. それに対して、VA を実現するための理論的な問題を解決することを目的としている研究であれば、手法は実験的に評価するのではなく、問題を解決できているかを論理的に評価するべきである. 本研究は最後のケースであるため、評価基準として第 3 章で示した 5 つの要件を定めている.

だがしかし、これらの要件の妥当性について懸念があるかもしれない. ここではこれらの要件が必要条件であることを示すために、もし要件のうち 1 つもしくは複数が満たされていなかった場合にどうなるか述べる. 結論から述べればどれか 1 つでも要件が欠ければ、集合データの VA としてシステムが機能しない、もしくはその機能が著しく制限されることは明白である. もし 1 つ目の要件もしくは 4 つ目の要件が満たされていなければ、システムは集合データの VA システムとして利用できない. もし 2 つ目の要件が達成されていなければ、システムは与えられたデータの限られた側面しか可視化することしかできない. 3 つ目の要件が達成されていなければ、システムを未来のケースについての意思決定支援という用途に利用できない. また、個々の応用先はそのドメイン特有の追加的な要件を要求することがしばしば生じるため、これに対応するためには 5 つ目の要件が重要である.

本研究においては、集合データを取り扱うことを考慮しながら理論的な課題を解決することに焦点を当てたが、他の評価基準を用いて構築された VA システムそのものの評価を

行うことは今後の課題である。特に、人間駆動という観点から構築された VA システムの操作性や利便性を評価することは、VA の領域において重要な課題である。

6.2 提案手法と他の学習パラダイムとの関連

MNM およびそれをデータから獲得するための GMM の結合による学習は、機械学習のいくつかの学習パラダイムと関連がある。

まずチームとメンバーの間の階層的なモデル化はマルチレベル分析 [111] と呼ばれるパラダイムに該当する。マルチレベル分析では、異なる集団からそれぞれの個体についてのデータが複数与えられるような状況で、各集団についての傾向や集団間の差異の分析を行う。提案手法ではメンバーの潜在変数によるメンバー全体の分析やチーム潜在変数によるチーム集合全体の分析、さらにメンバー潜在空間上の確率分布によるチーム間の差異の分析を提供しており、マルチレベル分析を実現していると言える。

次に複数の GMM を同時学習することで单一潜在空間と複数の多様体モデルを対応させるタスクは、マルチビュー学習 [112] の一種であると見なせる。マルチビュー学習は単一の分析対象に対して複数の情報源（ビュー）からデータが得られるようなデータセットに対する問題設定の総称である。その中でも提案手法は、複数の情報源に共通する潜在変数を推定しているため、マルチビュー表現学習 [113] に該当する。

また図 4.2(c) や (d) の例にあるような成果のモデリングの部分は、関係データ分析 [114] であると見なせる。通常の多変量データの場合、1 つのデータは 1 つの分析対象から得られたものあることを仮定している。これに対して関係データとは 1 つのデータは複数の分析対象の組合せから観測されたデータである。第 5 章で取り扱った NBA データも、成果を表現するチームのスタッフはそのチームと敵のチームという複数の分析対象から得られたデータであるため関係データと見なすことができる^{*1}。多変量データがしばしば行列（2 次元配列）として表現されるのに対して、関係データはしばしばマルチウェイデータ [115] もしくはテンソル [116] と呼ばれる多次元配列として表現される。そのため関係データの分析には、行列分解をテンソルに拡張したテンソル分解が用いられる。テンソル分解は各分析対象に対して潜在変数と複数の潜在変数を入力とする関数を仮定することで観測データの生成過程をモデル化する。線形手法としては Tucker 分解 [117] や CP 分解 [118] があり、非線形手法としては InfTucker [119] や NeuralCP [120], Tensor

^{*1} ただし、今回のデモではラインナップが異なれば違うチームと見なしており、同一のチームについてのデータが複数得られることはない。このようなデータは分析対象の間で類似度を定義できないため、テンソル分解を適用し潜在変数を得ることはできない。今回のデモの場合は、関係データ分析だけでなくマルチビュー表現学習のパラダイムも利用し、成果だけでなくメンバー構成についても統合することで潜在変数を推定している。

Self-Organizing Map (TSOM) [121] がある。この中でも TSOM はインタラクティブな可視化に特化している手法であり、提案手法の可視化方法は TSOM の可視化方法にインスパイアされている。

これらの学習パラダイムの先行研究からヒントを得ながら、モデルの実装の部分を改善することは今後の課題である。提案手法の MNM の枠組みは得られているデータセット全体の生成過程を包括的にモデル化するものであるが、それは即ち、MNM の実装の段階において、ここまで述べた学習タスクを独立に、もしくは協調させながら同時に解かなければならぬことを意味している。それぞれの学習タスクは未だ確立した手法というには存在せず、現在も研究が進められている領域である。本論文で示した提案手法の実装方法はあくまでも一例に過ぎず、それぞれの学習タスクの最新の学習テクニックを取り入れたり、同時学習に適したアプローチを検討したりすることは、データからより良い知識を発見できるモデルを獲得するために重要である。

6.3 提案手法の拡張

応用事例ごとにデータ構造や可視化したい情報は異なるため、実際の応用事例に提案手法を適用するためには可塑性は必要不可欠である。本手法のキーアイデアは生成過程を MNM を用いてモデル化することであり、MNM の構造をデータ構造に適応させることは容易である。また実装の面においては、MNM は GMM の組み合わせを用いて推定され、4 種の接続方法を一貫して用いる。したがって提案手法はアプリケーションの事例に対し柔軟に適応することができる。MNM は複数の多様体のベイジアンネットワークモデルであるとみることができ、この手法は general というよりも universal であると言えるだろう。本論文の範囲では実施していないが、ネットワーク構造を必要に応じて段階的に拡張できるソフトウェアライブラリを開発することは可能である。このようなオンデマンドの可塑性は、分析プロセスをよりダイナミックにすることが期待できる。

提案手法は、他のデータ分析手法を組み合わせた統合的な VA システムを開発することも可能にする。なぜならマップは様々な分析結果を可視化するプラットフォームとして利用することが可能だからである。例えば、PCA や因子分析の結果をマップ上で可視化することができる。この事例においては、ユーザは成分のリスト（例えばプルダウンメニューで表示される）から関心のある成分を TOI として選択するだけでよい。同様に他のブラックボックスシステムの出力もマップ上で可視化できる。

チーム編成支援では、バスケットボールにおけるポジションなど、メンバーの役割が与えられる場合もある。このような事例にも提案手法は容易に拡張できる。拡張のアプローチとしては次の 2 つが考えられる。1 つ目はチームのメンバーと役割の構成をメンバーの潜在変数と役割の同時確率 $q(\mu, r | \tau)$ で表現することである。ここで $r \in \{1, \dots, R\}$ は

全部で R 個ある役割を表現するカテゴリカルな変数である。2つ目は役割に対して連続潜在変数 $\{\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_R\}$ を導入し, $p(\mathbf{x} | \boldsymbol{\tau})$ を次のように表現する

$$p(\mathbf{x} | \boldsymbol{\tau}) = \int p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\rho}) p(\boldsymbol{\mu}, \boldsymbol{\rho} | \boldsymbol{\tau}) d\boldsymbol{\mu} d\boldsymbol{\rho} \quad (6.1)$$

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\rho}) = \mathcal{N}(\mathbf{x} | g(\boldsymbol{\mu}, \boldsymbol{\rho}), \beta^{-1} \mathbf{I}) \quad (6.2)$$

これは 2.2.5 節で紹介した e 型のモデル化と m 型のモデル化の組み合わせになっており, この場合の潜在変数 $\boldsymbol{\mu}$ は役割とは独立した個々のメンバーの性質を表現している。このように役割についても潜在変数を推定することで役割同士の類似関係をマップ上で把握することができる。また役割の数が多くなったときでも, $q(\boldsymbol{\mu}, \boldsymbol{\rho} | \boldsymbol{\tau})$ を可視化することでチームがどのような配分で役割を構成しているか, マップ上で直感的に把握できるという利点がある。この拡張はファンションアウトフィットに手法を適用する際にも有用であると考えられる。この場合, メンバーと役割はファンションアイテムとそのカテゴリー（例えばシャツ, パンツ, コートなど）に相当する [14].

時間の経過とともに新しいデータが逐次的に追加されても, オンライン学習を適用することでシステムを更新することができる。この事例においては, メンバーとチームはマップ上で固定されず, 時間経過とともにマップ上を移動することになる。例えば毎月選手のスタッフが得られれば, スタッフから得られる選手の特徴量とそれに対応する潜在変数はそれぞれ時系列 $\mathbf{x}_i(t), \boldsymbol{\mu}_i(t)$ となる。メンバーの潜在的なパフォーマンスを考慮したい場合には, この拡張は有用であると考えられる。

第 7 章

総括

本論文では集合データ VA の要件を満たすシステムの構築手法である、生成的多様体ネットワークモデリングを提案した。要件を満たすためには集合データに対して複数のタスクを同時に解かなければいけないという課題があり、本論文はその困難に対する最初の取り組みである。提案手法によって、集合データを含むさまざまなデータに対し、既存データ分析と新規パターンの予測を対話的かつシームレスに行える VA システムの構築が可能になった。つまり提案手法は、チーム編成やファッショナウトフィット選択に代表される複雑・重大な組合せ最適の場面において、人間自身による意思決定をデータ駆動で支援するための有効なアプローチであると言える。

提案手法の特筆すべき点はその汎用性の高さにある。提案手法ではデータセット全体の生成過程を多様体モデルのネットワーク構造で表現する。このネットワーク構造はデータ構造に合わせて柔軟に適応させることが可能である。即ち本研究はチーム編成支援のための特定のシステムを開発したにとどまらず、集合データを含むさまざまなデータに適応できる汎用的な VA システムの構築手法を提案したと言える。これは特定の場面に特化したシステムの開発が多い VA 領域においては稀有な試みであり、本論文が人間主導型データ駆動アプローチの新たな基盤の確立に寄与することを願っている。

参考文献

- [1] Hsinchun Chen, Roger H L Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, Vol. 36, No. 4, pp. 1165–1188, 2012.
- [2] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, Vol. 1, No. 1, pp. 51–59, 2013.
- [3] Dell Technologies. Digital transformation index 2020. Technical report, 2020.
- [4] Martina Barbero, Jo Coutuer, Regi Jackers, Karim Moueddene, Els Renders, Wim Stevens, Yves Toninato, Sebastian van der Peijl, and Dimitry Versteele. Big data analytics for policy making. Technical report, 2016.
- [5] Nada Elgendi and Ahmed Elragal. Big data analytics in support of the decision making process. *Procedia Computer Science*, Vol. 100, pp. 1071–1084, 2016.
- [6] David Gotz and David Borland. Data-driven healthcare: Challenges and opportunities for interactive visualization. *IEEE Computer Graphics and Applications*, Vol. 36, No. 3, pp. 90–96, 2016.
- [7] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.
- [8] Kar Yan Tam and Shuk Ying Ho. Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research*, Vol. 16, No. 3, pp. 271–291, 2005.
- [9] Michele Lombardi and Michela Milano. Boosting combinatorial problem modeling with machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5472–5478, 2018.
- [10] O C Ferrell and Linda Ferrell. The responsibility and accountability of CEOs : The last interview with ken lay. *Journal of Business Ethics*, Vol. 100, No. 2, pp. 209–219, 2011.
- [11] James J. Thomas and Kristing A. Cook. *Illuminating the path*. IEEE Computer

- Society Press, 2005.
- [12] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3391–3401. 2017.
 - [13] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97, pp. 3744–3753. PMLR, 2019.
 - [14] Yuncheng Li, LiangLiang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, Vol. 19, No. 8, pp. 1946–1955, 2017.
 - [15] Aharon Tziner and Dov Eden. Effects of crew composition on crew performance: Does the whole equal the sum of its parts? *Journal of Applied Psychology*, Vol. 70, No. 1, pp. 85–93, 1985.
 - [16] Bruno Travassos, Davids, Duarte Araujo, and Pedro Esteves. Performance analysis in team sports: Advances from an ecological dynamics approach. *International Journal of Performance Analysis in Sport*, Vol. 13, pp. 89–95, 2013.
 - [17] Duarte Araújo and Keith Davids. Team synergies in sport: Theory and measures. *Frontiers in Psychology*, Vol. 7, p. 1449, 2016.
 - [18] Julián Moreno, Demetrio A Ovalle, and Rosa M Vicari. A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. *Computers & Education*, Vol. 58, No. 1, pp. 560–569, 2012.
 - [19] Miguel Ángel Pérez-Toledano, Francisco J. Rodriguez, Javier García-Rubio, and Sergio José Ibañez. Players' selection for basketball teams, through performance index rating, using multiobjective evolutionary algorithms. *PLOS ONE*, Vol. 14, No. 9, p. e0221258, 2019.
 - [20] Fahimeh Rahmanniyay and Andrew Junfang Yu. A multi-objective stochastic programming model for project-oriented human-resource management optimization. *International Journal of Management Science and Engineering Management*, Vol. 14, No. 4, pp. 231–239, 2019.
 - [21] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20, No. 12, pp. 1604–1613, 2014.
 - [22] Wenqiang Cui. Visual analytics: A comprehensive overview. *IEEE Access*,

- Vol. 7, pp. 81555–81573, 2019.
- [23] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information Visualization. Lecture Notes in Computer Science*, Vol. 4950, pp. 154–175. 2008.
- [24] 伊藤貴之. 意思決定を助ける情報可視化技術. コロナ社, 2018.
- [25] Natalia Andrienko and Gennady Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, Vol. 12, No. 1, pp. 3–24, 2012.
- [26] Takanori Fujiwara, Jian Zhao, Francine Chen, and Kwan-Liu Ma. A visual analytics framework for contrastive network analysis. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 48–59, 2020.
- [27] Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmquist. Conceptvector: Text visual analytics via interactive lexicon building using word embedding. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 24, No. 1, pp. 361–370, 2018.
- [28] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. VitalITY: Promoting serendipitous discovery of academic literature with transformers & visual analytics. In *IEEE VIS*, 2021.
- [29] Ian T Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.
- [30] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374, No. 2065, p. 20150202, 2016.
- [31] Warren S Torgerson. *Theory and methods of scaling*. Wiley, 1958.
- [32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2625, 2008.
- [33] Kuno Kurzhals and Daniel Weiskopf. Space-time visual analytics of eye-tracking data for dynamic stimuli. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2129–2138, 2013.
- [34] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. ipca: An interactive system for pca-based visual analytics. *Computer Graphics Forum*, Vol. 28, No. 3, pp. 767–774, 2009.
- [35] A Endert, W Ribarsky, C Turkay, B L William Wong, I Nabney, I Díaz Blanco, and F Rossi. The state of the art in integrating machine learning into visual

- analytics. *Computer Graphics Forum*, Vol. 36, No. 8, pp. 458–486, 2017.
- [36] Yafeng Lu, Rolando Garcia, Brett Hansen, Michael Gleicher, and Ross Miejewski. The state-of-the-art in predictive visual analytics. *Computer Graphics Forum*, Vol. 36, No. 3, pp. 539–562, 2017.
- [37] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, Vol. 16, No. 3, pp. 31–57, 2018.
- [38] Iryna Korshunova, Jonas Degrave, Ferenc Huszar, Yarin Gal, Arthur Gretton, and Joni Dambre. Bruno: A deep recurrent model for exchangeable data. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 7190–7198, 2018.
- [39] Mengjiao Yang, Bo Dai, Hanjun Dai, and Dale Schuurmans. Energy-based processes for exchangeable data. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 10681–10692, 2020.
- [40] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [41] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2017.
- [42] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5099–5108, 2017.
- [43] Harrison Edwards and Amos Storkey. Towards a neural statistician. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
- [44] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [45] Yang Li, Christopher M. Bender, Haidong Yi, Siyuan Shan, and Junier B. Oliva. Exchangeable neural ODE for set modeling. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 6936–6946, 2020.
- [46] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 29–37, 2011.

- [47] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [48] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.
- [49] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, Vol. 17, pp. 1–40, 2016.
- [50] Barnabas Poczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 507–515, 2013.
- [51] Konstantinos Skianis, Giannis Nikolenzos, Stratis Limnios, and Michalis Vazirgiannis. Rep the set: Neural networks for learning set representations. In *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1410–1420, 2020.
- [52] Grégoire Mialon, Dexiong Chen, Alexandre d’Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [53] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [54] Siheng Chen, Chaojing Duan, Yaoqing Yang, Duanshun Li, Chen Feng, and Dong Tian. Deep unsupervised learning of 3D point clouds via graph topology inference and filtering. *IEEE Transactions on Image Processing*, Vol. 29, pp. 3183–3198, 2020.
- [55] Isaak Lim, Moritz Ibing, and Leif Kobbelt. A convolutional decoder for point clouds using adaptive instance normalization. *Computer Graphics Forum*, Vol. 38, No. 5, 2019.
- [56] Chun Liang Li, Manzil Zaheer, Yang Zhang, Barnabás Póczos, and Ruslan Salakhutdinov. Point cloud gan. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

- [57] Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zieba, and Tomasz Trzcinski. Hypernetwork approach to generating point clouds. In *Proceedings of the 3rd conference on Machine Learning and Systems (MLSys)*, pp. 6004–6013, 2020.
- [58] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [59] Michał Stypulkowski, Maciej Zamorski, Maciej Zięba, and Jan Chorowski. Conditional invertible flow for point cloud generation. In *NeurIPS Workshop Sets & Partitions*, 2019.
- [60] Bruno de Finetti. La prévision: ses lois logiques ses sources subjectives. *Annales de l'institut Henri Poincaré*, Vol. 7, pp. 1–68, 1937.
- [61] D. K. Kendall. *On Finite and Infinite Sequences of Exchangeable Events*. Defense Technical Information Center, 1967.
- [62] Persi Diaconis. Finite forms of de finetti’s theorem on exchangeability. *Synthese*, Vol. 36, No. 2, pp. 271–281, 1977.
- [63] P Diaconis and D Freedman. Finite exchangeable sequences. *Annals of Probability*, Vol. 8, No. 4, pp. 745–764, 1980.
- [64] G. Jay Kerns and Gábor J Székely. Definetti’s theorem for abstract finite exchangeable sequences. *Journal of Theoretical Probability*, Vol. 19, No. 3, pp. 589–608, 2006.
- [65] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [66] Hideaki Ishibashi, Masayoshi Era, and Tetsuo Furukawa. Hierarchical tensor manifold modeling for multi-group analysis. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E101A, No. 11, pp. 1745–1755, 2018.
- [67] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2095–2102, 2017.
- [68] Eva Ceulemans, Tom F Wilderjans, Henk A L Kiers, and Marieke E Timmerman. Multilevel simultaneous component analysis: A computational shortcut and software package. *Behavior Research Methods*, Vol. 48, No. 3, pp. 1008–

- 1020, 2016.
- [69] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.
- [70] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2506–2513, 2019.
- [71] Jozsef Nemeth. Adversarial disentanglement with grouped observations. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 10243–10250, 2020.
- [72] Shuangfei Zhai, Walter Talbott, Miguel Angel Bautista, Carlos Guestrin, and Josh M. Susskind. Set distribution networks: a generative model for sets of images. 2020.
- [73] Hideaki Ishibashi and Tetsuo Furukawa. Hierarchical tensor SOM network for multilevel - multigroup analysis. *Neural Processing Letters*, Vol. 47, No. 3, pp. 1011–1025, 2018.
- [74] Christopher Bender, Kevin O’Connor, Yang Li, Juan Garcia, Junier Oliva, and Manzil Zaheer. Exchangeable generative models with flow scans. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, No. 06, pp. 10053–10060, 2020.
- [75] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
- [76] Kashif Rasul, Ingmar Schuster, Roland Vollgraf, and Urs Bergmann. Set flow: A permutation invariant normalizing flow. 2019.
- [77] 内山治樹. チーム・パフォーマンスの生成にかかる前提要件の検討. *体育・スポーツ哲学研究*, Vol. 37, No. 2, pp. 115–131, 2015.
- [78] Louise Lemieux-Charles and Wendy L McGuire. What do we know about health care team effectiveness? a review of the literature. *Medical Care Research and Review*, Vol. 63, No. 3, pp. 263–300, 2006.
- [79] Gerald Goodwin, Nikki Blacksmith, and Meredith Coats. The science of teams in the military: Contributions from over 60 years of research. *American Psychologist*, Vol. 73, pp. 322–333, 2018.
- [80] Granville King. Crisis management & team effectiveness: A closer examination. *Journal of Business Ethics*, Vol. 41, No. 3, pp. 235–249, 2002.

- [81] Suzanne T Bell, Shanique G Brown, Neal B Outland, and Daniel R Abben. Critical team composition issues for long-distance and long-duration space exploration a literature review, an operational assessment, and recommendations for practice and research. Technical Report February, 2015.
- [82] Kara Hall, Amanda Vogel, Grace Huang, Katrina Serrano, Elise Rice, Sophia Tsakraklides, and Stephen Fiore. The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, Vol. 73, pp. 532–548, 2018.
- [83] Suzanne T Bell and Shanique G Brown. Selecting and composing cohesive teams. In *Team Cohesion: Advances in Psychological Theory, Methods and Practice*, Vol. 17 of *Research on Managing Groups and Teams*, pp. 181–209. Emerald Group Publishing Limited, 2015.
- [84] John E. Mathieu, Scott I. Tannenbaum, Jamie S. Donsbach, and George M. Alliger. A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management*, Vol. 40, No. 1, pp. 130–160, 2014.
- [85] Suzanne T Bell, Shanique G Brown, and Jake A Weiss. A conceptual framework for leveraging team composition decisions to build human capital. *Human Resource Management Review*, Vol. 28, No. 4, pp. 450–463, 2018.
- [86] Jon Calder and Ian Durbach. Decision support for evaluating player performance in rugby union. *International Journal of Sports Science and Coaching*, Vol. 10, pp. 21–38, 2015.
- [87] Jerry Freischlag. Team dynamics implications for coaching. *Journal of Physical Education, Recreation & Dance*, Vol. 56, No. 9, pp. 67–71, 1985.
- [88] Faez Ahmed, Kalyanmoy Deb, and Abhilash Jindal. Multi-objective optimization and decision making approaches to cricket team selection. *Applied Soft Computing*, Vol. 13, No. 1, pp. 402–414, 2013.
- [89] Erin L Fitzpatrick and Ronald G Askin. Forming effective worker teams with multi-functional skill requirements. *Computers & Industrial Engineering*, Vol. 48, No. 3, pp. 593–608, 2005.
- [90] Christoph Dorn, Florian Skopik, Daniel Schall, and Schahram Dustdar. Interaction mining and skill-dependent recommendations for multi-objective team composition. *Data and Knowledge Engineering*, Vol. 70, No. 10, pp. 866–891, 2011.
- [91] Ewa Andrejczuk, Filippo Bistaffa, Christian Blum, Juan A Rodríguez-Aguilar,

- and Carles Sierra. Synergistic team composition: A computational approach to foster diversity in teams. *Knowledge-Based Systems*, Vol. 182, p. 104799, 2019.
- [92] Haibin Liu, Mu Qiao, Daniel Greenia, Rama Akkiraju, Stephen Dill, Taiga Nakamura, Yang Song, and Hamid Motahari Nezhad. A machine learning approach to combining individual strength and team features for team recommendation. In *Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA)*, pp. 213–218, 2014.
- [93] Sandesh Bananki Jayanth, Akas Anthony, Gududuru Abhilasha, Noorni Shaik, and Gowri Srinivasa. A team recommendation system and outcome prediction for the game of cricket. *Journal of Sports Analytics*, Vol. 4, No. 4, pp. 263–273, 2018.
- [94] Anna Sapienza, Palash Goyal, and Emilio Ferrara. Deep neural networks for optimal team composition. *Frontiers in Big Data*, Vol. 2, p. 14, 2019.
- [95] William A. Young II and Gary Weckman. A team-compatibility decision support system for the national football league. *International Journal of Computer Science in Sport*, Vol. 19, No. 1, pp. 60–101, 2020.
- [96] Onwuachu Uzochukwu C. and P Enyindah. A machine learning application for football players' selection. *International Journal of Engineering Research & Technology*, Vol. 4, No. 10, pp. 459–465, 2015.
- [97] Michelle Cheatham and Kevin Cleereman. Application of social network analysis to collaborative team formation. In *International Symposium on Collaborative Technologies and Systems (CTS'06)*, pp. 306–311, 2006.
- [98] Lincoln Magalhaes Costa, Alinne C. Correa Souza, and Francisco Carlos M. Souza. An approach for team composition in league of legends using genetic algorithm. In *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pp. 52–61, 2019.
- [99] Jieqiong Zhao, Morteza Karimzadeh, Luke S. Snyder, Chittayong Surakitban-harn, Zhenyu Cheryl Qian, and David S. Ebert. Metricsvis: A visual analytics system for evaluating employee performance in public safety agencies. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 26, No. 1, pp. 1193–1203, 2020.
- [100] Miohk Ryoo, Namjung Kim, and Kyoungju Park. Visual analysis of soccer players and a team. *Multimedia Tools and Applications*, Vol. 77, No. 12, pp. 15603–15623, 2018.
- [101] *A Guide to the Project Management Body of Knowledge*. Project Management

- Institute, 5 edition, 2013.
- [102] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, H Hauser, Silvia Miksch, and P Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. *Proceedings of the 16th Eurographics Conference on Visualization (EuroVis)*, pp. 1–21, 2014.
 - [103] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, Vol. 6, pp. 1783–1816, 2005.
 - [104] Peter Meinicke, Stefan Klanke, Roland Memisevic, and Helge Ritter. Principal surfaces from unsupervised kernel regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 9, pp. 1379–1391, 2005.
 - [105] 持橋大地, 大羽成征. ガウス過程と機械学習. 講談社, 2019.
 - [106] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.
 - [107] Tomislav Horvat and Josip Job. The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, p. e1380, 2020.
 - [108] Hyunwoo Park, Marcus A Bellamy, and Rahul C Basole. Visual analytics for supply network management: System design and evaluation. *Decision Support Systems*, Vol. 91, pp. 89–102, 2016.
 - [109] Xun Zhao, Yanhong Wu, Weiwei Cui, Xinnan Du, Yuan Chen, Yong Wang, Dik Lun Lee, and Huamin Qu. Skylens: Visual analysis of skyline on multi-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 24, No. 1, pp. 246–255, 2018.
 - [110] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2277–2286, 2013.
 - [111] *Handbook of multilevel analysis*. Springer, 2008.
 - [112] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, Vol. 38, pp. 43–54, 2017.
 - [113] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 10, pp. 1863–1883, 2019.
 - [114] 石黒勝彦, 林浩平. 関係データ学習. 講談社, 2016.

-
- [115] Evrim Acar, Tamara G Kolda, and Daniel M Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. In *Proceedings of the 9th Workshop on Mining and Learning with Graphs (MLG)*, 2011.
 - [116] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, Vol. 51, No. 3, pp. 455–500, 2009.
 - [117] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, Vol. 31, No. 3, pp. 279–311, 1966.
 - [118] Richard A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis. *UCLA working papers in phonetics*, Vol. 16, pp. 1–84, 1970.
 - [119] Zenglin Xu, Feng Yan, and Yuan Qi. Infinite tucker decomposition: Nonparametric bayesian models for multiway data analysis. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1675–1682, 2012.
 - [120] Bin Liu, Lirong He, Yingming Li, Shandian Zhe, and Zenglin Xu. Neuralcpc: Bayesian multiway data analysis with neural tensor decomposition. *Cognitive Computation*, Vol. 10, No. 6, pp. 1051–1061, 2018.
 - [121] Tohru Iwasaki and Tetsuo Furukawa. Tensor SOM and tensor GTM : Nonlinear tensor analysis by topographic mappings. *Neural Networks*, Vol. 77, pp. 107–125, 2016.

研究業績リスト

I. 学術論文

(国際・査読あり)

1. R. Watanabe, H. Ishibashi, T. Furukawa, “Visual analytics of set data for knowledge discovery and member selection support”, *Decision support systems*, Vol. 152, 2022.

II. 国際会議

(口頭発表・査読あり)

1. R. Watanabe, H. Ishibashi, T. Iwasaki, T. Furukawa, “Non-parametric Continuous Self-Organizing Map”, *Joint International Conference on Soft Computing and Intelligent Systems and International Symposium on Advanced Intelligent Systems*, pp614-617, 2018.

III. 国内学会

(ポスター発表・査読あり)

1. H. Ishibashi, R. Watanabe, T. Iwasaki, T. Furukawa, “Continuous latent variable model by non-negative kernel smoother”, *日本神経回路学会 全国大会*, 2017.
2. 渡辺 龍二, 宮崎 一希, 古川 徹生, “ノンパラメトリック表現を用いた Tensor SOM の連続表現化”, *日本神経回路学会 全国大会*, 2019.

(口頭発表・査読なし)

1. 宮崎 一希, 渡辺 龍二, 古川 徹生, “関係データの直積空間への埋め込みによる可視化” *電子情報通信学会研究報告*, Vol.119, No.283, pp23-26, 2020.

謝辞

本研究ならびに本研究に至るまでの研究や大学院生活において、数多くの方々のお世話になりました。こちらに感謝の意を述べさせていただきます。

まず九州工業大学の古川徹生先生をはじめとする、古川研究室で研究を積み重ねてこられた皆さんに深く御礼を申し上げます。本論文で提案している生成的多様体ネットワークモデリングは、これまで研究室で蓄積してきた生成モデルによる情報処理の知見がなければ成し得なかった成果です。このバトンを私までつないでいただいたことに、大変感謝しております。

その上で改めて古川研究室で特にお世話になった方々にお一人ずつ感謝申し上げます。指導教員である古川徹生先生には、6年間の大学院生活を通して数多くのご指導・ご鞭撻をいただきました。先生のおかげで生成モデルや潜在変数推定の面白さと奥深さを知ることができました。また先生には研究のプレイヤとしての姿勢を教えていただいたように思います。先生には感謝の念が絶えません。深く感謝申し上げます。

本論文の元となったジャーナル論文の共著者であり、研究室のOBでもある石橋英朗さんにも本当にお世話になりました。私が博士後期課程2年のときに、本論文の元となったテーマを石橋さんに勧めていたいなれば、私は本論文を書くことはできていないと思います。また研究生活の中で石橋さんの研究に対するストイックな姿勢を間近で見ることができ、研究者はかくあるべきという姿を学ぶことができたように思います。本当にありがとうございました。

同じく研究室のOBであり、研究室と長年に渡り共同研究をさせていただいている岩崎亘さんにも大変お世話になりました。修士の時に岩崎さんと一緒に共同でアルゴリズム開発を進め、たくさんのご助言をいただいたことは、今振り返ってみると本当に大きな経験だったと思っています。本当にありがとうございました。

同じく研究室のOBであり研究生活でもプライベートでもお世話になりました米田圭佑さんと比嘉一志さんにもお礼申し上げます。米田さんには本研究にも深く関わるマルチビュー表現学習の重要さを、比嘉さんにはベイズ学習の面白さを教えていただけたように

思えます。本当にありがとうございました。

それから6年間を共に過ごした同期である石田琢朗氏に、深くお礼申し上げます。良き研究のディスカッション相手であると共に、精神的に支えてくれた友人であると誠に勝手ながら思っております。感謝申し上げます。

他の研究室のメンバーの皆さんにも大変お世話になりました。とても書ききれませんでしたが、諸先輩方、ならびに同期の皆さん、後輩の皆さん本当にありがとうございました。特に先輩である波田野創さんには古川研の社会人博士として先陣を切っていただいてとても心強かったです。また後輩である宮崎一希さんもいつもしっかりしていて頼もしかったです。それから秘書である小谷紗代さんにはいつも事務手続きでお世話になったことはもちろんのこと、研究以外の雑談も楽しくさせていただき、研究の合間にリラックスすることができました。皆様に大変感謝しております。

研究室外で研究の遂行にご協力いただいた皆様にもお礼申し上げます。まず本論文の予備調査会ならびに本審査でお世話になりました九工大の堀尾恵一先生、長隆之先生、田向権先生、九州産業大学の磯貝浩久先生に深く感謝申し上げます。皆様に予備調査会でご助言いただいたおかげで、本論文の内容を深め、成長させることができました。本当にありがとうございました。

またVAについて知るキッカケをくださったお茶の水女子大学の伊藤貴之先生にも深く感謝申し上げます。今のこの博士論文がVA研究であると気づけたのは伊藤先生のおかげです。また突然の本論文に対するご助言のお願いも快く引き受けてくださいり、研究に対して「紛れもなくVAの研究」「ユニークな試み」といったお墨付きをいただけて、本論文の内容に自信を持つことができました。重ねてお礼申し上げます。

また現在勤めさせていただいているGMOペパボ株式会社のパートナーの皆さん、特にペパボ研究所の皆さんには、博士号取得に向けて多大なるご支援を賜りまして大変感謝しております。入社以前に参加させていただいたインターンで感じた「この研究所で研究をやりたい」という気持ちが、辛い研究生活を心折れずに過ごすための大きな原動力となつたように思います。深くお礼申し上げます。

岡本直香さんは、プライベートにおいて精神的に私を支えてくださいました。辛い博論執筆を乗り越えられたのは直香さんのおかげです。感謝の念が絶えません。

最後にこれまでの学生生活を支えてくださった両親と祖父、そしてこれまで私たち家族の元に来てくれた愛猫と愛犬たちに深く感謝申し上げます。