

Title

細胞集団プロファイルからの 特徴量抽出に関するデータ解析的研究

岡田大瑚

京都大学大学院医学研究科ゲノム医学センター統計遺伝学分野

指導教員：山田亮教授 (MD, PhD)

この論文は、2021年3月に博士(医科学)の学位論文として統計遺伝学分野に提出されたものです。

Declaration

この論文の内容は研究倫理を遵守して作成されています。また、論文の内容はオリジナルのものであり、他の情報源より得た情報については論文中に情報源を記載しています。

Acknowledgment

本論文を完成させるに当たり御指導を賜りました京都大学大学院医学研究科山田亮教授に深く御礼を申し上げます。また、本研究の遂行を支えて頂いた京都大学大学院医学系研究科ゲノム医学センターの皆様に心より感謝いたします。

Table of Contents

序文

第1章: シングルセル時代の遺伝疫学のモデルと細胞集団プロファイル

1.1 表現型の多様性を表現するモデル

- ・遺伝疫学の基本モデル
- ・オミックス多様性に基づいたモデル
- ・細胞集団プロファイルの多様性に基づいたモデル

1.2 細胞内のオミックス情報とその取得方法

- ・エピゲノム
- ・トランスクリプトーム
- ・プロテオーム
- ・メタボローム
- ・体細胞ゲノム

1.3 生体分子の状態空間上の分布としての細胞集団プロファイル

- ・生体分子の状態空間
- ・シングルセルデータとバルクデータの違い
- ・細胞集団プロファイルの分布の空間情報

1.4 細胞集団プロファイル多様性と個体のラベルの関連解析

- ・生命科学研究における細胞集団プロファイルの特微量抽出の位置づけ
- ・多細胞生物における細胞の分類
- ・これまでの細胞集団プロファイルの特微量抽出法

1.5 細胞集団プロファイルのダイナミクス

1.6 本論文の目的と構成

第2章: サイトメトリーによる細胞集団プロファイルの特微量抽出

2.1 マニュアルゲーティングによる特微量抽出

2.2 計算サイトメトリーによる特微量抽出

- ・SPADE 法によるクラスタリング
- ・パラメトリックモデルによるクラスタリング
 1. EM アルゴリズム
 2. 混合正規分布フィッティング

2.3 ノンパラメトリックモデルを用いたデータドリブンな特微量抽出

2.4 これまでの手法の課題と情報幾何の応用

第3章: シングルセル発現解析のための情報幾何

3.1 双対平坦座標系

- ・リーマン多様体、計量、接続、測地線
- ・Bregman ダイバージェンスと双対平坦座標系
- ・双対平坦座標系の性質
 - ・双直交性
 - ・拡張ピタゴラスの定理
 - ・拡張射影定理

3.2 指数型分布族と統計多様体

- ・指数型分布族における双対平坦座標系
- ・指数型分布族の十分統計量
- ・内積行列としての Fisher 情報行列
- ・Kullback–Leibler ダイバージェンス

3.3 データ解析手法の情報幾何的解釈

- ・離散確率分布の指数型表現
- ・指数型分布族の母数パラメータの最尤推定
- ・EM アルゴリズムの情報幾何

第4章: 拡大指数型分布族分解法の提案

4.1 拡大指数型分布族分解法の提案

- ・動機
- ・理論とアルゴリズム

4.2 シミュレーションデータ解析

4.3 実データ解析

- ・材料と方法

- サイトメトリーデータからの確率行列 P の構築
- EGF stimulation data への適用
- SPADE による $F(x)$ 関数の可視化
- EGF 刺激データの次元削減とタイムコースの再構成
- ・結果
 - θ 座標と $F(x)$ 関数
 - 次元削減と時系列の再現

4.4 Appendix

第5章：総合考察

序文

本論文は、細胞集団プロファイルから特微量抽出を行うためのデータ解析手法についての研究論文である。

第1章では、論文の背景としてシングルセル時代の遺伝疫学のモデルと細胞集団プロファイルについて述べる。近年ではシングルセルオミックスデータが利用可能になり、これらのデータを取り入れた遺伝疫学研究が求められる。そこで、細胞集団プロファイルに基づく遺伝疫学モデルを提唱し、シングルセルデータを用いた細胞集団プロファイルからの特微量抽出が、シングルセルデータを用いた遺伝疫学研究を推進するうえでのデータ解析における主要なトピックであることを導入する。

第2章では、代表的なシングルセルデータとしてサイトメトリーデータを取り上げ、そこから細胞集団プロファイルの特微量抽出を行うこれまでの手法や考え方について紹介する。また、これらの手法を、特微量抽出の際の情報ロス、特微量の解釈のしやすさという2点から評価し、両者を満たす手法がいまだ存在しないことを問題提起する。また、確率分布に座標を与えその幾何学的構造を研究する情報幾何学の知見を用いて、これらの課題を解決する新規の手法を開発する着想について述べる。

第3章では、シングルセルデータを用いた細胞集団プロファイルの特微量抽出手法を考えるうえで必要となる、情報幾何学の基本的な事項について紹介する。情報幾何学は多くのトピックを持った広がりのある分野であるが、ここでは指数型分布族の性質に焦点を当てて基本的な事項についてまとめる。

第4章では、1-3章の内容を踏まえ、情報幾何学の知見を用いた新規のノンパラメトリックデータ解析手法である拡大指数型分布族分解法を提案する。

第5章は総合考察であり、提案手法に関する考察および今後の展望について述べる。

第1章:シングルセル時代の遺伝疫学のモデルと細胞集団プロファイル

岡田大瑚

2021年3月23日

概要

生物個体に現れる表現型の多様性について理解することは医学・生命科学において重要な課題である。遺伝疫学は、主として表現型の多様性に対する遺伝的効果の統計的な評価を行うが、近年さまざまなオミックスデータが取得できるようになったことにより急速に進歩している分野である。さらにここ数年では、オミックス情報をシングルセルレベルで取得することが可能になったため、これらに対応したシングルセル時代の遺伝疫学モデルが必要である。本研究では、表現型の多様性を表現するためのモデルを4種類にまとめ、シングルセル時代の遺伝疫学モデルとしての細胞集団プロファイルに基づくモデルについて議論する。次に、細胞集団プロファイルの情報は生体分子の状態空間上の分布であることに着目し、分布特徴量の観点から既存の生命科学データおよびその解析手法を整理する。さらに、細胞集団プロファイルの分布としての性質についてまとめ、細胞集団プロファイルからの特徴量抽出が、シングルセル時代の遺伝疫学における重要なデータ解析的トピックであることを示す。

1 表現型の多様性を表現するモデル

生物個体に現れる表現型の多様性について理解することは医学・生命科学において重要な課題である。遺伝疫学は、表現型を遺伝的効果と環境効果を組み合わせたモデルによって評価し、表現型の多様性に寄与するような遺伝的効果の定量や関連遺伝子の同定を行い、生命現象・病理現象の理解に大きく貢献してきた。

近年、ゲノム・オミックスデータが利用可能になったことによって、遺伝疫学分野は大きく発展している学問分野である。特に、網羅的なゲノム情報を用いた Genome-Wide Association Study (GWAS) は、疾患などの複雑形質に影響を与える遺伝的変異を数多く同定している [1]。また、ゲノム情報だけでなく他のオミックス情報も用いて、表現型の

多様性を解析することも可能である。このようなゲノム疫学は、複雑な生命現象・病理現象のメカニズムを解明する上で強力な手法である。さらに、ここ数年では、オミックスデータをシングルセルレベルで測定する技術が飛躍的に進歩しており、これらのオミックスデータを組み合わせた遺伝疫学的手法の開発がますます重要になっている。本章では、表現型の多様性を理解するための遺伝疫学モデルに焦点を当ててこれまでの手法を整理するとともに、シングルセル時代の新たな遺伝疫学モデルとして細胞集団プロファイルの多様性に基づくモデルを提案する。

1.1 遺伝疫学の基本モデル

表現型の多様性を遺伝的効果と環境効果の組み合わせとして表現するモデルは、遺伝疫学において最も基本的なフレームワークである (Standard Model:図 1(a))。GWAS を含めた遺伝疫学研究の多くはこのフレームワークに基づき、線形回帰や分割表検定といった統計モデルを用いて、遺伝因子の関連解析を行っている。この基本モデルでは、遺伝的多様性から表現型多様性への因果関係のみに着目しており、そのメカニズムについての洞察は含んでいないことが特徴である。

1.2 オミックス多様性に基づいたモデル

遺伝的多様性は、RNA やタンパク質などの生体分子の多様性を介して形質の多様性に影響している。網羅的な生体分子情報はオミックス情報と呼ばれ、ゲノム、トランスクリプトーム、プロテオーム、エピゲノム、メタボロームに分けられ、それぞれ実験によって計測することが可能である。そのため、近年の遺伝疫学においては、ゲノムデータだけではなくこのような各種オミックス情報の多様性を介して表現型の多様性を理解する研究も盛んに実施されている。

基本モデルにオミックス情報を加えて詳細化し、遺伝疫学領域におけるオミックス解析を表現したものが図 1(b) に示す Omics Model である。このモデルでは、遺伝的効果と環境効果は、生体分子情報の多様性を介して表現型の多様性に寄与する。トランスクリプトーム、プロテオーム、メタボローム、エピゲノムといった生体分子情報の多様性に対する遺伝的効果の研究としては、これらに関連する SNP(expression QTL, protein QTL, methylation QTL, metabolyte QTL) を同定が盛んに実施されている [2, 3]。例えば、eQTL 解析は、トランスクリプトームデータより得られる遺伝子発現量と関連する遺伝的変異を同定する。様々な組織で同定された eQTL は、データベースなどで公開されてい

る [4, 5]。また、オミックス多様性と表現型多様性の関連を調べる研究の例としては、トランスクリプトームデータを用いて、疾患のある個体と健康な個体における Differential expressed genes を同定する疾患オミックス解析が含まれる [6]。これらはオミックス研究においての代表的な研究デザインとして広く実施されている。

1.3 細胞集団プロファイルの多様性に基づいたモデル

医学・生命科学において対象となる表現型は、疾患など組織や個体レベルで発現する表現型である。このような表現型は、特定の細胞 1 個だけによって引き起こされるわけではなく、関連する組織における細胞集団全体の特性によって生じる。実際に、オミックス解析で用いる組織サンプルは、細胞の集団として成り立っており、サンプルを構成する細胞 1 つ 1 つがオミックス情報を持っている。ここ数年でのシングルセル解析技術の飛躍的な進歩により、オミックス情報をバルクレベルのみならずシングルセルレベルで取得することも可能になっている。以上のことより、Omics Model をシングルセルレベルの表現に拡張した遺伝疫学モデルが考えられる。

Cell Population Model は、Omics Model を拡張し、表現型の多様性を細胞集団プロファイルの多様性として表現したものである。生体内の細胞 1 つ 1 つがオミックス情報を持つておらず、そのどれも少しずつ違っている。このモデルでは、表現型の多様性は、そのような細胞集団のプロファイルで説明されると考える。遺伝的効果と環境効果は、この細胞集団プロファイルの多様性を介して表現型の多様性に影響を与える。このモデルは、複雑表現型の生物学的な発現機構としても妥当なものであると考えられる。

本研究に先立ち、数種類の細胞表面タンパク質マーカーのサイトメトリーデータを用いて末梢血中のリンパ球集団の多様性を対象とした GWAS を実施した [7]。その結果、比較的小さいサンプルサイズでの解析ではあったものの、その個人差に関連する SNP の検出に成功している。近年では、大規模にシングルセルデータを取得する研究も実施されており、このようなモデルの下での遺伝疫学研究は新たな知見をもたらすことが期待できる。

この Cell Population Model を複数組織に拡張したモデルが Multi Tissue Model である(図 2)。これは、関連する複数の組織の細胞集団プロファイルの影響の組み合わせとして表現型の多様性を表現したモデルである。全身性の疾患など、複数の組織、細胞集団が関連している形質ではこのようなモデルはその発現メカニズムの自然な表現になっている。Multi Tissue Model を用いた遺伝疫学研究は実施されていないものの、シングルセル時代の将来的な遺伝疫学モデルとして有意義なものであると考えられる。

2 細胞内のオミックス情報とその取得方法

細胞の持つ生体分子情報は、エピゲノムレイヤー、トランスクリプトームレイヤー、プロテオームレイヤー、メタボロームレイヤー、体細胞ゲノムレイヤーに分類できる [8, 9]。これらの生体分子情報の取得方法には、あらかじめ定めた特定の分子のみを調べる実験手法と、各レイヤーの情報を網羅的に調べる実験手法がある。また、実験対象とする生体組織に対して、バルクで測定する手法とシングルセルレベルで測定する手法がある。以下では、各オミックスレイヤーの生体分子情報の特徴とその取得方法について順に紹介する。

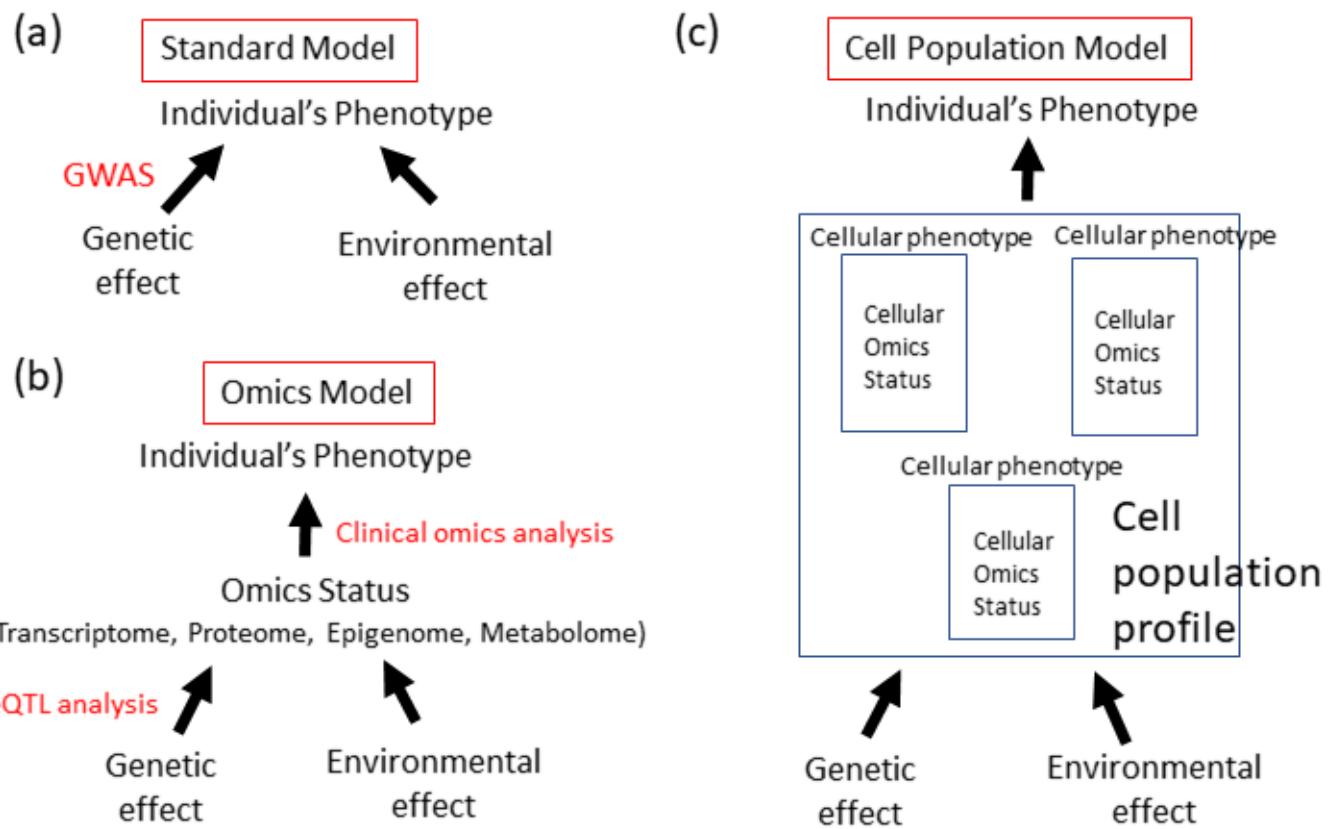


図1 遺伝疫学におけるモデル (a)Standard Model(b)Omics Model(c)Cell Population Model

Multi Tissue Model

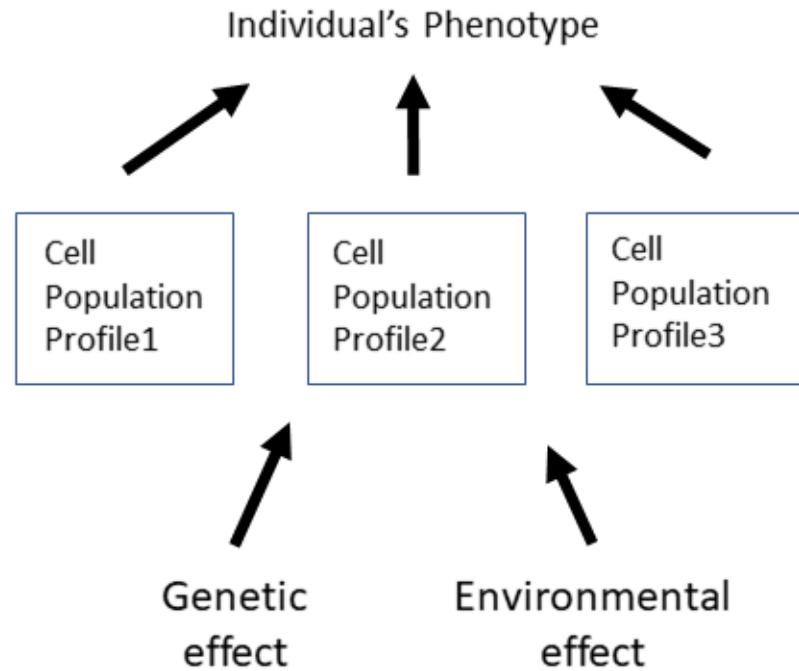


図 2 Multi Tissue Model の図

2.1 エピゲノム

エピジェネティクスは、DNA の塩基配列によらない遺伝子発現制御のメカニズムであり、そのようなエピジェネティクスの情報の総体をエピゲノムと呼ぶ。DNA のメチル化やヒストンタンパク質アセチル化といった DNA 分子やヒストンタンパク質の化学修飾がエピゲノム情報に該当する [10]。同一のゲノムを持つ細胞が異なる機能を持つ細胞に分化することができる原因是、これらの異なるエピゲノムプロファイルを持つことによるものである。

エピゲノムレイヤーにおける生体分子情報は、主に次世代シークエンサーを用いて網羅的に取得される。代表的な技術として、ヒストンタンパク質の化学修飾情報を取得する ChIP-seq やオープンクロマチン情報を取得する ATAC-seq がある。これらの実験は、バルクでの解析だけでなく、近年ではシングルセルレベルでの測定を行なうこともできる [11, 12, 13]。エピゲノム情報は、次世代シークエンサーを用いてゲノム上にマップされたショートリードとして定量され、ゲノム上の各塩基上のシグナル値として表現できる。それぞれの細胞が持つエピゲノム情報には、オープンクロマチン情報や種々のヒストン修飾情報など様々なものがあるため、細胞のエピゲノムレイヤーの情報はそれぞれの化学修飾に対応した複数のベクトルのセットとして存在している。

2.2 トランスクリプトーム

トランスクリプトームは転写された RNA の情報の総体である。ゲノム上の遺伝子が機能を発現するためには、RNA に転写される必要がある。生体内の RNA には、さらにタンパク質に翻訳されることによって機能を発現する mRNA と、RNA の状態で機能する non-coding RNA がある。non-coding RNA は、micro RNA や long non-coding RNA, piRNA などの種類があり、発生や疾患など様々な高次生命現象・病理現象に関与していることが知られている [14]。

トランスクリプトームレイヤーにおける生体分子定量手法も、あらかじめ定めた特定の転写物に対して定量を行う実験方法と、網羅的な解析を行う実験手法の 2 パターンが存在する。特定の遺伝子の転写物を定量する方法としてはノザンプロット法や RT-qPCR 法、トランスクリプトームを網羅的に測定する場合は DNA マイクロアレイや RNA-seq が盛んに用いられている。RNA-seq を 1 細胞に対して適用することで細胞 1 つ 1 つの遺伝子発現情報を網羅的に測定する single cell RNA-seq は、近年の代表的なシングルセル

オミックス技術である。single cell RNA-seq のデータ解析手法は計算生物学領域でも盛んに研究されている [15]。

2.3 プロテオーム

プロテオームはタンパク質の総体である。生体内のタンパク質は、mRNA が翻訳されることによって発現する。タンパク質はアセチル化、リン酸化などの翻訳語修飾を受けることによって機能が調節される。また、真核生物では、選択的スプライシングによって一つの遺伝子から複数の種類のタンパク質ができ、それぞれの isoform ごとに違った機能を持っている。したがって、タンパク質は mRNA よりも大きい多様性が存在する。

プロテオームレイヤーにおける生体分子定量手法も、あらかじめ定めた特定のタンパク質に対して定量を行う実験方法と、網羅的な解析を行う実験手法が存在する。前者のパターンでは、主として蛍光抗体が用いられ、バルクでの解析では Western blot や ELISA が分子生物学分野で広く用いられている。シングルセルでの解析技術には、数種類から数十種類程度の特定の細胞表面タンパク質発現量を細胞 1 つ 1 つについて測定するサイトメトリーがあり、幅広い生命科学分野で用いられている [16]。例えばリンパ球の分類は、細胞表面タンパク質マーカーの発現パターンの組み合わせによって定義されており、免疫学研究ではサイトメトリーデータを用いて T 細胞、B 細胞などのリンパ球サブセットの同定、定量が行われる。タンパク質発現量を網羅的に定量する場合は、主に質量分析計を用いた方法が用いられ、これはプロテオーム解析と呼ばれる [17]。

2.4 メタボローム

メタボロームは生体内の低分子化合物の総体である。メタボロームレイヤーの生体分子には、脂質、アミノ酸、その他の低分子化合物など様々なものがある [18]。タンパク質と同様、バルクレベルで特定の分子に対して定量を行う場合は ELISA などが用いられる。網羅的に取得する場合は質量分析計を用いた方法が使われ、メタボローム解析と呼ばれる。また、近年ではシングルセルレベルでメタボローム情報を取得することも可能になっている [19]。

2.5 体細胞ゲノム

ゲノムは DNA 配列情報の総体である。同一個体内では、細胞は同じゲノム情報を持ち、刺激や環境に応答して変化しない。しかしながら、DNA 複製の際のエラーや DNA

損傷などによって、体細胞ゲノムの中にも突然変異やDNA鎖の切断などが生じ、同一個体内においても体細胞ゲノム情報に多様性が生じてくる。病理現象の中には細胞に蓄積する体細胞ゲノムの変異や損傷が影響しているものもある。例えば、がんでは異常なゲノム情報をもった細胞の増加によって生じる疾患であり、癌細胞のゲノム解析によって疾患の発現に関わる遺伝子の同定などが行われている[20]。

3 生体分子の状態空間上の分布としての細胞集団プロファイル

3.1 生体分子の状態空間

生体内の細胞1つ1つが各オミックスレイヤーの生体分子情報を持っているため、それぞれの細胞は各生体分子データを座標軸とする状態空間の点として表現される。したがって、細胞集団プロファイルは状態空間上の分布として特徴づけられる(図3)。この分布は全ての遺伝子発現、タンパク質発現、体細胞ゲノム上の変異、エピゲノム修飾を含む生体分子マーカーの測定値の同時分布に相当するので、非常に高次元の分布になる。細胞はこの空間上に満遍なく観測されるわけではなく、ほとんどの部分では細胞が全く観測されない疎な領域になっている。生体分子量を測定する分子生物学実験データは、この細胞集団の分布の一部を捉えたものとして解釈できる。以降は、この全てのオミックスレイヤーの生体分子の状態空間をオミックス状態空間と呼び、細胞集団プロファイルをこの空間における分布として定義し、以降の議論を進める。

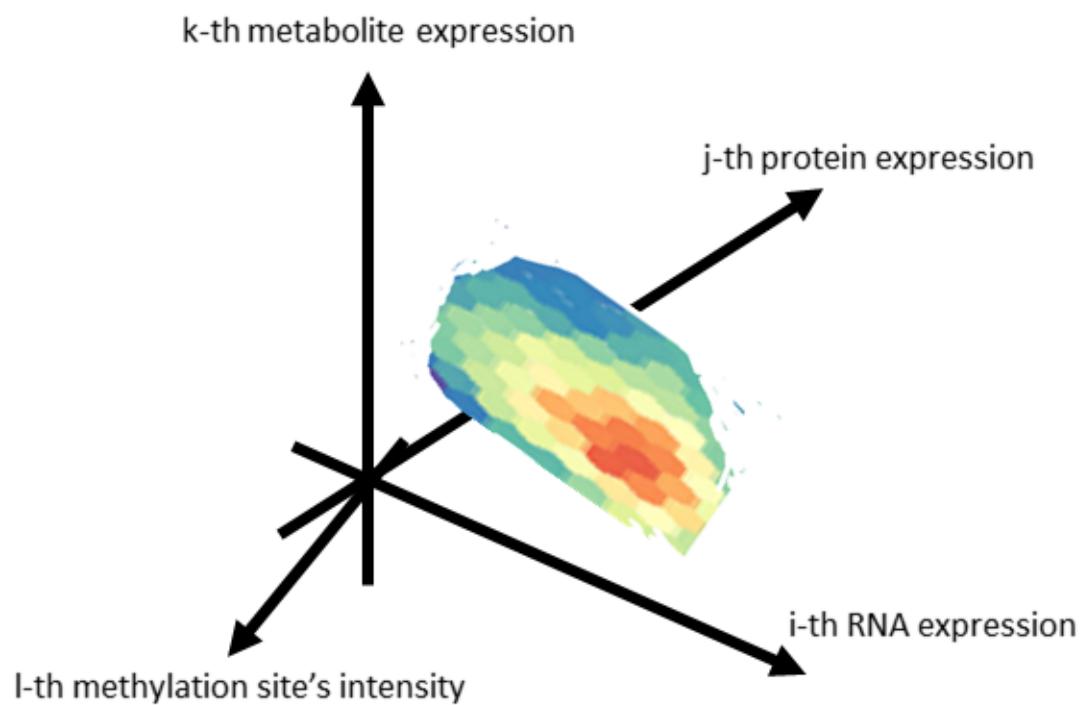


図3 オミックス状態空間のイメージ図

3.2 シングルセルデータとバルクデータの違い

シングルセルデータとバルクデータは、そのデータ構造や取得している情報の種類が異なっている。生体分子情報を測定したデータは、このオミックス状態空間上の分布を持つ情報の一部として整理することができる。

バルクの測定データは、オミックス状態空間における分布の特定の軸に関する平均値を推定している。例えば、Western blot によってタンパク質 A の発現量を定量したとき、その定量値はサンプル細胞集団の分布におけるタンパク質 A 軸の値の平均である。言い換えると、オミックス状態空間の分布において”タンパク質 A の発現量”以外の変数を積分消去し、得られた 1 次元分布の期待値を推定したものに相当する。確率分布における平均値は分布の代表値としては代表的かつ妥当な指標であるため、細胞集団が均質である場合には分布同士の比較指標としてバルクでの測定値は妥当である。

シングルセル発現データは特定の面にオミックス状態空間上の分布を射影した影を見ていることに相当する。例えば、 m 種類の細胞表面タンパク質を測定したサイトメトリーデータは、この分布からの標本に対して、この m マーカーの発現値のみを得ていることになる。バルクデータでは分布の形状に関する情報は失われているため、平均値だけでは説明できない分布の形状の情報のロスが発生する。一方で、シングルセルデータでは、バルクデータからは得ることできない分布の形に関する情報を得ることができ、細胞集団の heterogeneity やサブセットの同定や定量を行うことができる。

3.3 細胞集団プロファイルの分布の空間情報

細胞集団プロファイルを分布としてみる場合、各データ点を独立とみなしており、細胞の位置情報は消えている。末梢血中の血球細胞集団はこの前提が問題なく成立する。末梢血から採血した場合、細胞はそれぞれ独立かつランダム採取されたとみなすことができ、得られたシングルセルデータは母分布からの標本とみなすことができる。一方で、多くの解剖学的に定義された組織では、細胞が適切な生体分子の発現状態を維持しているだけではなく、それぞれの細胞が組織中で適切な位置を占めていることも組織の正常な機能発現において必要である。例えば、組織幹細胞はニッチと呼ばれる微小環境中において維持されることなどがあげられる [21]。このような細胞集団を分布としてみなすと、生物学的な情報を欠損することになる。

近年では、生体内での位置情報とオミックス情報を同時に取得する技術が注目を集めて

いる。例えば、spatial transcriptome は組織中の空間情報をのこしたままトランスクーリプトームデータを取得することができる [22]。このような spatial データによる空間情報は、分布として扱うことができる細胞集団の範囲の決定や、分布化することで消失する位置情報を補う上で有用であると考えられる。

4 細胞集団プロファイルの多様性と個体ラベルの関連解析

第 1 節では、細胞集団プロファイルの多様性を、表現型の多様性や遺伝的多様性と関連付けるシングルセル時代の遺伝疫学モデルとしての Cell Population Model やそれを多組織まで拡張した Multi Tissue Model について述べた。このようなモデルに基づく遺伝疫学研究をデザインするためには、個体の細胞集団プロファイルと個体の遺伝子型や表現型などの個体にラベルされる量との間の関連解析を実施する必要がある。

4.1 生命科学研究における細胞集団プロファイルの特徴量抽出の位置づけ

細胞集団プロファイルはオミックス状態空間上の複雑な高次元分布として表現されるため、このままでは従来の遺伝疫学の手法やオミックス解析の手法を適用できない。これらの統計解析の手法を適用し、個体の遺伝子型や表現型との関連解析を実施するためには、細胞集団プロファイルの特徴量抽出を行い、その代表値を得る必要がある。以降では、細胞集団プロファイルが持つ性質や特徴量抽出についてこれまでの研究について紹介する。

4.2 多細胞生物における細胞の分類

ヒトの体は数十兆個の細胞からできており、これらの細胞が協力、分担して機能を発現している。生命現象を理解する上では、この細胞たちの機能や働きを理解することが必要である。

細胞集団中の細胞は、サブセットに分類して理解することができる。細胞 1 つ 1 つはどれも少しずつ違っておりどの細胞も全く同じものはない。一方で、細胞集団は細胞が増殖・分化していく中で形成されるため、集団中には同じ性質、機能を持つ細胞サブセットが存在している。以上のことから、集団中の細胞をサブセットに分類し、それぞれのサブセットごとの機能を解明することで、細胞の機能について理解できると考えられる。細胞の分類は階層構造を取る [23]。例えば、”リンパ球”は 1 つの細胞サブセットの分類群であるが、T 細胞や B 細胞などさらに細かく分類することもできる。

顕微鏡下で観察される形態的特徴による分類は古くからある方法の1つである。例えば、白血球の下位の分類である好中球、リンパ球、好酸球、好塩基球、单球への分類は染色技術や顕微鏡技術を用いて細胞を観察し、その形態的特徴に基づいて同定することができる[24]。また、老化細胞においても細胞の肥大化など形態的な特徴をもつ[25]。近年では、情報科学的手法を用いて細胞の形についてのデータドリブンな特徴量を抽出する研究も行われている[26, 27]。

形態に基づく分類は効果的な手法であるが、より詳細な分類を行うためには生体分子情報を用いる必要がある。例えば、リンパ球サブセットは、CD3, CD4などの細胞表面マーカーの発現パターンによって分類されている。また、近年では、サイトメトリーデータやシングルセル RNA-seq データなどのシングルセルデータに対して情報科学的手法を適用することによってデータドリブンな細胞サブセット定量も行なわれている。このような計算生物学的な手法は、新規のサブセットの検出や、既存のサブセットのより詳細な分類を提供する上での有用な材料となる。

4.3 これまでの細胞集団プロファイルの特徴量抽出法

前節では、細胞集団プロファイルは複数の細胞サブセットから構成されることを示した。このような点から、細胞集団プロファイルの分布はサブセットごとの混合分布の形状を取っていると考えられる。この点を踏まえて、細胞集団プロファイルの特徴量抽出について既存の考え方を、バルクデータによる方法、細胞サブセットに基づく方法、ノンパラメトリックな方法に分けて紹介する。

まず、バルクデータによって取得される分布平均値は、細胞集団プロファイルのよく用いられる特徴量の1つと言える。例えば、バルクでの遺伝子発現解析は、組織特異的遺伝子を同定することに大きく貢献している[28]。ここでの組織特異的遺伝子は該当組織の細胞集団の分布での遺伝子発現量の平均値を各組織の細胞集団プロファイルの特徴量として採用し、組織ごとの分布の違いを代表するような特徴量を探索していると考えることができる。バルクデータを用いた生体分子マーカーの探索は数多く実施されている[29, 30]。

細胞サブセットに基づく特徴量抽出は、シングルセルデータを用いた解析の中で頻繁に用いられる手法である。細胞集団プロファイルは細胞サブセットの混合分布であるため、細胞の分類に基づく各サブセットがどのくらいの割合で存在するかの定量値も細胞集団プロファイルの妥当な特徴量である。免疫学では、主としてサイトメトリーデータを用いて、疾患や高次生命現象と関連するサブセットの同定を行っている。

しかしながら、このようなサブセット定量値による特徴量化もデータの持つ情報を失っ

ている。この理由の1つは、特微量抽出の結果が、細胞の分類についての事前の生物学的仮定の影響を受けるためである。細胞サブセットの数が正確に何種類あるのかもしくは何種類あるとすべきなのは不明であり、現在でも未知のサブセットが新たに同定されている。情報科学的な手法を用いたデータドリブンな分類であっても、適用するアルゴリズムや統計モデル中に含まれる仮定によって、このようなバイアスをなくすことはできない。加えて、細胞サブセット内でのばらつきや多様性に関する情報も失われている。

ノンパラメトリックな特微量抽出では、このような細胞サブセットの情報をもちいずに特微量を取得する方法である。ノンパラメトリックな統計モデルとは、パラメーターの数を気にすることなしにデータの従う確率分布をモデル化する手法であり、これまでに情報理論に基づいた分布間異同定量と多次元尺度構成法 (Multi-dimensional scaling: MDS) を組み合わせて行うものが提案されている [31, 32]。ここでは、それぞれのサンプルのシングルセルデータの母分布をノンパラメトリックに推定して分布間距離行列を計算し、この距離行列に MDS を適用して距離関係をよく反映する座標を取得する。ここで取得される座標は、細胞集団プロファイルのデータドリブンな特微量として扱うことができる。

第2章では、シングルセルデータを用いたこれらの方法に焦点を当て、それらの手法の詳細と課題について詳述する。

5 細胞集団プロファイルのダイナミクス

ここまででは、従来の遺伝疫学モデルから出発し、オミックス状態空間における分布の個体間での多様性によって、複雑表現型の個体間の多様性を表現するモデルを提案した。同様に、個人間での異同だけでなく、同一個人内での複雑表現型のダイナミクスを理解することは疾患などの制御において重要である。例えば、薬剤刺激に対して生体がどのように反応するかや、その分子的メカニズムを理解することは生命科学における基本的な研究テーマとなっている。このような個体内で生じる生命現象においても、関連する細胞集団においての細胞集団プロファイルの分布の変化を基盤として説明できる。ここでは、そのような同一個体内で生じる生命現象のダイナミクスについてオミックス状態空間の観点からまとめる。

細胞集団プロファイルは外部からの刺激に応答して時系列で変化することで生命現象を引き起こす。例えば、ワクチンを接種すると、末梢血中のリンパ球集団のプロファイル分布が変化して免疫応答を起こす。また、外部からシグナルを受けることで、細胞は増殖、分化などの応答を引き起こす。したがって、ある時点における細胞が次にどのような状態を取るかは、現時点での細胞の状態とその近傍の外部環境に依存していると考えられる。

このことをオミックス状態空間の観点から見ると、ある座標にある細胞が刺激の種類に応じて別の点に移動することで、全体の分布が変化するといえる。ここで分布の変化は、刺激を受けた細胞が細胞死または他の細胞集団に移動し集団からなくなることや、細胞分裂を行い増殖することによって細胞集団プロファイルに影響を与える場合もある。この変化は、細胞表面の受容体タンパク質が外部の物質を感知し活性が変化するというプロテオームレイヤーにおける変化をトリガーとして生じる。特定の刺激に応答して、細胞がオミックス状態空間において、どこの点からどこの点に細胞が移動するか、ある時刻にある点に存在した細胞が次の時刻にどこの点に移るのかは、生体分子間の制御関係によって規定される。

生命現象、病理現象を引き起こす生体分子間の制御関係をまとめた経路を生物学的パスウェイと呼ぶ。生物学的パスウェイは、生体分子の化学修飾や複合体の形成、遺伝子の転写制御、細胞内での核内から核外への移動など様々な生物学的過程を含む経路として表現され、生命現象・病理現象の発現を制御する分子的メカニズムを表す。生物学的パスウェイの解明は分子生物学の主要な研究対象となっており、これらの過程は生体分子の相互作用の様相をグラフィカルに描画したイラストとして表現されることが多い。実験的に決定された生物学的パスウェイの情報は、文献やKEGGなどのパスウェイデータベースにもまとめられている [33]。

まとめると、外部からの刺激を受けると細胞集団中のそれぞれの細胞は、オミックス状態空間における位置と外部環境に応じて、生物学的パスウェイで既定された遷移ルールに則って別の場所に移動する。その結果、細胞集団プロファイルの分布が変化し、それに伴い各時刻における表現型も変化する。このように、個体内で起こる生命現象も、オミックス状態空間の分布を中心に考えることができる。

このような細胞集団プロファイルのダイナミクスも、特徴量抽出を行うことによって通常のデータ解析の手法を適用することができる。サンプルごとの細胞集団プロファイルの変化は、その特徴量の時系列データとして可視化、解析することができる。時系列データの解析手法は様々なものが提案されており、生命科学データにおいても豊富な手法が利用可能である。

6 本論文の目的と構成

細胞集団プロファイルはオミックス状態空間における複雑な多次元分布として表現される。個人間での多様性を対象にした解析、個人内での変化を対象にした解析のどちらにおいても、オミックス状態空間上の分布としての細胞集団プロファイルからの特徴量抽出

を行うことで既存のデータ解析の手法が適用可能になる。

本論文では、シングルセル発現データからの細胞集団プロファイルの特微量抽出についてデータ解析手法の観点から研究する。第2章では、歴史が長く多くの研究で使用されているサイトメトリーデータを対象に、これまでの解析手法について紹介し、それらの手法の課題を提起する。第3章では、確率分布に理論的な座標を与え、その幾何学的な性質を研究する数学の一分野である情報幾何学の基本的な概念について、指数型分布族の性質を中心に紹介する。第4章では、情報幾何学の知見を用いて、現在までのシングルセル発現データ解析の課題を克服するような新規手法である拡大指数型分布族分解法を提案する。

参考文献

- [1] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derk. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, Vol. 27, No. 2, p. e1608, 2018.
- [2] Jia Wen, Conor Nodzak, and Xinghua Shi. Qtl analysis beyond eqtls. In *eQTL Analysis*, pp. 201–210. Springer, 2020.
- [3] Pascal Schlosser, Yong Li, Peggy Sekula, Johannes Raffler, Franziska Grundner-Culemann, Maik Pietzner, Yurong Cheng, Matthias Wuttke, Inga Steinbrenner, Ulla T Schultheiss, et al. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nature genetics*, Vol. 52, No. 2, pp. 167–176, 2020.
- [4] Maiko Narahara, Koichiro Higasa, Seiji Nakamura, Yasuharu Tabara, Takahisa Kawaguchi, Miho Ishii, Kenichi Matsubara, Fumihiko Matsuda, and Ryo Yamada. Large-scale east-asian eqtl mapping reveals novel candidate genes for ld mapping and the genomic landscape of transcriptional effects of sequence variants. *PloS one*, Vol. 9, No. 6, p. e100924, 2014.
- [5] GTEx Consortium, et al. Genetic effects on gene expression across human tissues. *Nature*, Vol. 550, No. 7675, p. 204, 2017.
- [6] Olga Zolotareva, Sahand Khakabimamaghani, Olga I Isaeva, Zoe Chervontseva, Alexey Savchik, and Martin Ester. Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics (Oxford, England)*, p.

btaa1038.

- [7] Daigo Okada, Naotoshi Nakamura, Kazuya Setoh, Takahisa Kawaguchi, Koichiro Higasa, Yasuharu Tabara, Fumihiko Matsuda, and Ryo Yamada. Genome-wide association study of individual differences of human lymphocyte profiles using large-scale cytometry data. *Journal of Human Genetics*, pp. 1–11, 2020.
- [8] Baiba Vilne and Heribert Schunkert. Integrating genes affecting coronary artery disease in functional networks by multi-omics approach. *Frontiers in cardiovascular medicine*, Vol. 5, p. 89, 2018.
- [9] Jingwen Yan, Shannon L Risacher, Li Shen, and Andrew J Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, Vol. 19, No. 6, pp. 1370–1381, 2018.
- [10] PatriziaD’ Aquila, Giuseppina Rose, Dina Bellizzi, Giuseppe Passarino. Epigenetics and aging. *Maturitas*, Vol. 74, No. 2, pp. 130–136, 2013.
- [11] Assaf Rotem, Oren Ram, Noam Shores, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, Vol. 33, No. 11, pp. 1165–1172, 2015.
- [12] Kevin Grosselin, Adeline Durand, Justine Marsolier, Adeline Poitou, Elisabetta Marangoni, Fariba Nemati, Ahmed Dahmani, Sonia Lameiras, Fabien Reyal, Olivia Frenoy, et al. High-throughput single-cell chip-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics*, Vol. 51, No. 6, pp. 1060–1066, 2019.
- [13] Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, Vol. 20, No. 1, pp. 1–25, 2019.
- [14] Ertan Altayli. Regulator non-coding rnas: mirna, sirna, pirna, lncrna, circrna. *J Clin Med Kaz*, Vol. 6, No. 60, pp. 29–39, 2020.
- [15] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, Vol. 21, No. 1, pp. 1–35, 2020.

- [16] Aysun Adan, Günel Alizada, Yağmur Kiraz, Yusuf Baran, and Ayten Nalbant. Flow cytometry: basic principles and applications. *Critical reviews in biotechnology*, Vol. 37, No. 2, pp. 163–176, 2017.
- [17] AF Maarten Altelaar, Javier Munoz, and Albert JR Heck. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, Vol. 14, No. 1, pp. 35–48, 2013.
- [18] David S Wishart. Metabolomics for investigating physiological and pathophysiological processes. *Physiological reviews*, Vol. 99, No. 4, pp. 1819–1875, 2019.
- [19] Rajesh Kumar, Mayukh Ghosh, Sandeep Kumar, and Minakshi Prasad. Single cell metabolomics: a future tool to unmask cellular heterogeneity and virus-host interaction in context of emerging viral diseases. *Frontiers in Microbiology*, Vol. 11, p. 1152, 2020.
- [20] Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome biology*, Vol. 21, No. 1, pp. 1–22, 2020.
- [21] David T Scadden. The stem-cell niche as an entity of action. *Nature*, Vol. 441, No. 7097, pp. 1075–1079, 2006.
- [22] Andreas E Moor and Shalev Itzkovitz. Spatial transcriptomics: paving the way for tissue-level systems biology. *Current opinion in biotechnology*, Vol. 46, pp. 126–133, 2017.
- [23] Hans-Werner Vohr. *Encyclopedia of immunotoxicology*. Springer Berlin Heidelberg, 2016.
- [24] Jaroonrut Prinyakupt and Charnchai Pluempiwiriyawej. Segmentation of white blood cells and comparison of cell morphology by linear and naïve bayes classifiers. *Biomedical engineering online*, Vol. 14, No. 1, pp. 1–19, 2015.
- [25] Alejandra Hernandez-Segura, Jamil Nehme, and Marco Demaria. Hallmarks of cellular senescence. *Trends in cell biology*, Vol. 28, No. 6, pp. 436–453, 2018.
- [26] Yusri Dwi Heryanto, Chin-Yi Cheng, Yutaka Uchida, Kazushi Mimura, Masaru Ishii, and Ryo Yamada. Integrated analysis of cell shape and movement in moving frame. *Biology Open*, 2021.
- [27] Chinyi Cheng, Yusri dwi Heryanto, and Ryo Yamada. The spherical harmonic based resolution increase and decrease method for cell mesh model with the vertex and face numbers consistency. In *2020 5th International Conference on*

Intelligent Informatics and Biomedical Sciences (ICIIBMS), pp. 225–228. IEEE, 2020.

- [28] Xiong Liu, Xueping Yu, Donald J Zack, Heng Zhu, and Jiang Qian. Tiger: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, Vol. 9, No. 1, pp. 1–7, 2008.
- [29] Jun Han, Meijun Chen, Yihan Wang, Boxuan Gong, Tianwei Zhuang, Lingyu Liang, and Hong Qiao. Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Scientific reports*, Vol. 8, No. 1, pp. 1–11, 2018.
- [30] Zobaer Akond, Munirul Alam, and Md Nurul Haque Mollah. Biomarker identification from rna-seq data using a robust statistical approach. *Bioinformation*, Vol. 14, No. 4, p. 153, 2018.
- [31] Kevin M Carter, Raviv Raich, William G Finn, and Alfred O Hero III. Fine: Fisher information nonparametric embedding. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 31, No. 11, pp. 2093–2098, 2009.
- [32] Julian A Gingold, Ed S Coakley, Jie Su, Dung-Fang Lee, Zerlina Lau, Hongwei Zhou, Dan P Felsenfeld, Christoph Schaniel, and Ihor R Lemischka. Distribution analyzer, a methodology for identifying and clustering outlier conditions from single-cell distributions, and its application to a nanog reporter rnai screen. *BMC bioinformatics*, Vol. 16, No. 1, p. 225, 2015.
- [33] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Vol. 28, No. 1, pp. 27–30, 2000.

第2章:サイトメトリーデータからの特微量抽出

岡田大瑚

2021年3月23日

概要

細胞集団プロファイルの情報を得るシングルセル解析技術として、サイトメトリーは最も頻繁に用いられてきた実験手法の1つである。この章では、サイトメトリーデータから細胞集団プロファイルの特微量を抽出する方法について、これまで用いられてきた手法を紹介する。サイトメトリーデータは分布として表現されるので、そこから何らかの特微量を抽出しない限り疾患など他のフェノタイプ情報との関連を調べることが難しい。近年では計算サイトメトリーとよばれる情報科学を用いたサイトメトリーデータの解析が注目されており、細胞サブセットを同定・定量するためのデータドリブンな解析手法も数多く提案されている。また一方で、数は少ないものの、細胞サブセットの存在を前提としないノンパラメトリックな手法も提案されている。この章ではサイトメトリーデータから細胞集団プロファイルの特微量を得るための手法について代表的なものについて詳述するとともに、現在までに提案されている手法の利点と欠点について議論する。

。

1 マニュアルゲーティングによる特微量抽出

大部分のサイトメトリーデータの解析研究では、特定の細胞サブセットの定量値を取得し、これを特微量として解析を行う。実験生物学におけるこれまでのサイトメトリー解析においては、ゲーティングと呼ばれる方法でこれを行ってきた。具体的には、多次元のマーカーを2マーカーずつ展開し密度分布を描く。そのプロット上で、あらかじめ定義された細胞集団を同定し、それらのサブセットの存在割合を定量するというプロセスである。このようにして特微量化されたデータと、ヒトにおける疾患の有無や実験動物における介入の有無などの関連を調べることで、特定の細胞サブセットの多寡とこれらの形質情報とを関連付けを行い、疾患や生命現象のメカニズムについての生物学的な解釈を行う。実験生物学における従来の解析では、研究者がソフトウェアを使いつつも目で見て細

胞集団を同定するマニュアルゲーティングが主に用いられてきた。マニュアルゲーティングは頻繁に用いられる方法であるが、同時に 2 つまでのマーカー数しか考慮できないことや研究者ごとにゲーティングの基準が異なるという欠点がある。

2 計算サイトメトリーによる特徴量抽出

近年ではサイトメトリーデータ解析に情報科学的なアプローチを適用し、データドリブンな情報抽出を行う計算サイトメトリーが盛んに研究されており、その中でデータドリブンに細胞集団の分類・定量を行う手法が多数提案されている [1]。このような手法を用いることで、従来のマニュアルゲーティングでは発見できなかった新規のサブセットを同定できる可能性があるだけでなく、サイトメトリーデータをデータドリブンに定量された有限個数の細胞サブセットの存在量で代表させることによる特徴量抽出も可能である。このような計算サイトメトリーの手法は数多く提案されているが、ここではクラスタリングアルゴリズムに基づく手法として SPADE 法を、パラメトリック統計モデルに基づく方法として EM アルゴリズムによる混合正規分布フィッティングを紹介する。

2.1 SPADE 法によるクラスタリング

情報科学分野では、多変量データに対するクラスタリングアルゴリズムが数多く提案されている。1 サンプルのサイトメトリーデータは細胞 \times マーカーの行列として表現できるため、計算サイトメトリーの分野においても、サイトメトリーデータの特徴を考慮したクラスタリングワークフローが多数提案されている。ここでは、その代表的な手法である SPADE 法について紹介する [2]。

SPADE 法は、代表的な計算サイトメトリーの手法の 1 つであり、多次元サイトメトリーデータを対象にデータドリブンな細胞集団のクラスタリングを行う。SPADE 法は、ダウンサンプリング、クラスタリング、アップサンプリングの 3 つの部分で構成される。最初のダウンサンプリングでは、希少なサブセットを検出するため、FACS データの重み付きリサンプリングを行う。ここではまず、細胞から n 個の細胞をピックアップする。次にピックアップした細胞の細胞 \times 細胞のマンハッタン距離行列に基づき、各細胞ごとに、どの程度細胞が密な場所に存在するかを示す Local Density value を計算し、この値に応じて重みを付けてリサンプリングを実施する。このようなルールでリサンプリングすることで細胞の数を減らしつつ希少なサブセットを検出しやすくしている。サイトメトリーデータが複数サンプルある場合は、それぞれに対してダウンサンプリングを行い、

残った細胞をマージしたデータセットを作成する。このデータセットに対してクラスタリングアルゴリズムを適用し、あらかじめ設定した K 個のクラスターに分類する。次に、これらのクラスターをノードとした最小全域木を構築することで、細胞のクラスタリングおよびその樹形図を作成する。最後に、アップサンプリングを行う。ここでは、各サンプルの全ての細胞に対して、最も近傍に位置するクラスターを割り当てる。結果として、ピックアップした全てのサンプルに対して樹形図上のサブセットへの分類と定量が行われる。各クラスターに所属する細胞の存在割合は、細胞集団プロファイルの特徴量として扱うことができる。このように、SPADE 法をはじめとする計算サイトメトリーの手法では、生命科学データの特性を生かしたクラスタリングの一連のワークフローが数多く提供されている。

2.2 パラメトリックモデルによるクラスタリング

サイトメトリーデータはランダムに採取した細胞集団に対して測定値を得るため、母確率分布からの標本を考えることができる。この性質に着目し、パラメトリック統計モデルを用いることで、細胞集団プロファイルの性質を調べることができる。計算サイトメトリーを用いた細胞サブセットの定量手法の中には、パラメトリックな統計モデルを母分布として仮定しそのパラメータの最尤推定を行い、推定された母確率分布に基づいて細胞サブセットの定量を行う手法が存在する。

ここではそのような手法の例として、EM アルゴリズムを用いたサイトメトリーデータへの混合正規分布フィッティングについて述べる。背後に仮定する統計モデルは混合正規分布モデルであり、そのパラメータ最尤推定をするための手法が EM アルゴリズムである。まず、文献 [3, 4, 5] の内容にを参考に、EM アルゴリズムの概要について紹介する。次に、文献 [3] を参考に、EM アルゴリズムを用いてサイトメトリーデータに混合正規分布をフィッティングし細胞集団の分類・定量を行う手法について順に紹介していく。

2.2.1 EM アルゴリズム

EM アルゴリズムは、隠れ変数のある統計モデルにおけるパラメーターの最尤推定において汎用的に使われる手法である。データを \mathbf{x} , 母分布のパラメータを $\boldsymbol{\theta}$ とおく。データ数を N 個とし、 i 番めのデータを x_i とおく。推定したい未知パラメータが $\boldsymbol{\theta}$ である。ここで、各 N 個のサンプルには観測できない隠れ変数が割り当てられている状況を考える。隠れ変数を長さ N のベクトルとして \mathbf{s} とおく。ここで、隠れ変数 s_i は i 番めのサンプルに割り当てられているクラスを表し、C 個のカテゴリのうちのいずれかをとるカテゴリ

変数である。

ここで、 θ の最尤推定は、以下の対数尤度で表される目的関数 $J(\theta)$ を最大化するような θ を計算する最適化問題として考えることができる。。

$$J(\theta) = \sum_i^N \log P(\mathbf{x}_i; \theta) = \sum_i^N \log \sum_{k=1}^C P(\mathbf{x}_i, s_k; \theta) \quad (1)$$

このような log-sum の形をした目的関数の最適化問題は一般に解析的に解くことができない。EM アルゴリズムはこのような最適化問題を解くための汎用的な反復解法である。まず、以下のような隠れ変数 s_i の事後分布を計算する。この事後分布 h_i は、長さ C の離散確率分布であり、以下のように定義される。

$$\mathbf{h}_i = P(s_i | \mathbf{x}_i; \theta) \quad (2)$$

ベイズの定理を用いて変形すると、このベクトルの k 番めの要素の値 (つまり、i 番めのサンプルがクラス k に属する事後確率) は以下のように表すことができる。

$$h_{ik} = P(s_i = k | \mathbf{x}_i; \theta) \quad (3)$$

$$= \frac{P(s_i = k, \mathbf{x}_i; \theta)}{\sum_{k=1}^C P(s_i = k, \mathbf{x}_i; \theta)} \quad (4)$$

ここで、暫定的なパラメータ θ^0 を与えると、全てのデータに対して $\mathbf{h}_i^0 = [h_{i1}^0 \cdots h_{iC}^0]$ を具体的に計算できる。これを用いて、目的関数 $J(\theta)$ はいかのよう変形できる。

$$J(\theta) = \sum_{i=1}^N \log \sum_{k=1}^C P(\mathbf{x}_i, s_k; \theta) = \sum_{i=1}^N \log \sum_{k=1}^C \frac{h_{ik}^0 P(\mathbf{x}_i, s_k; \theta)}{h_{ik}^0} \quad (5)$$

ここで、イエンセンの不等式を用いる。イエンセンの不等式とは任意の凸関数 f および確率分布 p について成り立つ以下のような不等式である。f が上に凸の場合は以下である。

$$\sum_{i=1}^{\infty} p_i f(x_i) \leq f\left(\sum_{i=1}^{\infty} p_i x_i\right) \quad (6)$$

f が下に凸の場合は以下のように不等号の向きが逆になる。

$$\sum_{i=1}^{\infty} p_i f(x_i) \geq f\left(\sum_{i=1}^{\infty} p_i x_i\right) \quad (7)$$

対数関数は上に凸の凸関数なので、式 5 にイエンセンの不等式を適用することで以下の不等式が成立する。

$$J(\boldsymbol{\theta}) = \sum_{i=1}^N \log \sum_{k=1}^C \frac{h_{ik}^0 P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} \geq \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} \quad (8)$$

次に、式 8 の不等式での右辺と左辺の差を計算する。 $\sum_{i=1}^N$ の中身に着目すると、

$$\begin{aligned} & \log \sum_{k=1}^C \frac{h_{ik}^0 P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} - \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} \\ &= \log \sum_{k=1}^C P(\mathbf{x}_i, s_k; \boldsymbol{\theta}) - \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} \\ &= \sum_{j=1}^C h_{ij}^0 \log \sum_{k=1}^C P(\mathbf{x}_i, s_k; \boldsymbol{\theta}) - \sum_{j=1}^C h_{ij}^0 \log \frac{P(\mathbf{x}_i, s_j; \boldsymbol{\theta})}{h_{ij}^0} \\ &= \sum_{j=1}^C h_{ij}^0 \left(\log \sum_{k=1}^C P(\mathbf{x}_i, s_k; \boldsymbol{\theta}) - \log \frac{P(\mathbf{x}_i, s_j; \boldsymbol{\theta})}{h_{ij}^0} \right) \\ &= \sum_{j=1}^C h_{ij}^0 \log \frac{h_{ij}^0 \sum_{k=1}^C P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{P(\mathbf{x}_i, s_j; \boldsymbol{\theta})} \\ &= \sum_{j=1}^C h_{ij}^0 \log \frac{h_{ij}^0}{h_{ij}^0} \\ &= KL(\mathbf{h}_i^0, \mathbf{h}_i) \end{aligned}$$

したがって、

$$J(\boldsymbol{\theta}) = \sum_{i=1}^N \log \sum_{k=1}^C \frac{h_{ik}^0 P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} = \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta})}{h_{ik}^0} + \sum_{i=1}^N KL(\mathbf{h}_i^0, \mathbf{h}_i) \quad (9)$$

また、ユーザーが設定した初期値 $\boldsymbol{\theta}_0$ の下での目的関数の初期値 $J(\boldsymbol{\theta}_0)$ は、

$$J(\boldsymbol{\theta}^0) = \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta}^0)}{h_{ik}^0} + \sum_{i=1}^N KL(\mathbf{h}_i^0, \mathbf{h}_i^0) = \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log \frac{P(\mathbf{x}_i, s_k; \boldsymbol{\theta}^0)}{h_{ik}^0}$$

(10)

ここで、 $KL(\mathbf{h}_i^0, \mathbf{h}_i^0) = 0$ を利用した。したがって、以下の式が成り立つ。

$$\begin{aligned} & J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^0) \\ &= \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log P(\mathbf{x}_i, s_k; \boldsymbol{\theta}) + \sum_{i=1}^N KL(\mathbf{h}_i^0, \mathbf{h}_i) - \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log P(\mathbf{x}_i, s_k; \boldsymbol{\theta}^0) \end{aligned}$$

ここで、既に値が決まっている $\boldsymbol{\theta}^0$ によって定められる $\boldsymbol{\theta}$ の関数、 $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0)$ を以下のように定義する。

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0) = \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log P(\mathbf{x}_i, s_k; \boldsymbol{\theta}) \quad (11)$$

$J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^0)$ は、Q 関数を用いて以下のように表すことができる。

$$J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^0) = Q(\boldsymbol{\theta}^0, \boldsymbol{\theta}) - Q(\boldsymbol{\theta}^0, \boldsymbol{\theta}^0) + KL(\mathbf{H}^0, \mathbf{H}) \quad (12)$$

ここで、初期値 $\boldsymbol{\theta}^0$ を、 $\boldsymbol{\theta}^1$ に更新することを考える。KL ダイバージェンスは常に 0 以上の値を取るため、 $Q(\boldsymbol{\theta}^1; \boldsymbol{\theta}^0) - Q(\boldsymbol{\theta}^0; \boldsymbol{\theta}^0) \geq 0$ ならば $J(\boldsymbol{\theta}^1) \geq J(\boldsymbol{\theta}^0)$ が成り立つ。したがって、初期値の元での Q 関数を大きくするように $\boldsymbol{\theta}^1$ を定めれば、目的関数も大きくなる。以上より、目的関数の代わりに Q 関数の最適化を行う反復解法アルゴリズムを構成できる。

1. パラメータの初期値 $\boldsymbol{\theta}^0$ を定める。

2. すべてのデータ i およびすべての隠れ変数のクラス k について h_{ik}^0 を計算する。これは、 $h_{ik}^0 = P(s_i = k | \mathbf{x}_i; \boldsymbol{\theta}^0) = \frac{P(s_i = k, \mathbf{x}_i; \boldsymbol{\theta}^0)}{\sum_{k=1}^C P(s_i = k, \mathbf{x}_i; \boldsymbol{\theta}^0)}$ で計算できる。このステップは、隠れ変数に関する期待値を計算していることに相当するので Expectation ステップ (E ステップ) と呼ばれる。

3. $Q(\theta; \theta^0)$ を最大にする θ を求め、 θ^1 とする。このステップは、Q 関数の最大化をしているので Maximization ステップ (M ステップ) と呼ばれる。

4, 全ステップで求めた θ^1 を θ^0 とし、ステップ 2 に戻る。

上記の反復計算を繰り返すことで、目的関数は増加 (非減少) していく。一定の基準をみたした段階で反復計算を終了し、その時のパラメータ θ^1 を θ の推定値、 \mathbf{h}_i をデータ \mathbf{x}_i が所属する隠れ変数 s_i の確率分布として採用する。EM アルゴリズムは、目的関数の最適化が難しい一方で、Q 関数の最適化が容易に計算できる時に効果を発揮する。これは、隠れ変数が既知の場合には問題が簡単になるような状況である。

生命科学において EM アルゴリズムが用いられる例として、統計遺伝学におけるハプロタイプ頻度の推定問題を示す。まず最初に、アレル頻度の推定を考える。N 人を対象にある多型のジェノタイプを調べた結果遺伝子型が AA, Aa, aa の人数がそれぞれ n_{AA}, n_{Aa}, n_{aa} 人とする。この時、このサンプル中においては、A のアレルは $2n_{AA} + n_{Aa}$ 個、a のアレルは $2n_{aa} + n_{Aa}$ 個存在する。A のアレル頻度を p とし、観測データが 2 項分布に従っているとすると、アレル頻度 p の推定値は $\frac{2n_{AA} + n_{Aa}}{2N}$ となる。次にディプロタイプデータからハプロタイプ頻度を推定する問題を考える。ある座位での遺伝子型 AA, Aa, aa、もう 1 つの座位の遺伝子型を BB, Bb, bb とすると、ディプロタイプはこれらの組み合わせなので 9 通り存在する。ハプロタイプ AB, Ab, aB, ab の本数は、それぞれのディプロタイプを持つ人数を用いて以下のように表される。

$$N_{AB} = \frac{2n_{AABB} + n_{AABb} + n_{AaBB} + rn_{AaBb}}{2N}$$

$$N_{Ab} = \frac{2n_{AAbb} + n_{AABb} + n_{Aabb} + (1-r)n_{AaBb}}{2N}$$

$$N_{aB} = \frac{2n_{aaBB} + n_{AaBB} + n_{aaBb} + (1-r)n_{AaBb}}{2N}$$

$$N_{ab} = \frac{2n_{aabb} + n_{Aabb} + n_{aaBb} + rn_{AaBb}}{2N}$$

ここで、AaBb のディプロタイプを持つ個体のうち、AB/ab のハプロタイプの組み合わせの割合を r 、Ab/aB のハプロタイプの組み合わせの割合を $1 - r$ とした。この時、 r の値はわからないため、ハプロタイプ頻度 $f_{AB}, f_{Ab}, f_{aB}, f_{ab}$ の推定はアレル頻度推定問題のように 1 回の計算で求めることはできない。しかしながら、AaBb のディプロタイプを持つ個体のうち、AB/ab と Ab/aB のハプロタイプの組み合わせを持つ人数がわかれば、ハ

プロタイプ頻度は簡単に推定できる。従って、AaBb の個体がどちらのハプロタイプの組み合わせをとるかを 2 カテゴリの隠れ変数とした EM アルゴリズムを適用することでこの問題を解くことができる。

今回推定するパラメータ θ は、4 種類のハプロタイプ頻度 $[f_{AB}, f_{Ab}, f_{aB}, f_{ab}]$ である。観測できるデータ x_i は、 i 番めの個体のディプロタイプであり、9 通りのうちのどれかをとる。また、隠れ変数 s_i は各個体のハプロタイプの組み合わせを表す。これは、AB/AB, AB/Ab, AB/ab, AB/aB, Ab/Ab, Ab/aB, Ab/ab, aB/aB, aB/ab の合計 10 通りのパターンがある。パラメーター最尤推定のための目的関数 $J(\theta)$ は以下のような形式になっており、EM アルゴリズムが適用できる形である。

$$J(\mathbf{f}) = \sum_i^N \log P(x_i; \mathbf{f}) = \sum_i^N \log \sum_{k=1}^{C=10} P(x_i, s_i = k; \mathbf{f}) \quad (13)$$

以下で、EM アルゴリズムの各ステップを具体的に見ていく。

ステップ 1: ハプロタイプ頻度の初期値 $\mathbf{f}^0 = [f_{AB}^0, f_{Ab}^0, f_{aB}^0, f_{ab}^0]$ を設定する。

ステップ 2: \mathbf{h}_i^0 を計算する。これは各個体が 9 通りのディプロタイプのうちどれであるかによって異なる。AaBb 以外の個体では、ディプロタイプが定まるとハプロタイプは一意に定まる。例えば、AABb の個体では、ハプロタイプの組み合わせは AB/Ab になるため、 \mathbf{h}_i^0 は AB/Ab で 1, それ以外で 0 の離散確率分布となる。AaBb のディプロタイプの個体では、AB/ab と Ab/aB の 2 パターンが存在し、これらのパターンをとる事後確率は

$$h_{i,AB/ab}^0 = P(s_i = AB/ab | x_i = AaBb; \mathbf{f}^0) \quad (14)$$

$$= \frac{P(s_i = AB/ab, x_i = AaBb; \mathbf{f}^0)}{\sum_k^C P(s_i = k, x_i = AaBb; \mathbf{f}^0)} \quad (15)$$

$$= \frac{f_{AB}^0 f_{ab}^0}{f_{AB}^0 f_{ab}^0 + f_{Ab}^0 f_{aB}^0} \quad (16)$$

同様に、

$$h_{i,Ab/aB}^0 = \frac{f_{Ab}^0 f_{aB}^0}{f_{AB}^0 f_{ab}^0 + f_{Ab}^0 f_{aB}^0} \quad (17)$$

と表現できる。これらは、ステップ 1 で設定されている \mathbf{f}^0 から具体的に計算できる。

ステップ3: $Q(\mathbf{f}; \mathbf{f}^0)$ を最大化する \mathbf{f} を計算する。

$$\begin{aligned}
Q(\mathbf{f}; \mathbf{f}^0) &= \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log P(x_i, s_k; \mathbf{f}) \\
&= N_{AABB} \log f_{AB}^2 + N_{AABb} \log(2f_{AB}f_{Ab}) + N_{AAbb} \log f_{Ab}^2 + \\
&\quad N_{aaBB} \log f_{aB}^2 + N_{aaBb} \log(2f_{aB}f_{ab}) + N_{aabb} \log f_{ab}^2 + \\
&\quad N_{AaBB} \log(2f_{AB}f_{aB}) + N_{Aabb} \log(2f_{Ab}f_{ab}) + \\
&\quad N_{AaBb} \left(\frac{f_{AB}^0 f_{ab}^0}{f_{AB}^0 f_{ab}^0 + f_{Ab}^0 f_{aB}^0} \log(2f_{AB}f_{ab}) + \frac{f_{Ab}^0 f_{aB}^0}{f_{AB}^0 f_{ab}^0 + f_{Ab}^0 f_{aB}^0} \log(2f_{Ab}f_{aB}) \right)
\end{aligned}$$

このうち、変数である \mathbf{f} に関する部分に着目して整理すると、

$$Q(\mathbf{f}; \mathbf{f}^0) = \log(f_{AB}^{\hat{N}_{AB}} f_{Ab}^{\hat{N}_{Ab}} f_{aB}^{\hat{N}_{aB}} f_{ab}^{\hat{N}_{ab}}) + const \quad (18)$$

ここで、

$$\begin{aligned}
\hat{N}_{AB} &= 2N_{AABB} + N_{AABb} + N_{AaBB} + r_0 N_{AaBb} \\
\hat{N}_{Ab} &= 2N_{AAbb} + N_{AABb} + N_{Aabb} + (1 - r_0) N_{AaBb} \\
\hat{N}_{aB} &= 2N_{aaBB} + N_{AaBB} + N_{aaBb} + (1 - r_0) N_{AaBb} \\
\hat{N}_{ab} &= 2N_{aabb} + N_{Aabb} + N_{aaBb} + r_0 N_{AaBb} \\
r_0 &= \frac{f_{AB}^0 f_{ab}^0}{f_{AB}^0 f_{ab}^0 + f_{Ab}^0 f_{aB}^0}
\end{aligned}$$

と置いている。

ここで、式18は単に多項分布のパラメータのを最尤推定する際の対数尤度関数なので、これを最大化するような $\hat{\mathbf{f}}$ は、 $\hat{N}_{AB} + \hat{N}_{Ab} + \hat{N}_{aB} + \hat{N}_{ab} = 2N$ より、以下のように簡単に計算できる。

$$\hat{f}_{AB} = \frac{\hat{N}_{AB}}{2N} \quad (19)$$

$$\hat{f}_{Ab} = \frac{\hat{N}_{Ab}}{2N} \quad (20)$$

$$\hat{f}_{aB} = \frac{\hat{N}_{aB}}{2N} \quad (21)$$

$$\hat{f}_{ab} = \frac{\hat{N}_{ab}}{2N} \quad (22)$$

ステップ4: 上記で求めた $\hat{\mathbf{f}}$ を \mathbf{f}^0 として、ステップ2に戻る。

上記の反復計算を繰り返すことで、ハプロタイプ頻度の最尤推定値を計算することができる。

2.2.2 混合正規分布フィッティング

上記のEMアルゴリズムを用いて、サイトメトリーデータのマーカー発現値の分布の母分布に混合正規分布を仮定しそのパラメータ推定を行うことで、細胞集団の分類・定量を行う手法について紹介する。ここでは、それぞれのマーカーごとに1次元の混合正規分布を当てはめて陰性・陽性を定めるカットオフ値をデータドリブンに算出する方法を示す。これまでの例と同様に、N個の細胞に対してマーカータンパク質Aの発現値 x_i がデータとして得られているとする。ここに混合正規分布を当てはめる。

ここで $r_1 \cdots r_C$ はそれぞれの正規分布の混合割合を表しており、 $\sum^C r_k = 1$ である。推定されるパラメータ $\boldsymbol{\theta}$ は、混合正規分布中の $r_1 \cdots r_C, \mu_1 \cdots \mu_C, V_1 \cdots V_C$ の3C個である。パラメータを最尤推定するための目的関数はこれまでと同様に以下である。

$$J(\boldsymbol{\theta}) = \sum_i^N \log P(x_i; \boldsymbol{\theta}) \quad (23)$$

ここで、各データ点がどの正規分布由来であるかが特定されれば、これらのパラメータは通常の正規分布のときと同様に計算することができる。そこで、各データ点がどの正規分布からのサンプルかを隠れ変数とするEMアルゴリズムでこの問題を解くことができる。これまでと同様、目的関数を以下のようにおく。

$$J(\boldsymbol{\theta}) = \sum_i^N \log P(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_i^N \log \sum_{k=1}^C P^k(x_i, s_i = k; \boldsymbol{\theta}) \quad (24)$$

ここで、 $P^k(x)$ は混合正規分布に含まれる k 番目の正規分布の確率密度関数である。

EMアルゴリズムの各ステップを具体的に見ていく。

ステップ1: パラメータの初期値 $\boldsymbol{\theta}^0$ を定める。

ステップ2: すべてのデータ i およびすべての隠れ変数のクラス k について h_{ik}^0 を計算

する。

$$\begin{aligned}
h_{ik}^0 &= P(s_i = k \mid x_i; \boldsymbol{\theta}^0) \\
&= \frac{P(s_i = k, x_i; \boldsymbol{\theta}^0)}{\sum_{k=1}^C P(s_i = k, x_i; \boldsymbol{\theta}^0)} \\
&= \frac{P(s_i = k; \boldsymbol{\theta}^0) P(x_i \mid s_i = k; \boldsymbol{\theta}^0)}{\sum_{k=1}^C P(s_i = k; \boldsymbol{\theta}^0) P(x_i \mid s_i = k; \boldsymbol{\theta}^0)} \\
&= \frac{r_k N(x_i; \mu_k^0, V_k^0)}{\sum_{k=1}^C r_k N(x_i; \mu_k^0, V_k^0)}
\end{aligned}$$

ステップ 3: $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0)$ を最大化する $\boldsymbol{\theta}$ を計算する。

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0) &= \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log P(x_i, s_i = k; \boldsymbol{\theta}) \\
&= \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log(r_k N(x_i; \mu_k^0, V_k^0))
\end{aligned}$$

ここで、 $\boldsymbol{\theta}$ は $r_1 \cdots r_C, \mu_1 \cdots \mu_C, V_1 \cdots V_C$ なので、これらで偏微分することで Q 関数を最大化するようなパラメータの値を計算する。ここで、 $\sum^C r_k = 1$ の制約条件があるので、ラグランジュの未定乗数法より以下のコスト関数 $R(\boldsymbol{\theta}; \boldsymbol{\theta}^0)$ を目的関数に用いる。

$$\begin{aligned}
R(\boldsymbol{\theta}; \boldsymbol{\theta}^0) &= Q(\boldsymbol{\theta}; \boldsymbol{\theta}^0) + \lambda \left(\sum_k r_k - 1 \right) \\
&= \sum_{i=1}^N \sum_{k=1}^C h_{ik}^0 \log(r_k N(x_i; \mu_k^0, V_k^0)) + \lambda \left(\sum_{k=1}^C r_k - 1 \right)
\end{aligned}$$

これを解いて、 r_k の最適値は以下のように定まる。

$$\hat{r}_k = \frac{1}{N} \sum_i^N h_{ik}^0 \tag{25}$$

これはデータ点が k 番めの正規分布に所属している事後確率の、全 N 個のデータ点における平均となっている。

同様に、 $\frac{\partial R}{\partial \mu_k} = 0$ より、 μ_k の最適値を計算すると、

$$\hat{\mu}_k = \frac{\sum_i^N h_{ik}^0 x_i}{\sum_i^N h_{ik}^0} \tag{26}$$

これも、それぞれのサンプルの各クラスの事後確率で重みづけをして計算した平均値に相当している。

同様に、 $\frac{\partial R}{\partial \sigma_k} = 0$ より、 σ_k の最適値を計算すると、

$$\hat{\sigma}_k = \sqrt{\frac{\sum_i^N h_{ik}^0 (x_i - \mu_k)^2}{\sum_i^N h_{ik}^0}} \quad (27)$$

これも、 μ の場合と同様、それぞれのサンプルの各クラスの事後確率で重みづけをして計算した標準偏差に相当している。

ステップ 4: ステップ 3 で計算した $\hat{\theta} = [\hat{r}_1 \dots \hat{r}_k, \hat{\mu}_1 \dots \hat{\mu}_k, \hat{\sigma}_1 \dots \hat{\sigma}_k]$ を $\hat{\theta}^0$ にセットし、ステップ 2 に戻る。

このような EM アルゴリズムを適用し、一定の基準を満たすまで反復することによって混合正規分布を当てはめることができる。この、データ点の母分布として推定された混合正規分布に基づいて、陰性とみなす正規分布に所属する事後確率と陽性とみなす正規分布に所属する事後確率がちょうど等しくなる値を陰性・陽性を定めるカットオフ値として採用することで、データドリブンに細胞集団の分類と定量を行うことができる。例えばマーカー数が m 個で、 $C=2$ の混合正規分布を当てはめる場合には、各マーカーの陰性・陽性的パターンによって 2^m 通りの細胞サブセットの割合を定量できるため、これらは細胞集団プロファイルの特徴量として利用可能である。母分布の統計モデルとしては、混合正規分布の代わりに混合 t 分布も提案されている [6]。

3 ノンパラメトリックモデルを用いたデータドリブンな特徴量抽出

混合正規分布モデルによるサイトメトリーデータからの特徴量抽出はパラメトリックな統計モデルを用いた手法であったが、サイトメトリーデータからノンパラメトリックに特徴量抽出を行う研究も行われている。ノンパラメトリックな統計モデルとは、パラメーターの数を気にすることなしにデータの従う確率分布をモデル化する手法である。計算サイトメトリーにおいて提案されているノンパラメトリックな特徴量抽出は、情報理論に基づいた分布間異同定量と多次元尺度構成法 (Multi-dimensional scaling: MDS) を組み合わせて行うものである。まず、最初の STEP として各サイトメトリーデータからノンパラメトリックに母分布推定を行う。ここでは、 k 最近傍法 (k -nearest neighbor) やカーネル密度推定法を用いて、あらかじめ設定したグリッドに対して確率密度を推定する。次

に、情報理論に基づく分布間距離の指標を用いて分布間距離行列を作成する。ここで用いられる分布間距離の指標としては Kullback–Leibler (KL) ダイバージェンスに基づいた指標が良く用いられる。そのような指標の例としては以下で示すような Jensen-Shannon divergence がある。

$$JS(p||q) = \frac{1}{2}(KL(p\|\frac{p+q}{2}) + KL(q\|\frac{p+q}{2})) \quad (28)$$

次に、ここで作成したサンプル \times サンプルの分布間距離行列に、多次元尺度構成法 (MDS) を適用する。多次元尺度構成法は距離行列をインプットとしてとり、その距離関係を最もよく反映するようなユーグリッド座標を返す古典的統計手法である。MDS は分布間距離行列の情報をできるだけ保つように各サンプルに MDS 座標の値を付与する。MDS 座標は寄与率とともに得られるため、寄与の大きい座標軸の値はサイトメトリーデータのデータドリブンな特徴量と捉えることができる。このような手法に関する文献としては、文献 [7, 8] などがある。

4 これまでの手法の課題と情報幾何の応用

細胞集団プロファイルの比較は、これまでほとんどの場合、細胞のサブセットの存在量に基づいて行われてきた。例えば、血中のリンパ球を対象とした解析の場合、ゲーティングを用いてナチュラルキラー細胞やヘルパー T 細胞といった既知の細胞サブセットがどれだけあるかを数え、疾患との関連を調べる。このような解析では、細胞サブセットに関しての事前情報をあらかじめ知っている必要がある。SPADE などのクラスタリングアルゴリズムやパラメトリック統計モデルを用いることで、サイトメトリーデータに含まれる細胞を自動で複数種類のサブセットに分類することができる。このような場合でも何種類のサブセットに分類するかの数は解析者が指定する必要がある。

しかしながら、サンプル細胞集団に含まれる細胞が実際に何種類の細胞サブセットに分類されるべきなのは不明であり、最近でも新規の細胞サブセットが報告されることは數多くある。その他にも、このようなクラスタリングに基づく手法のほとんどはそれぞれの手法ごとにいくつもの仮定を置いている。例えば、混合正規分布フィッティングでは、マーカー発現量の分布が混合正規分布になるという強い仮定を置いているが、そのような仮定が常に適切であるとは限らない。

一方でノンパラメトリックな方法はそのような細胞サブセットに関する事前知識や仮定を用いずにサイトメトリーデータ同士を比較することができる。代表的なノンパラメト

表1 それぞれの手法の解釈のしやすさと情報ロスの少なさ

	解釈のしやすさ	情報ロスの少なさ
ゲーティング	○	×
クラスタリング	○	△
MDS	×	△
期待される新手法	○	○

リックな手法である MDS を用いた方法では、サイトメトリーデータの分布間の距離情報のみから座標を与えるため、事前の仮定による影響を受けずデータドリブンに特徴量を与えることができる点が大きなメリットである。

しかしながら、このようなノンパラメトリックな方法では、特徴量として得られた座標とともに分布との対応関係が不明なため得られた座標の生物学的解釈が難しいという欠点がある。また、確率分布同士の距離関係を完全に満たすようなユークリッド座標が常に存在するとは限らず、そのような場合には情報のロスが発生してしまう。

表1は、古典的なゲーティング、パラメトリックモデルによる方法、クラスタリングによる方法それぞれでの生物学的解釈のしやすさ、情報ロスの少なさにおける強みと弱みをまとめたものである。両者を両立させるような手法は現在存在せず、新規手法の開発が期待される。

一方で、数学の分野では、確率分布に座標を与えてその幾何学的な性質を研究する情報幾何学が研究されている。そこで、情報幾何学の知見を用いて細胞集団プロファイルの確率分布に座標を与えることで、前述の課題を解決するような新規の手法を開発できるのではないかと着想した。次章では、そのような背景を踏まえ、情報幾何学の基本的な概念についてまとめる。

参考文献

- [1] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, Vol. 16, No. 7, p. 449, 2016.
- [2] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data

with spade. *Nature biotechnology*, Vol. 29, No. 10, p. 886, 2011.

- [3] 上田修功石井健一郎. 続・わかりやすいパターン認識 教師なし学習入門. 続・わかりやすいパターン認識 教師なし学習入門. オーム社.
- [4] 平井有三. はじめてのパターン認識. 森北出版株式会社.
- [5] 山田亮. 遺伝統計学の基礎 r による遺伝因子解析・遺伝子機能解析. オーム社.
- [6] Kenneth Lo, Ryan Remy Brinkman, and Raphael Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, Vol. 73, No. 4, pp. 321–332, 2008.
- [7] Kevin M Carter, Raviv Raich, William G Finn, and Alfred O Hero III. Fine: Fisher information nonparametric embedding. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 31, No. 11, pp. 2093–2098, 2009.
- [8] Julian A Gingold, Ed S Coakley, Jie Su, Dung-Fang Lee, Zerlina Lau, Hongwei Zhou, Dan P Felsenfeld, Christoph Schaniel, and Ihor R Lemischka. Distribution analyzer, a methodology for identifying and clustering outlier conditions from single-cell distributions, and its application to a nanog reporter rnai screen. *BMC bioinformatics*, Vol. 16, No. 1, p. 225, 2015.

第3章:シングルセル発現解析のための情報幾何

岡田大瑚

2021年3月23日

概要

この章では、情報幾何の基礎については文献 1-6 の内容をもとに述べる。まず、リーマン多様体について紹介したうえで、多様体上に θ 座標および凸関数 $\psi(\boldsymbol{\theta})$ を設定することで、点と点同士の距離関係を定める Bregman ダイバージェンスが導入でき、ここからリーマン計量が定まることを示す。さらに、凸関数に基づいて θ 座標と 1 対 1 対応する η 座標を導入すると、これらの 2 つの座標系は双対かつ平坦であり、幾何学的に都合のいい性質を持つことを示す。次に、情報幾何で主な考察対象となる指数型分布族 $p = \exp(C(x) + \sum F_i \theta_i - \psi(\boldsymbol{\theta}))$ においては、式中の θ を分布に付与される座標、 $\psi(\boldsymbol{\theta})$ をダイバージェンスを定めるポテンシャル関数、とみなして上記の議論を展開することで、多様体上の内積行列として Fisher 情報行列、Bregman ダイバージェンスとして KL ダイバージェンスが導かれることが示す。最後に、実際のデータ解析の手法を例に、双対平坦座標系の性質を用いてデータ解析のアルゴリズムが幾何学的に解釈できることを見る。

1 双対平坦座標系

1.1 リーマン多様体、計量、接続、測地線

多様体とは、局所的な座標を張り合わせてできる空間である。n 個のパラメータを用いてその多様体上の 1 点を指定できるとき、これを n 次元多様体と呼ぶ。ここでは、情報幾何で主に取り扱うリーマン多様体について述べる。リーマン多様体では、座標 $\boldsymbol{\theta}$ を持つ点 P とその近傍の点 Q (座標は $\boldsymbol{\theta} + d\boldsymbol{\theta}$) の距離の 2 乗が以下で与えられる。

$$ds^2 = \sum g_{ij}(\boldsymbol{\theta}) d\theta_i d\theta_j = d\boldsymbol{\theta}^T \mathbf{G} d\boldsymbol{\theta} \quad (1)$$

ここでの $\mathbf{G} = g_{ij}$ は正定値行列であり、これをリーマン計量と呼ぶ。一般に、多様体上の点ごとに計量は違ってよいが、ユークリッド空間では全ての点において \mathbf{G} が単位行列に

なっている。リーマン計量は内積を定める行列でもあり、 θ_i の基底 e_i と θ_j の基底 e_j との内積 $\langle e_i, e_j \rangle$ は g_{ij} になる。多様体の構造を確定させるためには、各点での計量に加えて接続を与える必要がある。リーマン多様体は一般に曲がっているので、ある点での接空間と、そこから少し離れた点での接空間は一般にずれている。このずれは、それぞれの接空間を対応づけるための接続係数を用いて表現される。接続係数が全て 0 になるような座標系を取れる空間を平坦な空間と呼ぶ。平坦な空間では、多様体上のまっすぐな線である測地線が、座標上の直線としてあらわすことができる。つまり、 θ 座標系がこの平坦な座標系である場合、点 P と点 Q を結ぶ測地線は、ユーグリッド空間と同様に変数 t を用いて $(1-t)\theta^P + \theta^Q$ と表すことができる。また、 θ 座標系が平坦な座標系であるとき、 $A\theta + b = 0$ の線形束縛で定義される部分空間を平坦な部分空間という。このように平坦な空間・座標系は、ユークリッド空間における平面の一般化したものといえるので、計算を行う際に扱いが楽になる。一方で、式 1において内積行列 G が正定値行列にならないような多様体を擬リーマン多様体とよび、物理学における一般相対論で重要なローレンツ多様体などがこれに該当する。

1.2 Bregman ダイバージェンスと双対平坦座標系

多様体上に凸関数が定義されると、その凸関数に基づき双対平坦性と呼ばれる良い性質を持つ 2 種類の座標系を定めることができ、さらに空間を定めるリーマン計量もそこから導出されることを示す。まず、多様体上に凸関数 $\psi(\theta)$ が定められている。ここで、 θ 座標系はこの多様体上に設定される 1 つめの座標系である。次に、この多様体上に凸関数に基づいて θ 座標系での Bregman ダイバージェンスを定義する。ダイバージェンスは距離を非線形、非対称に拡張したものであり、Bregman ダイバージェンスもそのようなダイバージェンスの指標のうちの 1 つである。Bregman ダイバージェンスは、凸関数とその接平面とのずれによって定義され(図 1)、以下の式で定義される。

$$D(P||Q) = \psi(\theta^P) - \psi(\theta^Q) - \langle \nabla \psi(\theta^Q), \theta^P - \theta^Q \rangle \quad (2)$$

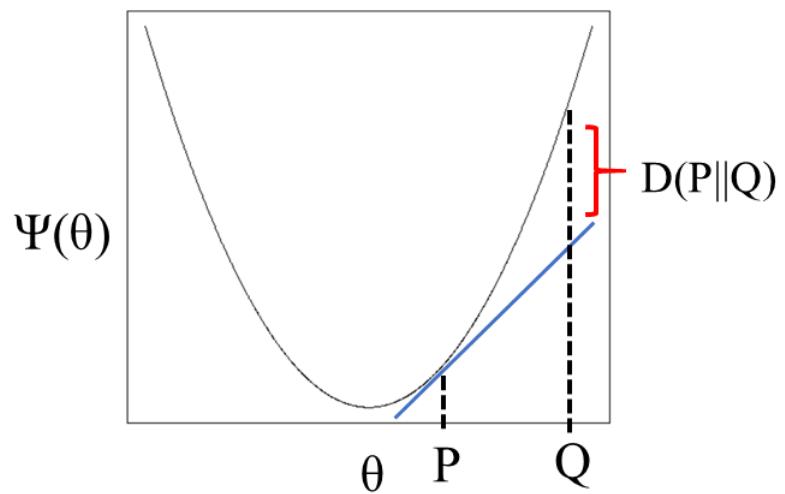


図 1 Bregman ダイバージェンス

座標が $\boldsymbol{\theta}$ である点とそこから微小な距離離れた $\boldsymbol{\theta} + d\boldsymbol{\theta}$ の間の Bregman ダイバージェンスを考える。

$$D(\boldsymbol{\theta} || \boldsymbol{\theta} + d\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta} + d\boldsymbol{\theta}) + \langle \nabla \psi(\boldsymbol{\theta} + d\boldsymbol{\theta}), d\boldsymbol{\theta} \rangle \quad (3)$$

この式に含まれるそれぞれの項をテイラー展開し、2次の項までを考える。

$$\psi(\boldsymbol{\theta} + d\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) + \sum_i d\theta_i \frac{\partial \psi}{\partial \theta_i} + \frac{1}{2} \sum_{ij} \frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j} d\theta_i d\theta_j \quad (4)$$

$$\langle \nabla \psi(\boldsymbol{\theta} + d\boldsymbol{\theta}), d\boldsymbol{\theta} \rangle = \sum_k d\theta_k \frac{\partial \psi}{\partial \theta_k} + \sum_{ij} \frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j} d\theta_i d\theta_j \quad (5)$$

これらを (3) 式に代入し、計算すると

$$D(\boldsymbol{\theta} || \boldsymbol{\theta} + d\boldsymbol{\theta}) = \frac{1}{2} \sum_{ij} \frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j} d\theta_i d\theta_j \quad (6)$$

微小な距離 ds^2 はダイバージェンスの 2 倍と一致するため式 (1) と比較すると、内積行列 g_{ij} と $\psi(\boldsymbol{\theta})$ の関係が得られる。

$$g_{ij} = \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \xi_i \partial \xi_j} \quad (7)$$

次に、以下のように η 座標を定義すると、この θ 座標と η 座標は双対平坦性と呼ばれる都合のいい性質を持つことが知られている。 η 座標は以下のように定義される。

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}) \quad (8)$$

この変換は Legendre 変換と呼ばれ、 $\psi(\boldsymbol{\theta})$ が凸関数の場合 1 対 1 対応する。そのため、 θ が座標系として使用できる場合、 η も座標系として使用できる。次に、多様体上の η 座標の関数 $\psi^\eta(\boldsymbol{\eta})$ を以下のように定義する。

$$\psi^\eta(\boldsymbol{\eta}) = \langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle - \psi(\boldsymbol{\theta}) \quad (9)$$

ここで、 $\psi^\eta(\boldsymbol{\eta})$ を η_i で偏微分すると、

$$\begin{aligned}
\frac{\partial \psi^\eta(\boldsymbol{\eta})}{\partial \eta_i} &= \theta_i + \eta_i \frac{\partial \theta_i}{\partial \eta_i} - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \eta_i} \\
&= \theta_i + \eta_i \frac{\partial \theta_i}{\partial \eta_i} - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \\
&= \theta_i
\end{aligned}$$

したがって、

$$\frac{\partial \psi^\eta(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \boldsymbol{\theta}$$

このように、 θ 座標と η 座標は、以下に示すような対称な関係で結ばれている。

$$\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\eta} \tag{10}$$

$$\frac{\partial \psi^\eta(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \boldsymbol{\theta} \tag{11}$$

この関係を双対性と呼ぶ。凸関数が設定された多様体上には、双対な 2 種類の座標系を定めることができ、これらは種々の性質において対称な関係性を持つ。また、ここでの θ 座標系は平坦な座標系であり、Legendre 変換によって生成される η 座標系もまた平坦な座標系である。このような双対かつ平坦な座標系を双対平坦座標系と呼び、そのような座標系が設定できる空間を双対平坦空間と呼ぶ。

また、Bregman ダイバージェンスもこれらの双対座標を用いて以下のように表すことができる。

$$\begin{aligned}
D(P||Q) &= \psi(\boldsymbol{\theta}^P) - \psi(\boldsymbol{\theta}^Q) - \langle \nabla \psi(\boldsymbol{\theta}^Q), \boldsymbol{\theta}^P - \boldsymbol{\theta}^Q \rangle \\
&= \psi(\boldsymbol{\theta}^P) - \psi(\boldsymbol{\theta}^Q) - \langle \boldsymbol{\eta}^Q, \boldsymbol{\theta}^P - \boldsymbol{\theta}^Q \rangle \\
&= \psi(\boldsymbol{\theta}^P) - (\langle \boldsymbol{\theta}^Q, \boldsymbol{\eta}^Q \rangle - \psi^\eta(\boldsymbol{\eta}^Q)) - \langle \boldsymbol{\eta}^Q, \boldsymbol{\theta}^P - \boldsymbol{\theta}^Q \rangle \\
&= \psi(\boldsymbol{\theta}^P) + \psi^\eta(\boldsymbol{\eta}^Q) - \langle \boldsymbol{\theta}^P, \boldsymbol{\eta}^Q \rangle
\end{aligned}$$

したがって、

$$D(P||Q) = \psi(\boldsymbol{\theta}^P) + \psi^\eta(\boldsymbol{\eta}^Q) - \langle \boldsymbol{\theta}^P, \boldsymbol{\eta}^Q \rangle \tag{12}$$

1.3 双対平坦座標系の性質

双対平坦座標系は、幾何学的に都合の良い性質を持ち合わせている。ここでは、双直行性、拡張ピタゴラスの定理、拡張射影定理の3つについて述べる。

1.3.1 双直交性

θ_i における基底ベクトルを e_i , η_i における基底ベクトルを e'_i とおく。一般に、 e_i と e_j は直交ではない。なぜなら、これらの基底ベクトル同士の内積行列は式1におけるG行列であり、これは一般には対角行列ではないからである。これは、 η 座標系においての基底ベクトル e'_i と e'_j の関係においても同様である。しかしながら、双対平坦座標系においては、 e_i と e'_j は直交する、つまり、 $\langle e_i, e'_j \rangle = 0$ が成り立つという性質がある。これを双直交性と呼び、双対平坦座標系における重要な性質である。

1.3.2 拡張ピタゴラスの定理

双直交性に基づき、双対平坦座標系における直角三角形においても、ユーグリッド空間におけるピタゴラスの定理と同等の性質が成り立つ。PとQが測地線、QとRが η 座標上の測地線（双対測地線）で結ばれており、それらが直交しているとする。この時、以下が成り立つ。

$$D(P||R) = D(P||Q) + D(Q||R) \quad (13)$$

これは、以下の手順で証明できる。式13の右辺-左辺を計算し、式12を用いて変形すると、

$$\begin{aligned} & D(P||Q) + D(Q||R) - D(P||R) \\ &= \psi(\boldsymbol{\theta}^P) + \psi^\eta(\boldsymbol{\eta}^Q) - \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^Q + \psi(\boldsymbol{\theta}^Q) + \psi^\eta(\boldsymbol{\eta}^R) - \boldsymbol{\theta}^Q \cdot \boldsymbol{\eta}^R - \psi(\boldsymbol{\theta}^P) - \\ & \quad \psi^\eta(\boldsymbol{\eta}^R) + \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^R \\ &= \psi^\eta(\boldsymbol{\eta}^Q) - \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^Q + \psi(\boldsymbol{\theta}^Q) - \boldsymbol{\theta}^Q \cdot \boldsymbol{\eta}^R + \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^R \end{aligned}$$

ここで、Legendre変換の定義より $\psi(\boldsymbol{\eta}^Q) = \boldsymbol{\eta}^Q \cdot \boldsymbol{\theta}^Q - \psi(\boldsymbol{\theta}^Q)$ を代入し、

$$\begin{aligned}
& \psi^\eta(\boldsymbol{\eta}^Q) - \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^Q + \psi(\boldsymbol{\theta}^Q) - \boldsymbol{\theta}^Q \cdot \boldsymbol{\eta}^R + \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^R \\
&= (\boldsymbol{\eta}^Q \cdot \boldsymbol{\theta}^Q - \psi(\boldsymbol{\theta}^Q)) - \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^Q + \psi(\boldsymbol{\theta}^Q) - \boldsymbol{\theta}^Q \cdot \boldsymbol{\eta}^R + \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^R \\
&= \boldsymbol{\eta}^Q \cdot \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^Q - \boldsymbol{\theta}^Q \cdot \boldsymbol{\eta}^R + \boldsymbol{\theta}^P \cdot \boldsymbol{\eta}^R \\
&= (\boldsymbol{\theta}^Q - \boldsymbol{\theta}^P) \cdot (\boldsymbol{\eta}^Q - \boldsymbol{\eta}^R)
\end{aligned}$$

P と Q を結ぶ θ 座標上での測地線と、Q と R を結ぶ η 座標上での測地線（双対測地線）が直交するとき、この内積は 0 になるため、式 13 が成り立つ。

同様に、P と Q が η 座標上の測地線（双対測地線）、Q と R が測地線で結ばれており、それらが直交している時、以下が成り立つ。

$$D^\eta(P||R) = D^\eta(P||Q) + D^\eta(Q||R) \quad (14)$$

証明は、式 13 と同様に右辺-左辺を計算すると、

$$\begin{aligned}
& D^\eta(P||R) - D^\eta(P||Q) + D^\eta(Q||R) \\
&= D(R||P) - D(Q||P) + D(R||Q) \\
&= (\boldsymbol{\eta}^Q - \boldsymbol{\eta}^P) \cdot (\boldsymbol{\theta}^Q - \boldsymbol{\theta}^R)
\end{aligned}$$

P と Q を結ぶ η 座標上での測地線（双対測地線）と、Q と R を結ぶ θ 座標上での測地線が直交するとき、この内積は 0 になるため、式 14 が成り立つ。拡張ピタゴラスの定理が成り立つ 2 つのパターンを図 2 に示した。

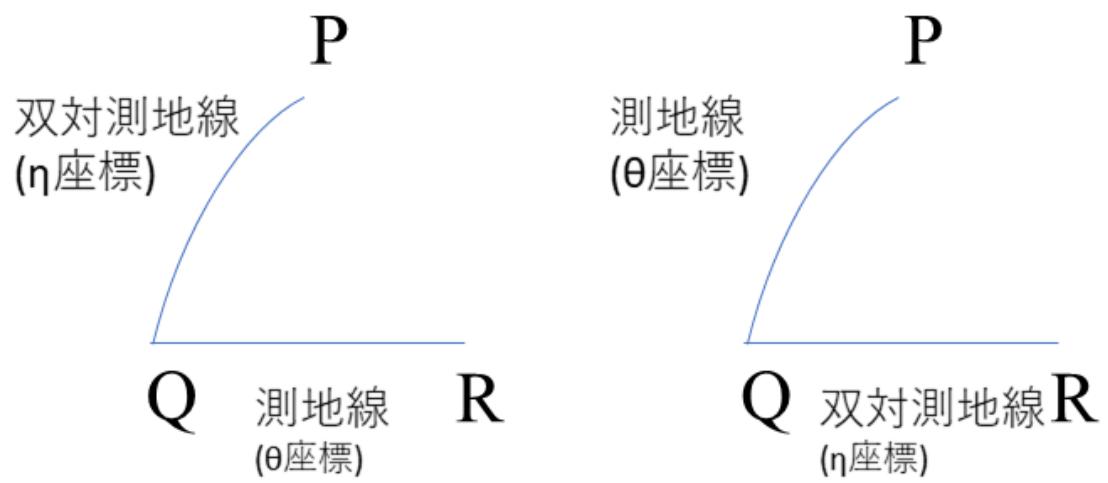


図 2 拡張ピタゴラスの定理が適用される 2 つのパターン

1.3.3 拡張射影定理

双対平坦座標系における拡張射影定理は、ユーグリッド空間における点と曲面 S の最短距離の求め方を拡張したものである。ユーグリッド空間において、点 P と曲面 S の最短距離は、点 P から曲面 S に直交する直線をひいた交点として求めることができる。直線と曲面が直交するというのは、交点における S の接平面のすべてとこの直線が直交するという意味である。この考え方は、双対平坦座標系においても成り立つ。ある点 P から、曲面 S への θ 座標でのダイバージェンスを最も小さくする S 上の点 R は、 P からの θ 座標上の測地線が曲面 S と直交するような点である。 S が平坦な部分空間である場合、これは一意に定まる。このことは、ユーグリッド空間においては、点と平面の間の距離を最短にする射影が、点から平面に垂線を下した時の交点として一意に定まることと同様である。拡張射影定理のイメージ図を図 3 に示した。

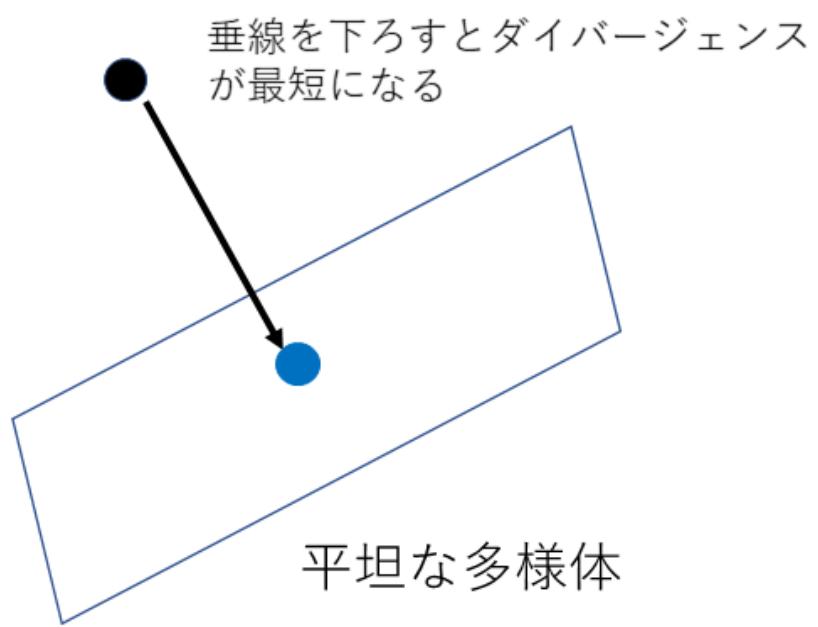


図 3 拡張射影定理のイメージ図

2 指数型分布族と統計多様体

このセクションでは、指数型分布族について上記の理論を適用することで、双対平坦座標系上の点として分布が表現できることを示す。指数型分布族は、自然パラメータ θ 、ポテンシャル関数と呼ばれる凸関数 $\psi(\theta)$ を用いて以下の式で表すことのできる確率分布のクラスである。

$$p(x) = \exp\left(\sum_i \theta_i F_i(x) + C(x) - \psi(\theta)\right) \quad (15)$$

正規分布、二項分布などの多くの確率分布が、この形式で表現することができ、指数型分布族として表現することでこれらの分布を統一的に扱うことができる。ここで、 θ 座標は分布に付与される座標であり、確率分布がつくる多様体を統計多様体と呼ぶ。統計多様体上の点は、それぞれのパラメータを持つ確率分布を示す。

(例) 正規分布 $P(x|m, s) = \frac{1}{\sqrt{2\pi}s^2} e^{-\frac{(x-m)^2}{2s^2}}$ 確率密度関数の対数をとって変形すると、

$$\begin{aligned} \log P(x|m, s) &= -\frac{(x-m)^2}{2s^2} - \frac{1}{2} \log 2\pi s^2 \\ &= \frac{m}{s^2}x - \frac{1}{2s^2}x^2 - \left(\frac{m^2}{2s^2} + \frac{1}{2} \log 2\pi s^2\right) \\ &= \left(\frac{m}{s^2}, -\frac{1}{2s^2}\right) \cdot \begin{pmatrix} x \\ x^2 \end{pmatrix} - \left(\frac{m^2}{2s^2} + \frac{1}{2} \log 2\pi s^2\right) \end{aligned}$$

これを指数型分布族の式 15 と比較すると、正規分布の指数型表現が得られる。

$$\log P(x|\theta_1, \theta_2) = (\theta_1, \theta_2) \cdot \begin{pmatrix} x \\ x^2 \end{pmatrix} - \psi(\theta_1, \theta_2) \quad (16)$$

$$= (\theta_1, \theta_2) \cdot \begin{pmatrix} F_1(x) \\ F_2(x) \end{pmatrix} - \psi(\theta_1, \theta_2) \quad (17)$$

$$\theta_1 = \frac{m}{s^2} \quad (18)$$

$$\theta_2 = -\frac{1}{2s^2} \quad (19)$$

$$\psi(\theta_1, \theta_2) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log \frac{\pi}{\theta_2} \quad (20)$$

$$F_1(x) = x \quad (21)$$

$$F_2(x) = x^2 \quad (22)$$

$$C(x) = 0 \quad (23)$$

$$(24)$$

上記が、正規分布の指数型表現である。次節以降では、指数型分布族がつくる統計多様体上のポテンシャル関数 $\psi(\boldsymbol{\theta})$ に対して、前節の議論を当てはめていく。

2.1 指数型分布族における双対平坦座標系

まず指数型分布族における θ の双対座標である η 座標を計算する。

指数型分布族におけるポテンシャル関数 $\psi(\boldsymbol{\theta})$ は、以下のように導ける。

$$\psi(\boldsymbol{\theta}) = \log \int \exp(\sum \theta_i F_i + C) \quad (25)$$

η 座標系の定義である

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}) \quad (26)$$

を計算すると、 η_k は、以下のように $F_k(x)$ の期待値になることがわかる。

$$\begin{aligned}
\eta_k &= \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\frac{\partial}{\partial \theta_k} \int \exp(\sum \theta_i F_i + C) dx}{\int \exp(\sum \theta_i F_i + C) dx} \\
&= \frac{\int F_k \exp(\sum \theta_i F_i + C) dx}{\int \exp(\sum \theta_i F_i + C) dx} \\
&= \frac{\int \exp(\psi) F_k \exp(\sum \theta_i F_i + C - \psi) dx}{\int \exp(\psi) \exp(\sum \theta_i F_i + C - \psi) dx} \\
&= \int F_k p(x) dx \\
&= E(F_k)
\end{aligned}$$

よって、

$$\eta_k = E(F_k) \quad (27)$$

のことから、指指数型分布族における η 座標は期待値パラメータと呼ばれる。指指数型分布族は、 θ 座標、 η 座標の双対平坦座標系に埋め込まれる。 θ 座標系、 η 座標系での平坦性をそれぞれ e 平坦、m 平坦と呼ぶ。平坦な座標系での測地線は、それぞれの座標上では以下のような直線として表される。

$$\theta(t) = at + b \quad (28)$$

$$\eta(t) = at + b \quad (29)$$

したがって、2 つの指指数型分布族（例えば正規分布 P と正規分布 Q）の間には、 θ 座標系と η 座標系それぞれで測地線がかかる。 θ 座標系での測地線 $t\theta^P + (1-t)\theta^Q$ を e 測地線、 η 座標系での測地線 $t\eta^P + (1-t)\eta^Q$ を m 測地線と呼ぶ。

例えば、正規分布は平均、標準偏差の 2 パラメータによって同定される指指数型分布族であり、全ての正規分布は θ_1, θ_2 によって一意に特定され、さらに対応する η_1, η_2 によっても一意に特定される。正規分布の η 座標は、 $\eta_i = \frac{\partial \psi}{\partial \theta_i}$ を計算すると、

$$\eta_1 = \frac{-\theta_1}{2\theta_2} \quad (30)$$

$$\eta_2 = \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} \quad (31)$$

双対平坦座標系では双直交性が成り立つため、 θ_1 と η_2 , θ_2 と η_1 によっても一意に特定され、これらは直交座標系となる。

2.2 指数型分布族の十分統計量

十分統計量とは、母数パラメータの統計的推定を行うために十分な情報を持つ統計量である。

指数型分布族の母数パラメータの最尤推定の手順を以下に示す。観測されたデータ点 x_1, x_2, \dots, x_n から、母分布となる指数型分布族 $\log P(x) = C(x) + F_1(x)\theta_1 + F_2(x)\theta_2 - \psi(\boldsymbol{\theta})$ の未知パラメータ θ_1, θ_2 を最尤推定することを考える。簡単のため、ここでは 2 パラメータの指数型分布族を例とする。対数尤度関数 $L(\theta_1, \theta_2)$ は、

$$L(\theta_1, \theta_2) = \sum_i^n (C(x_i) + F_1(x_i)\theta_1 + F_2(x_i)\theta_2 - \psi(\boldsymbol{\theta})) \quad (32)$$

$$= \theta_1 \sum_i^n F_1(x_i) + \theta_2 \sum_i^n F_2(x_i) - n\psi(\boldsymbol{\theta}) \quad (33)$$

したがって、最尤推定値は、 L の θ_1, θ_2 での偏微分がそれぞれ 0 になることから、

$$\hat{\eta}_1 = \frac{\sum_i^n F_1(x_i)}{n} \quad (34)$$

$$\hat{\eta}_2 = \frac{\sum_i^n F_2(x_i)}{n} \quad (35)$$

$$(36)$$

となり、母分布を一意に定めるパラメータが推定できる。このように、推定値 $\hat{\eta}_1$ を得るために、 $F_1(x_i)$ の値が得られれば十分である。双対座標系の性質より θ_k と η_k は相互に変換可能なので、 $F_k(x)$ は指数型分布族における十分統計量として扱うことができる。つまり、観測データから母分布の指数型分布族の母数パラメーターの最尤推定を行う際、それぞれの θ_k に対応するような $F_k(x)$ の値が得られれば十分である。

2.3 内積行列としての Fisher 情報行列

次に、指数型分布族のなす統計多様体におけるリーマン計量 $g_{ij} = \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$ を計算する。

$$\begin{aligned}
g_{ij} &= \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_j} \int F_i p(x) dx \\
&= \int F_i (F_j - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_j}) p(x) dx \\
&= \int F_i F_j p dx - \int F_i E[F_j] p dx \\
&= \int F_i F_j p dx - \int F_i E[F_j] p dx - \int E[F_i] F_j p dx + \int E[F_i] E[F_j] p dx \\
&= E[(F_i - E[F_i])(F_j - E[F_j])] \\
&= E[(F_i - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_i})(F_j - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_j})] \\
&= E[\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j}]
\end{aligned}$$

よって、

$$g_{ij} = E[\frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j}] \quad (37)$$

この行列は、Fisher 情報行列と呼ばれる統計学でよく知られた行列であり、データから分布の母数パラメーターを推定する際の難易度に関係する。

サンプルセットが観測されている時、その母分布のパラメータを推定する場合を考える。その際、最尤推定は以下のような好ましい性質をもつため、最尤推定はよく用いられる手法である。

まず、パラメータ推定値の分散がクラーメル・ラオの不等式の下限として与えられ、これはサンプル数とフィッシャー情報行列 \mathbf{I} の逆行列に依存することが知られている。ただし、一様分布の母分布推定のようなパラメータが積分範囲を指定するケースでは、クラーメル・ラオの不等式は必ずしも成り立つわけではない。

$$Var(\hat{\boldsymbol{\theta}}) \geq \frac{\mathbf{I}^{-1}}{n} \quad (38)$$

また、一般に好ましいとされる推定量として、サンプル数を増やしていくと、パラメータの推定値 $\hat{\theta}$ は真値 θ に近づいていくような推定量（一致推定量）や、パラメータ推定値の分布の期待値がパラメーターの真値となる推定量（不偏推定量）がある。最尤推定量は、サンプル数が十分に多い時には、近似的に以下の式が成り立ち、上記の一致性、不偏

性を近似的に満たす。

$$\hat{\boldsymbol{\theta}} \sim Normal(\boldsymbol{\theta}, \frac{\mathbf{I}^{-1}}{n}) \quad (39)$$

その中で、フィシャー情報行列の逆行列は、推定されるパラメータ間の依存関係を表している。指指数型分布族において各分布が埋め込まれる θ 座標系におけるリーマン計量は Fisher 情報行列で表現される。

指指数型分布族の場合、 θ 座標系や η 座標系のかわりに、それらを混ぜ合わせた混合座標系を用いることで、双直交性により、Fisher 情報行列が、対角行列やブロック対角行列になる。これはパラメーター推定を行う際に有用である場合がある。図 4 は $\mu = 3, \sigma = 2$ の正規分布を例にこのことをシミュレーションで示した結果である。この正規分布の θ 座標は $\theta_1 = 3, \theta_2 = -0.5$ となり、 η 座標は $\eta_1 = 3, \eta_2 = 10$ となる。この正規分布から乱数を 1000 個発生させ、データ点から式 27 を用いて η 座標を推定し、変換式を用いて θ 座標の推定値を得る作業を 1000 回繰り返し、推定値をプロットした。 θ 座標同士、 η 座標同士の推定値の間には相関関係がみられる一方で、 $[\theta_1, \eta_2], [\theta_2, \eta_1]$ のプロットでは推定値同士の相関がみられないことから、混合座標系の双直交性を確認できる。

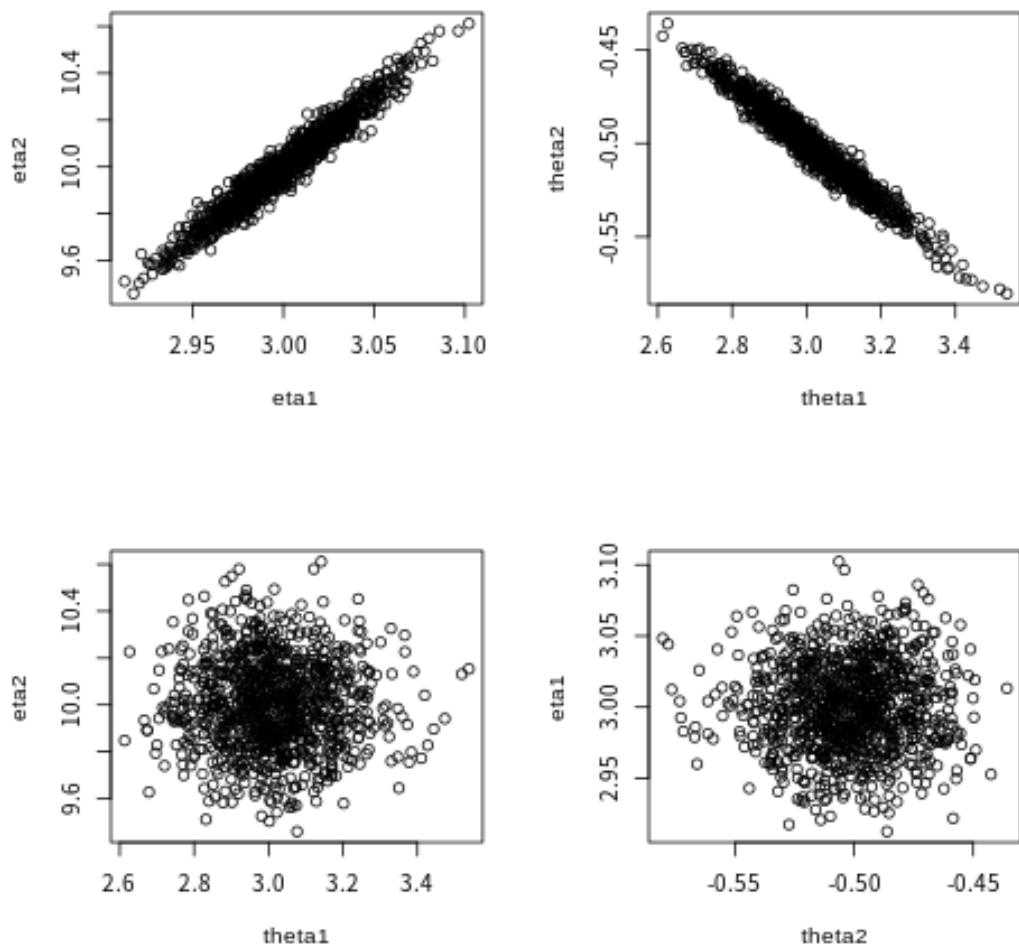


図 4 双直交性のシミュレーション。 θ 座標同士、 η 座標同士の推定値は相互に強く相関しているが、混合座標系では推定値に相関がみられない。

2.4 Kullback–Leibler ダイバージェンス

確率分布同士の似通いを測るダイバージェンスとして、以下の式で定義される f ダイバージェンスと呼ばれるクラスに属するものが一般によくつかわれる。

$$D_f[P||Q] = \sum p_i f\left(\frac{q_i}{p_i}\right) \quad (40)$$

f ダイバージェンスは f の選び方によっていろいろなものが考えられ、Hellinger 距離、Kullback–Leibler(KL) ダイバージェンス、 α ダイバージェンスなどがある。そのような f ダイバージェンスの中でも KL ダイバージェンスは最もよく用いられる確率分布間のダイバージェンス指標の 1 つであり、多くのデータ解析手法で用いられている。

指数型分布族がなす統計多様体上で、 P から Q への KL ダイバージェンスと Q から P への Bregman ダイバージェンスと一致する。まず、KL ダイバージェンスの定義より

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (41)$$

$$= \int p(x)(\log p(x) - \log q(x))dx \quad (42)$$

$$= \int p(x)(C + \sum F_i \theta_i^P - \psi(\boldsymbol{\theta}^P) - C - \sum F_i \theta_i^Q + \psi(\boldsymbol{\theta}^Q)) \quad (43)$$

$$= \int p(x)(\sum F_i \theta_i^P - \psi(\boldsymbol{\theta}^P) - \sum F_i \theta_i^Q + \psi(\boldsymbol{\theta}^Q)) \quad (44)$$

$$(45)$$

ここで、Bregman ダイバージェンス $D_B(q||p)$ は定義より

$$D_B(q||p) = \psi(\boldsymbol{\theta}^Q) - \psi(\boldsymbol{\theta}^P) - \langle \nabla \psi(\boldsymbol{\theta}^P), \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \rangle \quad (46)$$

これを代入すると

$$KL(p||q) = \int p(x)(D_B(q||p) + \langle \nabla \psi(\boldsymbol{\theta}^P), \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \rangle - \langle \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P, F(x) \rangle) \quad (47)$$

$$= D_B(q||p) + \langle \nabla \psi(\boldsymbol{\theta}^P), \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \rangle - \int p(x) \langle \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P, F(x) \rangle dx \quad (48)$$

この式の第 2 項は、 $\langle \nabla \psi(\boldsymbol{\theta}^P), \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \rangle = \langle \int p(x)F(x)dx, \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P \rangle$ なので、これは $\sum_i (\theta_i^Q - \theta_i^P) \int p(x)F_i(x)dx$ となる。また、第 3 項は $\int p(x) \langle \boldsymbol{\theta}^Q - \boldsymbol{\theta}^P, F(x) \rangle$

) $dx = \int p(x) \sum_i (\theta_i^Q - \theta_i^P) F_i(x) dx = \sum_i (\theta_i^Q - \theta_i^P) \int p(x) F_i(x) dx$ であり、第 2 項と等しくなる。従って、

$$KL(p||q) = D_B(q||p) \quad (49)$$

同様に、

$$KL(q||p) = D_B(p||q) \quad (50)$$

さらに、 θ 座標上での Bregman ダイバージェンスである $D_B(p||q)$ と、 η 座標上での Bregman ダイバージェンス $D_B^\eta(p||q) = \psi^\eta(\boldsymbol{\eta}^Q) - \psi^\eta(\boldsymbol{\eta}^P) - \langle \nabla \psi^\eta(\boldsymbol{\eta}^P), \boldsymbol{\eta}^Q - \boldsymbol{\eta}^P \rangle$ の間には、 $D_B(p||q) = D_B^\eta(q||p)$ の関係がある。これらをまとめると、指數型分布族による統計多様体においては、以下の関係式が成り立つ。

$$D_B(p||q) = D_B^\eta(q||p) = KL(q||p) \quad (51)$$

$$D_B(q||p) = D_B^\eta(p||q) = KL(p||q) \quad (52)$$

以上のことから、点 P から平面 S への e 射影した点を Q とすると、拡張射影定理より θ 座標上での Bregman ダイバージェンス、すなわち $KL(Q||P)$ を最小化するような射影になっている。点 P から平面 S への m 射影した点を Q' とすると、 η 座標上での Bregman ダイバージェンス、すなわち $KL(P||Q')$ を最小化するような射影になっている。

指數型分布族による統計多様体は双対平坦性を持つため、前節で議論した拡張ピタゴラスの定理が成り立つ。式 51 より、 η 座標上での Bregman ダイバージェンスが KL ダイバージェンスと等しくなるので、式 53 より、P と Q が η 座標上の測地線、Q と R が η 座標上の測地線で結ばれており、それらが直交している時以下が成り立つ。

$$KL(P||R) = KL(P||Q) + KL(Q||R) \quad (53)$$

ここでは、正規分布を例にそれらを示す。まず、平均 3, 標準偏差 2 の正規分布を考え、これを正規分布 P とおく。 $[\mu, sd]$ 座標、 $[\theta_1, \theta_2]$ 座標、 $[\eta_1, \eta_2]$ 座標はそれぞれ変換可能であるため、正規分布 P の θ 座標、 η 座標を計算すると、 $[\theta_1^P, \theta_2^P] = [-0.75, -0.125]$ 、 $[\eta_1^P, \eta_2^P] = [3, 13]$ である。

次に、正規分布 P から η 座標上で $[-1, 3]$ 進んだ点を正規分布 Q とする。正規分布 Q の η 座標は $[\eta_1^Q, \eta_2^Q] = [3 - 1, 13 + 3] = [2, 16]$ である。残り 2 つの座標系での位置を計算すると、 $[\theta_1^Q, \theta_2^Q] = [0.167, -0.0417]$ 、 $[\mu^Q, sd^Q] = [2, 3.46]$ になる。

次に、正規分布 Q から θ 座標上で $[-6, -2]$ 進んだ点を正規分布 R とする。ここで、 $[-1, 3]$ と $[-6, -2]$ は内積が 0 であり直交している。したがって、正規分布 R の θ 座標は

$[\theta_1^R, \theta_2^R] = [-5.83, -2.04]$ となる。同様に、残り 2 つの座標系での位置を計算すると、 $[\eta_1^R, \eta_2^R] = [-1.43, 2.29]$ 、 $[\mu^R, sd^R] = [-1.43, 0.495]$ と計算できる。これらの 3 つの正規分布 P,Q,R 間の KL ダーバージェンスを計算すると、 $KL(P||R) \approx KL(P||Q) + KL(Q||R)$ となっていることが数値的に確認できる。

3 データ解析手法の情報幾何的解釈

多くのデータ解析手法は、観測データに対して何らかの統計モデルを設定し、その統計モデルにおけるパラメータを推定する手続として考えることができる。情報幾何を用いることで、これを統一的な枠組みで幾何学的に説明できることが知られている。このセクションでは、まず離散確率分布が指数型分布族であることから確率分布全体の集合で双対平坦座標系の性質が成り立つことを示し、データ解析手法の情報幾何的解釈の枠組みについて述べる。次に、生命科学分野でもよく用いられるデータ解析の例として、指数型分布族の母数パラメータ推定と EM アルゴリズムについて、その情報幾何的解釈について述べる。

3.1 离散確率分布の指数型表現

離散確率分布は、指数型分布族であることから、離散確率分布の集合は双対平坦座標系に埋め込むことができる。 x が $0, 1, 2, \dots, n$ の $n+1$ カテゴリのどれかをとる離散確率分布の確率質量関数は以下のように表せる。

$$\mathbf{p} = [p_0, p_1, \dots, p_n] \quad (54)$$

$$\sum_{i=0}^n p_i = 1 \quad (55)$$

ここで、 δ_i を以下のように定義する。

$$\delta_i(x) = \begin{cases} 0, & x \neq i \\ 1, & x = i \end{cases} \quad (56)$$

指数型表現は以下のように表すことができる。

$$p(x) = \sum_{i=0}^n \delta_i(x) p_i \quad (57)$$

$$= \exp\left(\sum_{i=0}^n \delta_i(x) \log \frac{p_i}{p_0} + \log p_0\right) \quad (58)$$

この分布は以下のような指数型分布族の形になっている。

$$p(x) = \exp(C(x) + \sum_{i=1}^n F_i(x) \theta_i - \psi(\boldsymbol{\theta})) \quad (59)$$

$$\theta_i = \log \frac{p_i}{p_0} \quad (60)$$

$$C(x) = 0 \quad (61)$$

$$F_i(x) = \delta_i(x) \quad (62)$$

$$\psi(\boldsymbol{\theta}) = -\log p_0 \quad (63)$$

$$p_0 = \frac{1}{\sum_{i=1}^n \exp(\theta_i) + 1} \quad (64)$$

ここで、 p_0 を θ 座標を用いて表すと、

$$\begin{aligned} \theta_i &= \log \frac{p_i}{p_0} \\ \exp(\theta_i) &= \frac{p_i}{p_0} \\ \sum_{i=1}^n \exp(\theta_i) &= \sum_{i=1}^n \frac{p_i}{p_0} = \frac{1-p_0}{p_0} \\ p_0 &= \frac{1}{\sum_{i=1}^n \exp(\theta_i) + 1} \end{aligned}$$

よって、

$$\psi(\boldsymbol{\theta}) = -\log p_0 = \log\left(\sum_{i=1}^n \exp(\theta_i) + 1\right) \quad (65)$$

ポテンシャル関数 $\psi(\boldsymbol{\theta})$ から η 座標を計算すると、

$$\begin{aligned}
\eta_i &= \frac{\partial \psi(\theta)}{\partial \theta_i} \\
&= \frac{\exp(\theta_i)}{\sum_{i=1}^n \exp(\theta_i) + 1} \\
&= \frac{p_i}{p_0} p_0 \\
&= p_i
\end{aligned}$$

となり、各カテゴリの実現確率 p_1, \dots, p_n がそのまま η_1, \dots, η_n 座標の値になる。したがって、 $n+1$ カテゴリの離散確率分布全体はこの双対平坦空間の点に対応する。離散確率分布だけでなく、任意の連続確率分布も無限次元の離散確率分布と捉えることができるため、任意の確率分布は双対平坦空間の点として考えられる。そして、パラメトリックな統計モデルは、この空間の部分空間として定義できる。観測データをそのまま解釈して得られる確率分布を「データが定める分布」とし、この空間上の 1 点であらわされる。このような分布が複数ある場合は、これらは多様体をなし、データ多様体と呼ばれる。一方で、統計モデルが定める分布の集合をモデル多様体と呼ぶ。一般に、データが定める分布がモデル多様体上の点であるとは限らないため、多くのデータ解析手法は「データが定める分布からモデル多様体への射影」として説明することができる。また、正規分布などの多くのパラメトリックモデルも指數型分布族であるため、その双対平坦座標系上でこれを考えることもできる。

射影を行う際に、部分多様体の平坦性についての情報は重要である。なぜなら、拡張射影定理より、 e 平坦な空間への m 射影、 m 平坦な空間への e 射影は一意に定まるからである。そうでない場合は、ダイバージェンスを最小化するような射影が複数存在する可能性があり推定が難しくなる。

ある部分多様体 D 上の点と点の間に θ 座標上での測地線を引いた時、その測地線上の点も必ず D に含まれるのであれば、その多様体は e 平坦であると判断できる。同様に、ある部分多様体 D 上の点と点の間に η 座標上での測地線を引いた時、その測地線上の点も必ず D に含まれるのであれば、その多様体は m 平坦になる。例えば、正規分布は双対平坦空間にマッピングすることができ、指數型表現した際の θ 座標、 η 座標はともに平坦な座標系であるが、正規分布 P と正規分布 Q の θ 座標を結んだ測地線上の全ての点に対して対応する異なる正規分布が存在する。同様のことが η 座標でも成り立つ。

また、セクション 1 で述べたように、 θ 座標系が平坦な座標系であるとき、部分多様体 D が、 $A\theta + b = 0$ の線形束縛で定義されるなら D も e 平坦である。同様に、 η 座標系が

平坦な座標系であり D が、 $A\eta + b = 0$ の線形束縛で定義されるなら D も m 平坦である。

3.2 指数型分布族の母数パラメータの最尤推定

ここでは、生命科学分野でもよく用いられる手法として指数型分布族の母数パラメータの最尤推定について、その情報幾何的解釈を示す。

データ点 x_1, x_2, \dots, x_n が観測されたとき、母分布となる指数型分布族のパラメータの推定値は、十分統計量 $F(x)$ を用いて以下のように書きあらわすことができた。

$$\hat{\eta}_k = \frac{\sum_i^n F_k(x_i)}{n} \quad (66)$$

これは、情報幾何的にはデータが定める分布 P の座標から、モデル多様体への m 射影をした際の交点の座標 Q として求めることができるなどを以下に示す。ここでデータが定める分布は、以下の確率密度関数となる。

$$P(x) = \frac{1}{n} \sum_{i=1}^n \delta_i(x - x_i) \quad (67)$$

これは、ノンパラメトリックにデータをそのまま解釈した分布であり、観測点に $\frac{1}{n}$ が立っているような分布である。拡張射影定理より、この射影は $KL(P||Q)$ を最小化するので、

$$L(P||Q) = KL(P||Q) = \int P \log \frac{P}{Q} dx = \int (P \log P - P \log Q) dx \quad (68)$$

このうち、 $Q(x)$ に関する部分だけを取り出すと

$$\begin{aligned} L(\theta_1, \theta_2) &= \int (P \log Q) dx = \int P(x)(C(x) + F_1(x)\theta_1 + F_2(x)\theta_2 + \psi(\boldsymbol{\theta})) dx \\ &= \frac{1}{n} \sum_i^n (C(x_i) + F_1(x_i)\theta_1 + F_2(x_i)\theta_2) - \psi(\boldsymbol{\theta}) \end{aligned} \quad (69)$$

$$= \frac{1}{n} \sum_i^n (C(x_i) + F_1(x_i)\theta_1 + F_2(x_i)\theta_2) - \psi(\boldsymbol{\theta}) \quad (70)$$

したがって、これを最小にするような分布では、 L の θ_1, θ_2 での偏微分がそれぞれ 0 に

なることから

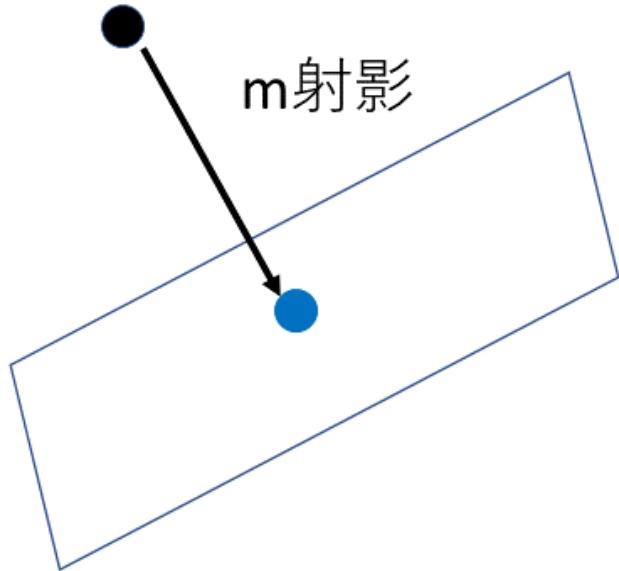
$$\hat{\eta}_1 = \frac{\sum_i^n F_1(x_i)}{n} \quad (71)$$

$$\hat{\eta}_2 = \frac{\sum_i^n F_2(x_i)}{n} \quad (72)$$

$$(73)$$

となり、最尤推定と同様の結果が導ける。

データが定める分布



モデル多様体

図 5 母数パラメータの最尤推定の情報幾何のイメージ図

3.3 EM アリゴリズムの情報幾何

ここでは EM アルゴリズムについて、その情報幾何的解釈を示す。

EM アルゴリズムは、隠れ変数のある統計モデルにおけるパラメーターの最尤推定において汎用的に使われる手法である。データを \mathbf{x} , 隠れ変数を \mathbf{s} , パラメータを $\boldsymbol{\alpha}$) とおくと、これは以下の対数尤度 $J(\boldsymbol{\alpha}) = \log P(\mathbf{x}; \boldsymbol{\alpha}) = \log \sum_{\mathbf{s}} P(\mathbf{x}, \mathbf{s}; \boldsymbol{\alpha})$ を最大化するような $\boldsymbol{\alpha}$ を計算する。これを情報幾何的に考える。

まず、データ \mathbf{x} と隠れ変数 \mathbf{s} の同時確率分布 $P(\mathbf{x}, \mathbf{s})$ の空間を考える。 \mathbf{x} を n グリッドの離散確率分布（連続の場合 n は無限大）, \mathbf{s} を m グリッドの離散確率分布と考えると、あらゆる $P(\mathbf{x}, \mathbf{s})$ からなる空間は $n \times m$ 次元の離散確率分布となり、これは双対平坦空間 S とする。

データが定める分布は $P(x, s) = P(x)P(s|x)$ より、 S の部分多様体 D として以下のような式で表すことができる。

$$P^D(x, s) = \frac{1}{n} \sum_i^n \delta(x - x_i)P(s|x_i) \quad (74)$$

隠れ変数 s は観測できないことから $P(s|x_i)$ が様々な値を取りうるため、データが定める分布は 1 点ではなく広がりを持ったデータ多様体で与えられる。

ここで、簡単のため $n=2, m=2$ の場合を考える。双対平坦空間 S は 4 グリッドの離散確率分布の集合で構成される。各グリッドは、 $[x, s]=[1,1], [1,2], [2,1], [2,2]$ の 4 パターンに対応しているとし、それぞれの確率を p_0, p_1, p_2, p_3 とおく。離散分布の指數型表現の式より、 S の η 座標系は $\eta_1 = p_1, \eta_2 = p_2, \eta_3 = p_3$ の 3 つである。ここで、 $x_i = 1$ の時、 $\delta(x - x_i)P(s|x_i)$ は $\eta_2 = \eta_3 = 0$ の直線に対応する。 $x_i = 2$ の時、 $\delta(x - x_i)P(s|x_i)$ は $\eta_1 = 0$ の平面に対応する。 $P^D(x, s)$ はこの平面と直線の重み付き平均なので、 S における η 座標系の線形束縛で表すことができる。したがって、データ多様体 D は m 平坦である。一方で、モデル多様体 M は一般に平坦であるとは限らない。

指數型分布族の統計多様体において拡張射影定理が成り立つ。モデル多様体 M 上の点 Q^0 からデータ多様体 D に対しての e 射影した点を P^1 とすると、この射影は $D_B(Q||P) = KL(P||Q)$ を最小化するような P である。また、データ多様体 D 上の P からモデル多様体 M に M 射影した点を Q' とすると、この射影は $D_B^\eta(P||Q') = KL(Q'||P)$ を最小化している。ここで、 Q_0 の初期値からスタートし、 e 射影と m 斜影を順番に繰り返すと、 $KL(P_t||Q_t)$ は単調減少していくことが知られている。これを、 e 射影と m 斜

影を繰り返すため em アルゴリズムと呼ばれる。EM アルゴリズムの E ステップと M ステップを反復されることによって、隠れ変数がある場合でのパラメータの最尤推定を行う手順は、em アルゴリズムによってデータ多様体とモデル多様体の間の KL ダーバージェンスを最小化するような点を探索することとほぼ同等な処理になる。

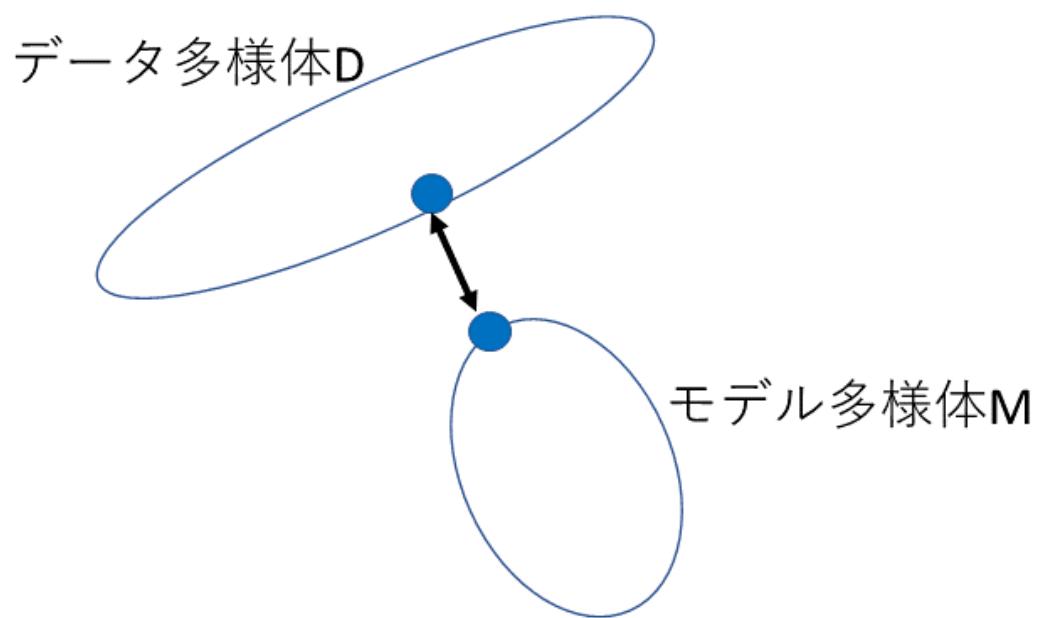


図 6 EM アルゴリズムの情報幾何のイメージ図

参考文献

- [1] 甘利 俊一. ”新版 情報幾何学の新展開” サイエンス社, 2019.
- [2] 甘利俊一. ”講座 情報幾何とその応用-I: 情報幾何とは何か-入門編.” システム/制御/情報 48.6 (2004): 227-235.
- [3] 赤穂昭太郎. ”情報幾何と機械学習.” 計測と制御 44.5 (2005): 299-306.
- [4] 赤穂昭太郎, 渡辺一帆, and 岡田真人. ”指數型分布族の空間におけるデータ解析法について.” (2010).
- [5] Miura, Keiji. ”An introduction to maximum likelihood estimation and information geometry.” Interdisciplinary Information Sciences 17.3 (2011): 155-174.
- [6] Nielsen, Frank, and Richard Nock. ”Entropies and cross-entropies of exponential families.” 2010 IEEE International Conference on Image Processing. IEEE, 2010.

第 4 章：拡大指数型分布族分解法

2021 年 3 月 23 日

概要

本章では、学位論文の主論文である文献 [1] の内容についてまとめる。本研究では、新規のノンパラメトリック統計解析手法である拡大指数型分布族分解法 (Decomposition into Extended Exponential Family: DEEF) を提案する。この手法は、数マーカーを測定したシングルセル発現データセットの各サンプルを低次元空間に埋め込み、サンプルセットの heterogeneity を説明するような構成分布を同定する。本章ではまず、理論的な確率分布セットを用いて DEEF が適切な分布セットの分解を行うことを示す。次に乳腺上皮細胞の培養細胞に対する EGF 刺激の影響を調べたサイトメトリーデータセットに DEEF を適用する。DEEF は 2 つのパラメータで細胞集団プロファイルの複雑なダイナミクスを捉え、それを低次元座標上の軌跡として描画した。この 2 つのパラメータによって事前の仮定なしに細胞集団プロファイルの主要なパターンが同定された。次に、DEEF のさらなる応用として、サイトメトリーデータセットの次元削減およびタイムコース補完を行った。DEEF を用いることで、上位の座標軸のみを用いて分布を再構成することができた。この性質を使うことで、実データに基づいたサイトメトリー人工データを作成することができる。DEEF によって得られた座標軸の値を用いることで、サンプルの分布の特徴とサンプルの形質情報との関連を、既存のデータ解析手法を用いて調べることが可能になる。

1 拡大指数型分布族分解法の提案

1.1 動機

本論文の第 2 章では、細胞集団プロファイルからの特微量抽出に関するこれまでのアプローチについて紹介し、その利点と欠点について述べ、情報ロスがなくかつ解釈性も高い手法が求められることを述べた。そこで、第 3 章で紹介した情報幾何の知見を取り入れることで、サイトメトリーデータからの新たなノンパラメトリック特微量抽出手法が

開発できると考え、新規の解析手法である拡大指数型分布族分解法 (Decomposition into Extended Exponential Family: DEEF) を提案する。まず、情報幾何で主に研究対象とされている指数型分布族の定義を拡張した新規の確率分布族である拡大指数型分布族を定義し、任意の確率分布を拡大指数型分布族として表現する新規のアルゴリズムを提案する。

1.2 理論とアルゴリズム

情報幾何学で主に研究対象とされている指数型分布族は、その確率密度関数・確率質量関数が以下で表現される確率分布のクラスである。

$$\log P(x, \boldsymbol{\theta}) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi(\boldsymbol{\theta}) \quad (1)$$

ここで、 $C(x)$ 、 $F_k(x)$ 、 $\psi(\boldsymbol{\theta})$ は固定された関数であり ($\psi(\boldsymbol{\theta})$ は凸関数)、 $\boldsymbol{\theta}$ は分布を特定するパラメータである。正規分布や二項分布など、多くのパラメトリック確率分布が指数型分布族に含まれている。一方で、混合正規分布など、生命科学でよく用いられる確率分布の中には指数型分布族に含まれていないものもある。そこで、本研究では指数型分布族の定義を拡張した拡大指数型分布族 (Extended Exponential Family: EEF) を次のように定義した。

$$\log P(x, \boldsymbol{\theta}) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi'(\boldsymbol{\theta}) \quad (2)$$

$$\psi'(\boldsymbol{\theta}) = \sum_{k=1} h_k \theta_k^2 \quad where \quad h_k = -1 \quad or \quad 1 \quad (3)$$

ここで、EEF は式 1 とほぼ同じ形式をしているが、ポテンシャル関数 $\psi(\boldsymbol{\theta})$ が式 3 に示すように変更されている。 $\psi'(\boldsymbol{\theta})$ は、 h_k がすべて 1 でない限り、凸関数にはならない。そのため、EEF は、指数型分布族の定義から、ポテンシャル関数 $\psi(\boldsymbol{\theta})$ の凸性に関する制約を条件付きで除外した分布族であるといえる。

ポテンシャル関数が凸関数であるかどうかに関わらず、指数型表現によってあらわされた確率分布 $P(x)$ と $Q(x)$ の間の関数内積は、 θ 座標とポテンシャル関数のみを使用して次のように表現できる（証明は Appendix の定理 1）。

$$\langle P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) \rangle = \frac{e^{\psi(\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \quad (4)$$

さらに、 $P(x)$ と $Q(x)$ が両方とも EEF である場合、 $P(x)$ と $Q(x)$ の間の関数内積と θ 座標の間には以下の関係が成り立つ（証明は Appendix の定理 2）。

$$\frac{1}{2} \log \langle P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) \rangle = \sum_{k=1} h_k \theta_k^P \theta_k^Q \quad (5)$$

ここで、 \mathbf{M} を、(i, j) 番目の要素 $m_{i,j}$ を $\frac{1}{2} \log q_{i,j}$ とする $n \times n$ の行列とする。 $q_{i,j}$ は、i 番目と j 番目の分布の間の関数内積を表す。式 5 は以下のように書くことができる。

$$\mathbf{M} = \boldsymbol{\Theta} \mathbf{S} \boldsymbol{\Theta}^T \quad (6)$$

$\boldsymbol{\Theta}$ は、(i,j) 番目の要素が EEF 式の i 番目の分布の j 番目の座標値を表す θ 座標行列であり、 \mathbf{S} は、(k,k) 番目の要素が h_k となる対角行列である。

一方で、 \mathbf{M} の i 番目の固有値を λ_i とおく。次に、 \mathbf{M} は、次のように固有値分解で表すことができる。

$$\mathbf{M} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \quad (7)$$

ここで、 \mathbf{V} の i 番目の列は \mathbf{M} の i 番目の固有ベクトルを表し、 $\boldsymbol{\Lambda}$ は i 番目の対角要素が λ_i である対角行列である。 \mathbf{M} の固有値には負の値が含まれる。次に、 $\mathbf{M} = \mathbf{V} \boldsymbol{\Lambda}' \mathbf{S}' \mathbf{V}^T = \mathbf{V} \sqrt{\boldsymbol{\Lambda}'} \mathbf{S}' (\mathbf{V} \sqrt{\boldsymbol{\Lambda}'})^T$ と変形する。ここで \mathbf{S}' 、 $\boldsymbol{\Lambda}'$ 、 $\sqrt{\boldsymbol{\Lambda}'}$ は $n \times n$ の対角行列であり、その i 番目の対角要素はそれぞれ $sign(\lambda_i)$ 、 $|\lambda_i|$ 、および $\sqrt{|\lambda_i|}$ である。したがって、式を比較することにより、

$$\boldsymbol{\Theta} = \mathbf{V} \sqrt{\boldsymbol{\Lambda}'} \quad (8)$$

$$\mathbf{S} = \mathbf{S}' \quad (9)$$

ここで、 $\mathbf{S} = \mathbf{S}'$ は $h_i = sign(\lambda_i)$ を意味している。以上より、関数内積の情報のみから EEF 表現における θ 座標の値を固有値分解によって求めることができる。

次に、 $C(x)$ と $F_i(x)$ を計算する。この計算をコンピュータで離散的に扱うためには、拡大指数型分布族表現を行列で表す必要がある。

$$\mathbf{P}^{log} = \mathbf{C} + \Theta \mathbf{F} - \Psi \quad (10)$$

ここで \mathbf{P}^{log} は、n 個のサンプルの m 個の格子の離散化された確率質量関数の要素ごとの対数をとった $n \times m$ 行列、 \mathbf{C} は、 $C(x)$ に対応する $n \times m$ 行列であり、その行の全てのベクトルが \mathbf{c} である。 Θ は既に計算されている座標行列であり、 \mathbf{F} は行ベクトルが離散化された $F_i(x)$ に対応する行列、 Ψ は、 $n \times m$ 行列で、その列ベクトルは $\sum h_k \theta_k^2 \mathbf{1}$ であり、既に計算されている。

$$\mathbf{P}^{log} = \begin{pmatrix} \log(P_{11}) & \cdots & \log(P_{1j}) & \cdots & \log(P_{1m}) \\ \vdots & \ddots & & & \vdots \\ \log(P_{i1}) & & \log(P_{ij}) & & \log(P_{im}) \\ \vdots & & & \ddots & \vdots \\ \log(P_{n1}) & \cdots & \log(P_{nj}) & \cdots & \log(P_{nm}) \end{pmatrix} \quad (11)$$

$$\mathbf{C} = \begin{pmatrix} c_1 & \cdots & c_j & \cdots & c_m \\ \vdots & \ddots & & & \vdots \\ c_1 & & c_j & & c_m \\ \vdots & & & \ddots & \vdots \\ c_1 & \cdots & c_j & \cdots & c_m \end{pmatrix} \quad (12)$$

$$\Theta = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1j} & \cdots & \theta_{1n} \\ \vdots & \ddots & & & \vdots \\ \theta_{i1} & & \theta_{ij} & & \theta_{in} \\ \vdots & & & \ddots & \vdots \\ \theta_{n1} & \cdots & \theta_{nj} & \cdots & \theta_{nn} \end{pmatrix} \quad (13)$$

$$\mathbf{F} = \begin{pmatrix} F_{11} & \cdots & F_{1j} & \cdots & F_{1m} \\ \vdots & \ddots & & & \vdots \\ F_{i1} & & F_{ij} & & F_{im} \\ \vdots & & & \ddots & \vdots \\ F_{n1} & \cdots & F_{nj} & \cdots & F_{nm} \end{pmatrix} \quad (14)$$

$$\Psi = \begin{pmatrix} \sum_{k=1} h_k \theta_{1k}^2 & \cdots & \sum_{k=1} h_k \theta_{1k}^2 & \cdots & \sum_{k=1} h_k \theta_{1k}^2 \\ \vdots & \ddots & & & \vdots \\ \sum_{k=1} h_k \theta_{ik}^2 & & \sum_{k=1} h_k \theta_{ik}^2 & & \sum_{k=1} h_k \theta_{ik}^2 \\ \vdots & & & \ddots & \vdots \\ \sum_{k=1} h_k \theta_{nk}^2 & \cdots & \sum_{k=1} h_k \theta_{nk}^2 & \cdots & \sum_{k=1} h_k \theta_{nk}^2 \end{pmatrix} \quad (15)$$

この式を整理すると、この式は以下のように表すことができる。

$$\mathbf{P}' = \Theta' \mathbf{F}' \quad (16)$$

ここで、 $\mathbf{P}' = \mathbf{P}^{\log} + \Psi$, \mathbf{F}' は $[\mathbf{F}^T, \mathbf{c}]^T$, Θ' は $[\Theta, \mathbf{1}]$ を表している。

$$\mathbf{P}' = \begin{pmatrix} \log(P_{11}) + \sum_{k=1} h_k \theta_{1k}^2 & \cdots & \log(P_{1j}) + \sum_{k=1} h_k \theta_{1k}^2 & \cdots & \log(P_{1m}) + \sum_{k=1} h_k \theta_{1k}^2 \\ \vdots & \ddots & & & \vdots \\ \log(P_{i1}) + \sum_{k=1} h_k \theta_{ik}^2 & & \log(P_{ij}) + \sum_{k=1} h_k \theta_{ik}^2 & & \log(P_{im}) + \sum_{k=1} h_k \theta_{ik}^2 \\ \vdots & & & \ddots & \vdots \\ \log(P_{n1}) + \sum_{k=1} h_k \theta_{nk}^2 & \cdots & \log(P_{nj}) + \sum_{k=1} h_k \theta_{nk}^2 & \cdots & \log(P_{nm}) + \sum_{k=1} h_k \theta_{nk}^2 \end{pmatrix} \quad (17)$$

$$\Theta' = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1j} & \cdots & \theta_{1n} & \cdots & 1 \\ \vdots & \ddots & & & \vdots & & \vdots \\ \theta_{i1} & & \theta_{ij} & & \theta_{in} & \cdots & 1 \\ \vdots & & & \ddots & \vdots & & \vdots \\ \theta_{n1} & \cdots & \theta_{nj} & \cdots & \theta_{nn} & \cdots & 1 \end{pmatrix} \quad (18)$$

$$\mathbf{F}' = \begin{pmatrix} F_{11} & \cdots & F_{1j} & \cdots & F_{1m} \\ \vdots & \ddots & & & \vdots \\ F_{i1} & & F_{ij} & & F_{im} \\ \vdots & & & \ddots & \vdots \\ F_{n1} & \cdots & F_{nj} & \cdots & F_{nm} \\ c_1 & \cdots & c_j & \cdots & c_m \end{pmatrix} \quad (19)$$

したがって \mathbf{F}' は Moore-Penrose の擬似逆行列 $Ginv(\Theta')$ を用いて以下のように書ける。

$$\mathbf{F}' = Ginv(\boldsymbol{\Theta}') \mathbf{P}' \quad (20)$$

ここで、 F' は $[\mathbf{F}^T, \mathbf{c}]^T$ と定義されているので、分布集合の EEF 表現に必要な全ての項目が得られる。

以上の理論に基づいて、確率行列 \mathbf{P} を分解し、任意の分布集合の EEF 表現を得るための簡単な行列演算に基づくアルゴリズムを構築することが可能である。入力は、確率行列 \mathbf{P} であり、その行は確率質量関数を表す。最初のステップは、 \mathbf{P} から行列 \mathbf{M} を計算する。第2ステップは、 \mathbf{M} の固有値分解である。 h_i を求めて、 $\psi'(\boldsymbol{\theta})$ を決定し、全てのサンプルを埋め込むために、n 個のサンプルの座標行列 $\boldsymbol{\Theta}$ を求める。第3ステップでは、 \mathbf{c} と \mathbf{F} を計算し、EEF 表現式の全構成要素を決定する。

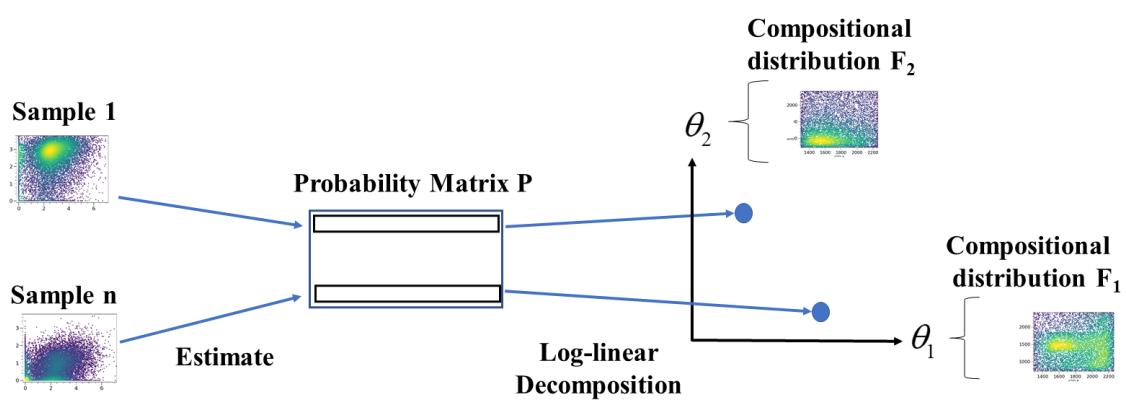


図 1 Fig 1.DEEF のアウトライン

この方法を分布セットに適用することで、各分布を EEF 空間に埋め込むことができ、分布の特徴として θ を得ることができる。 θ 座標は、固有値分解から計算されるので、固有値の上位にあるいくつかの座標が確率分布セットの情報を多く持っていることになる。また、 \mathbf{F} 行列は、元の空間における構成分布を与える。本アルゴリズムは、R パッケージ "deef" として GitHub (<https://github.com/DaigoOkada/deef>) でソフトウェアとして公開している。

1.3 シミュレーションデータ解析

提案手法のパフォーマンスを検証するために、指数族分布族の分布のセットと、指數型分布族の分布のパラメトリックな混合分布のセットに DEEF を適用した。まず、理論的な分布の 4 つのセットを生成した。4 つのセットを 2D、Random、1D、Mixture と呼ぶ(図 2)。分布セット 2D は 900 個の正規分布で、平均値は -1 から 1、sd は 2 から 4 の範囲を等間隔に区切ることで生成した。分布セット Random は、2D からランダムにサンプリングした 50 個の分布で構成されている。1D は、2D と同じ空間に 1 次元の多様体を作る正規分布セットである。Mixture は 2 つの正規分布の混合分布であり、一方の正規分布は $N(-1,1)$ であり、もう一方の正規分布は平均値が 4~5、sd が 2~4 である 900 個の分布のいずれかから構成されている。2 つの分布の混合比は 0.5 とした。離散化のグリッド数は 10,000 とした。離散化の範囲は、全分布の 0.5 パーセンタイルから 99.5 パーセンタイルまでの区間が含まれるように決定した。平均値と sd で定義された関数内積の値を用いて、DEEF の θ 座標を計算した。DEEF では、 C, F_i, θ_i, h_i が得られれば、任意の数の座標を用いて分布を再現することができる。上位 K 個の絶対固有値を持つ座標を選択し、 $\hat{P}(x, \boldsymbol{\theta}) = \exp(C(x) + \sum_{k=1}^K F_k(x)\theta_k - \psi(\boldsymbol{\theta}))$ を総和 1 に正規化することで確率質量関数を再構成し分布を再現した。本研究では、真の確率質量関数と再構成された確率質量関数の誤差の二乗和で定義される指標 (Performance Index: PI) で性能を評価した。ここで、二乗誤差が小さいほど再現性が高いことを示す。特に、この値が 0 の場合、元の分布と再構成された分布は全く同じであることを意味する。

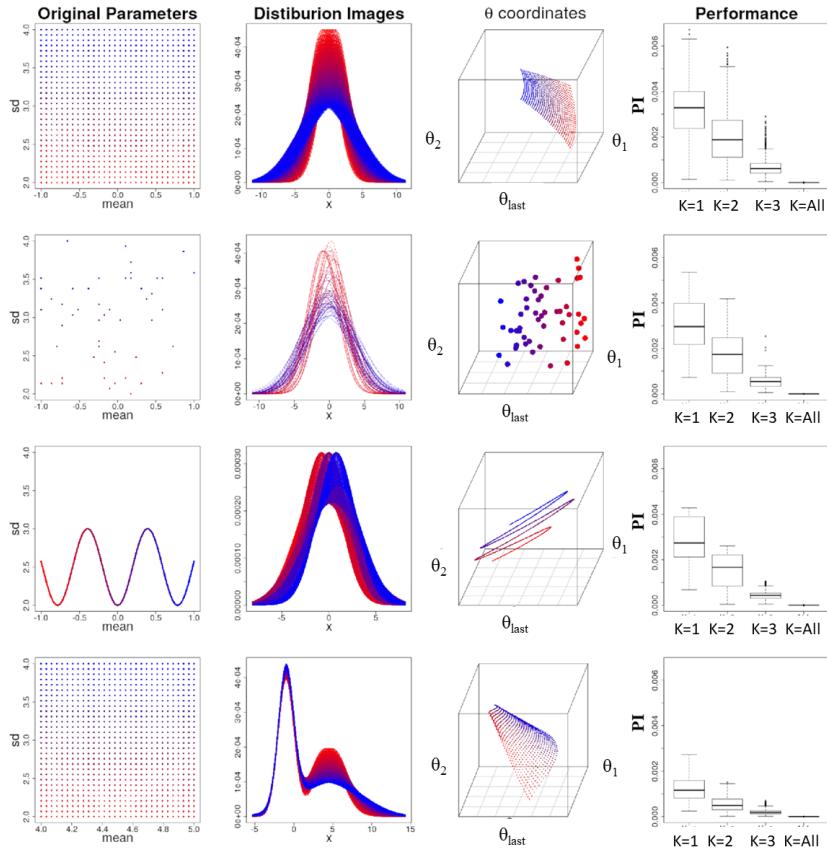


図 2 4 つの分布セットにおける元のパラメータ構造 (1 列目), 分布セットの描画 (2 列目), θ 座標 (3 列目), およびパフォーマンスの箱ひげ図 (4 列目). θ 座標空間に埋め込むことで、元のパラメータ構造が歪んだ形で再現されている。第 4 列目のパネルは、各分布セットについて、絶対固有値の高い上位 K 座標のみを用いて再構成された分布との二乗誤差の和で定義された PI の箱ひげ図を示しており、 K が増加するにつれて、再構成された分布セットは、元の分布セットに近づいている。すべての θ 座標が使われるとき、再構成された分布セットに属するすべての分布が元の分布と同じになる。

これらの 4 つの分布セットに DEEF を適用した結果を図 2 に示し、各 θ 座標に対応する固有値を図 3 に示した。すべての分布セットについて、最大固有値は負であった。 θ 座標は、固有値の降順で θ_i で表した。

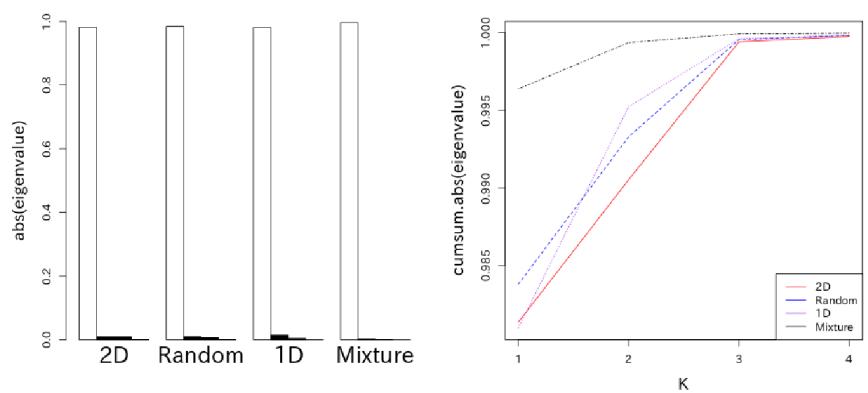


図 3 各分布セットでの固有値プロット。左図は合計値が 1 になるように標準化した固有値の絶対値を示しており、黒棒は正の固有値、白棒は負の固有値を表す。すべての分布セットにおいて、最大固有値は負であった。右図はその累積和を示している。

θ_{last} は、絶対値が最大である最小固有値に対応する座標である。この方法を使用して計算された固有値には、負の値が含まれる。多くの場合では上位 2 つまたは 3 つの固有値のみが意味のある寄与を持っており、それ以下はノイズと考えられる。分布セット 2D における異同を説明するために使用されるパラメーターの数は既知であり、2 (mean および sd) である。同様に Random、1D、および Mixture では、それぞれ 2、1、および 2 であり、これらの数は、意味のある絶対値を持つ正の固有値の数に対応している。

図 2 の 3 番目の列では、すべての分布を上位 3 つの絶対固有値を持つ 3 次元 θ 座標空間に埋め込んだ結果である。 θ 座標空間では、カラーパターンで示される元のパラメーター構造が、4 つのセットすべてで維持されており、分布が元のパラメーター構造の次元の多様体に埋め込まれていることも確認できる。

次に、各分布セットのにおいて $C(x)$ と $F(x)$ を調べた。図 4 は、2D、Random、1D、および Mixture について計算された $C(x)$ 、 $F_{last}(x)$ 、 $F_1(x)$ 、および $F_2(x)$ を示している。分布セット 2D の場合、 $C(x)$ は凸型で、 x 座標の中心にピークがある。これは、分布セット 2D の平均的なパターンと考えられる。 θ_1 と θ_2 は、正の固有値に対応する座標を表す。 $F_1(x)$ は、 θ_1 の値が大きいほど、両端の確率質量が大きくなることを示している（図 4 (a)、青線）。これは、分布 Normal (-1, 4) および Normal (1, 4) が図 2 の最大の θ_1 座標値を持つ領域に埋め込まれたという事実と一致している。 $F_2(x)$ は、 θ_2 の値が大きいほど、右端の確率質量が大きくなり、左端の確率質量が小さくなる（図 4 (a)、紫線）。 θ_2 の値が最大の分布は Normal (1, 2) である。これは、図 2 の分布セット 2D で mean が最大で sd 値が最小の分布である。これらの結果は、 $F_i(x)$ 、各 θ 座標が元の分布で差異のどの部分を説明するかについての情報を示している。 $F_{last}(x)$ は、最大の負の固有値を持つ座標である θ_{last} に対応するが、 x 軸にほぼ平行であり、分布の特徴に関する情報がほとんど保持されていなかった（図 4 (a)、赤線）。このような負の固有値に対応している軸は、内積と多様体上の点間の距離の歪みに関連していると考えられる。分布セット Random の場合も、2D の場合とわずかな歪みはあるものの同様の結果が得られた（図 4 (b)）。興味深いことに、分布セット 1D の $F_1(x)$ と $F_2(x)$ は、それぞれ $F_2(x)$ と $F_1(x)$ によく似ている（図 4(c)）。また、分布セット Mixture の場合、エッジ領域で非常に大きな値が推定される傾向があった（図 4 (d)）。Mixture (sub) は、Mixture の中央部分の拡大図であり、 $C(x)$ が混合正規分布の二峰性を捉えていることを示している。各 $F_i(x)$ は、正規分布セットの場合と同様に、分布セット内の異同を説明する複雑なパターンを表している。

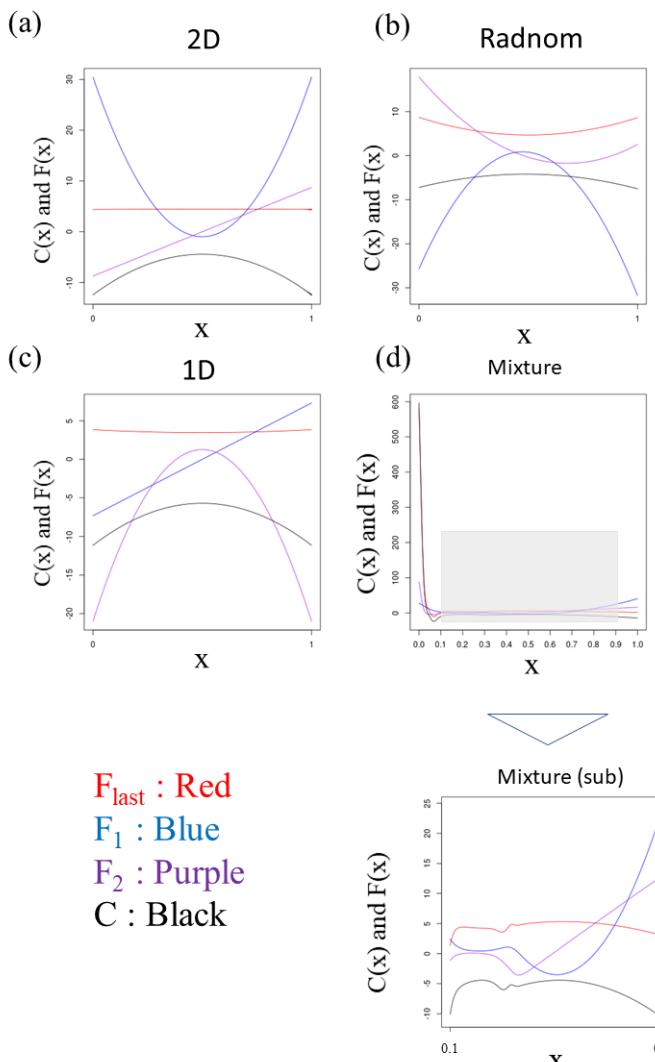


図 4 分布セット 2D, Random, 1D, Mixture における $C(x)$, $F_{\text{last}}(x)$, $F_1(x)$, $F_2(x)$. 。x 軸の範囲は 0 から 1 にスケーリングされている。Mixture での全領域の 0.1~0.9 の間のプロット (灰色の矩形で示されている) を抽出し、Mixture(sub) とした。

最後に、上位の θ 座標を用いて分布を再構成し、PI を用いて性能を評価した。座標数が多くなるほど、再構成された分布は全体として元の分布に近づく傾向にあった(図 2 の 4 列目)。全ての座標を用いれば、元の分布を正確に再現することができる。

分布セット 2D、Random、1D、および Mixture の固有値プロットは、有意な正の固有値の数が分布セットの自由度であることを意味していた。つまり、2D、Random、1D、および Mixture の元のパラメーター構造の自由度は、それぞれ 2、2、1、および 2 であり、有意な正の固有値の数に対応していることが示された。そこで、他の混合正規分布セットに対して DEEF を適用し、分布の潜在的なパラメータの自由度を推定できるかどうかを定量的に調査した。混合成分の数が 2 から 10 の間で変更し、すべての成分の正規分布は同じ $sd (=1)$ をとり、構成成分の正規分布の平均値は、-10 から 10 の間で等間隔に配置した。有意な正の固有値の数は、合計がすべての正の固有値の合計の 90% を超える固有値の最小数として定義した。分布セットは 100 の分布で構成され、その中で混合比のみが異なるような分布セットである。この場合、混合比のみが変化したため、分布セットの潜在的な自由度は成分数-1 として定義される。

図 5 は、真の自由度と無視できない正の固有値の数のプロットを示している。このプロットは、成分の分布の数が少ない場合、自由度が無視できない正の固有値の数に対応することを示している(赤線)。一方で、パラメータの自由度が大きい領域では、必要な固有値の数が予測される数よりも減少した。これは、データセットの複雑さに比べてサンプルサイズが比較的小さいため、提案手法による分解の解像度が不十分であったためと考えられる。この結果は、提案手法がデータセット内の意味のある情報の量に基づいて、データセットの潜在的な自由度を識別できることを示唆している。

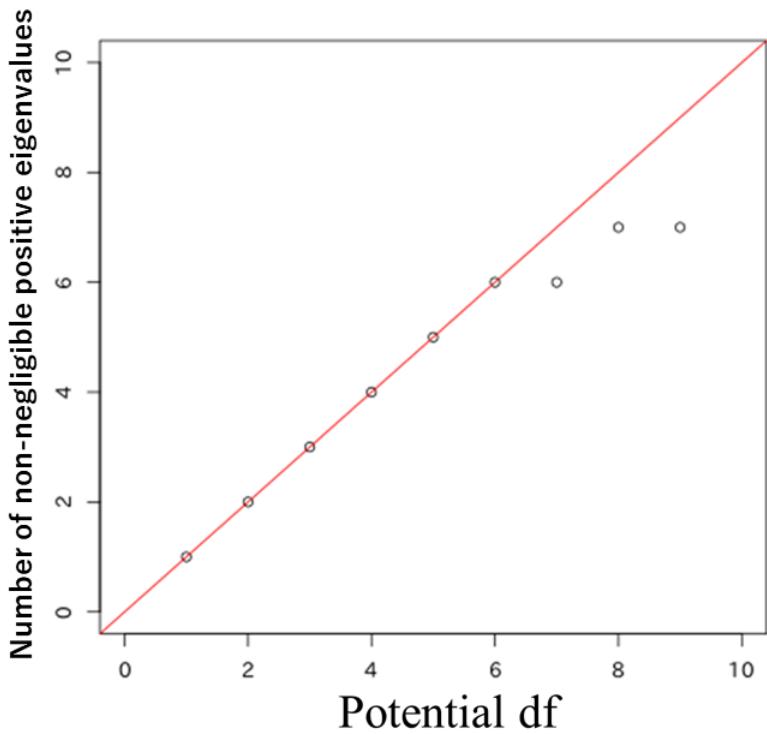


図 5 潜在的な自由度と有意な正の固有値の数との co-plot. 混合成分の数は 2 から 10 まで変化させた。パラメータの自由度は混合成分の数-1 と定義することができ、有意な正の固有値の数は分布セットの自由度を反映している一方で、提案手法の分解能が限られているため、自由度が大きい領域ではこの関係は満たされていないと考えられる。

これらの一連の理論的な分布に対する結果は、提案手法が指数型分布族においてもそうでない分布族においても、分布特徴量およびその特徴量が説明する構成分布をそれぞれ θ 座標、 $F(x)$ 関数として同定できることを示している。

2 実データ解析

次に、サイトメトリー実データに対して DEEF を適用した。

2.1 材料と方法

2.1.1 サイトメトリーデータからの確率行列 P の構築

実データでは、シミュレーションデータとは異なり、データ点が得られた母集団の分布が不明である。そこで、DEEF を適用する前に、母分布となる細胞集団プロファイルの確率行列 P を以下のように推定した。各サイトメトリーデータは、未知のマーカーの数 (d) 次元の母集団分布からのサンプルである。まず、各マーカーの範囲を決定した。各サンプルについて、各マーカー発現の α パーセンタイルと $1-\alpha$ パーセンタイルを計算した。すべてのサンプルが、 α パーセンタイルと $1-\alpha$ パーセンタイルの間の発現範囲を含むように、すべてのサンプル間の最小の α パーセンタイル値と最大の $1-\alpha$ パーセンタイル値の間に対象範囲を設定した。次に、この範囲を等間隔の m 個のグリッドで分割した。グリッドの数は、 m^d である。得られたグリッドについて、k 最近傍 (k-nearest neighbor: kNN) 法を用いて確率密度を推定した。 m^d グリッドの kNN 法に基づく密度を表すベクトルは、その和が 1 となるように正規化した。

2.1.2 EGF stimulation data への適用

EGF 刺激のヒト乳腺上皮細胞の培養細胞への影響に関する研究で用いられたマスサイトメトリーデータを使用した [2]。データは Flow Repository (ID:FR-FCM-ZYBC) から入手した。実験では、EGF 刺激後と対照条件をそれぞれ 2 回ずつの Replicate で、それぞれ 10 タイムポイント (0, 0.5, 1, 3, 6, 10, 15, 30, 60, 120 分) の測定を行っている。そのため、データセットは合計 40 サンプルからなる (2 条件 x2 反復 x10 時点)。マーカータンパク質は、論文中で EGF 刺激に反応することが示されている pAkt, pERK, pS6, pPLC γ 2 の 4 つのタンパク質を選んだ。前処理として、マーカーの発現レベルを原論文と同様に $\text{asinh}(\text{発現値}/5)$ を用いて変換した。このデータセットの細胞数は 8,089 個から 22,221 個の範囲であった。サイトメトリーデータから確率行列 P を構築した。こ

の時の、確率行列 \mathbf{P} を構築するためのハイパーパラメーターは、 $m = 20$, $\alpha = 0.05$, $k = 800$ である。次に、DEEF を \mathbf{P} に適用した。 θ 座標は、固有値の高い順に $\theta_1, \theta_2 \dots \theta_{last}$ とした。

2.1.3 SPADE による $F(x)$ 関数の可視化

複数マーカーでの発現プロファイルを同時に可視化するために、サイトメトリーデータに SPADE アルゴリズムを適用した [3]。クラスター数は 10 個、その他のハイパーパラメーターは原論文と同じものを使用した。最小全域木を作成するステップでは、R パッケージ “ape” の mst 関数を使用した。クラスタリングステップでは、完全連結法を用いた。得られたクラスターのマーカー発現の代表値は、ツリー上の各クラスターに属する細胞の中央値とした。

2.1.4 EGF 刺激データの次元削減とタイムコースの再構成

DEEF は固有値の絶対値が上位の θ 座標のみを用いて分布を再構成することができる。細胞集団プロファイルの次元を小さくするために、主要なパターンの合成和のみを用いて細胞集団プロファイルを再現した。再構築された分布 ($K=3$) は、シミュレーションデータ解析に記載されている手順を使用して得た。4 つのマーカー (pAKT、pERK、pS6、および pPLC γ 2) のそれぞれについて、元のマーカー発現と再構成されたマーカー発現の発現値の中央値の変化を可視化した。元のマーカー発現については、各サンプルについて、細胞の発現値から各マーカーの発現値の中央値を算出した。再構築されたマーカー発現については、再構築された分布から着目したマーカー以外の全てのマーカー (3 個) を消去することで算出した。そして、グリッド間の値を線形補間することにより、対象マーカーの発現値の 50 パーセンタイル値を中心値として推定した。次に、Replicate1 の EGF 刺激データセットに対してタイムコース再構成を行った。各時間点における θ 座標の値は、各時点間の θ 座標の値を線形補間して 10 等分し、合計 91 点の時点における θ 座標を再構成することで推定した。この推定値をもとに、 $K=3$ の分布を計算し、動画として再現した。

2.2 結果

2.2.1 θ 座標と $F(x)$ 関数

Replicate1 と Replicate2 について、EGF 刺激後の各時点の前処理済みマーカー発現データを図 6 に示す。構築された \mathbf{P} に対して、DEEF を適用し、対応する θ 座標を計算

した。

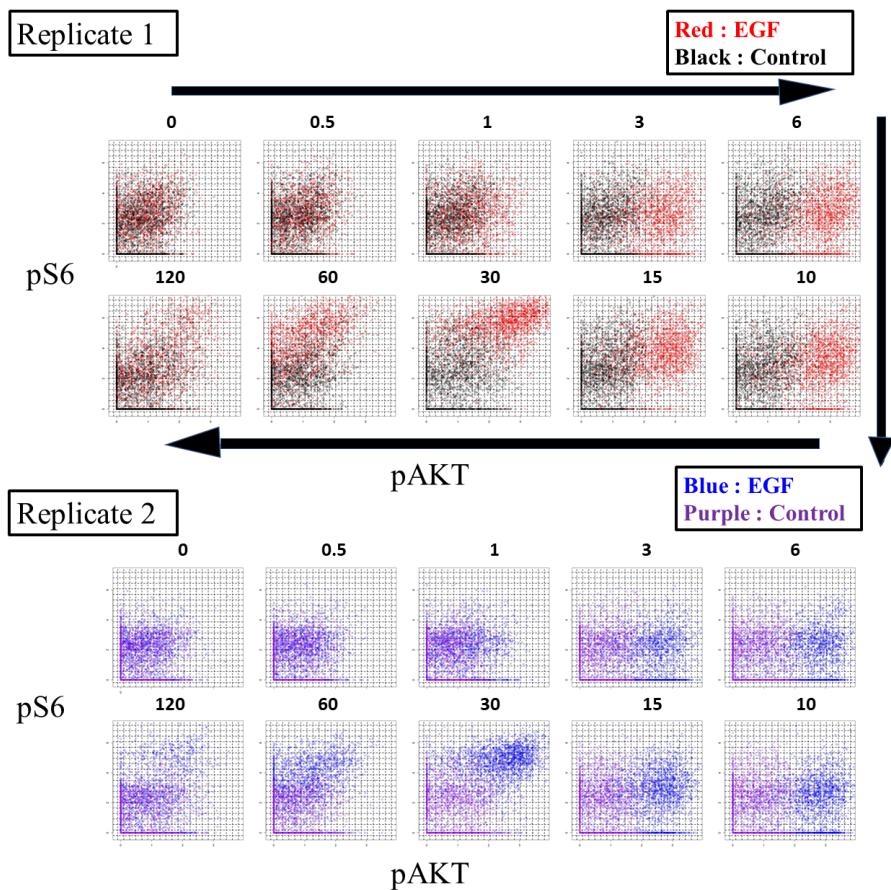
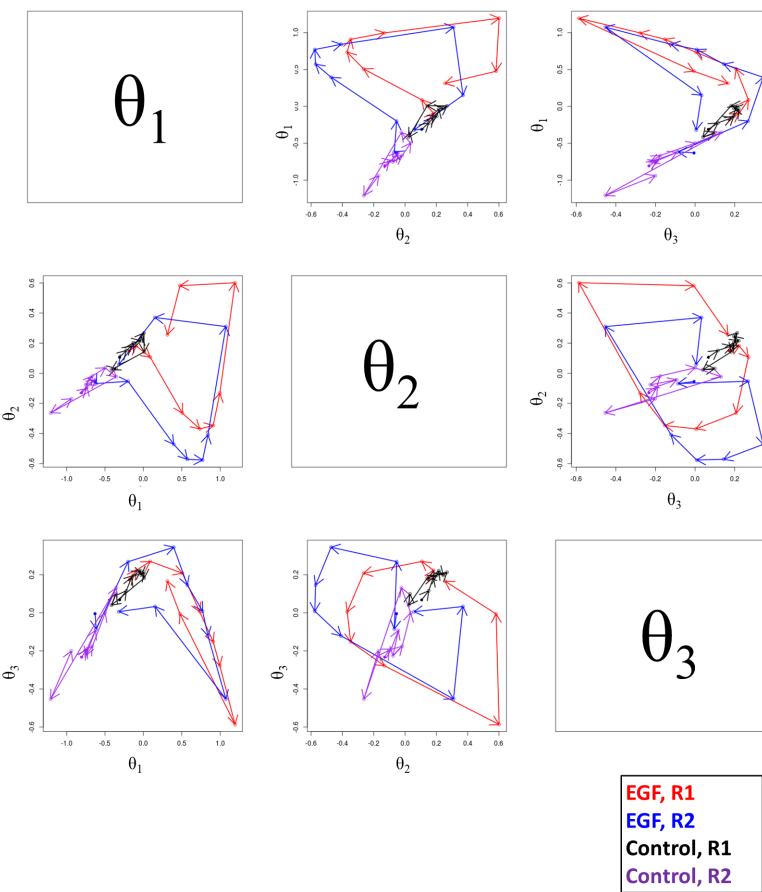


図 6 pAKT と pS6 の 2 マーカーでの発現値の散布図。各レプリケートと条件について、ランダムに選択された 2,000 個の細胞をプロットしている。黒い点線はグリッドを表している。細胞集団のプロファイルは、EGF 刺激後に動的に変化するが、生データから定量的にこれを捉えて評価することは困難である。

全ての細胞集団を低次元の座標空間に埋め込み、DEEF を用いて可視化した。正の固有値の和のうち、 θ_1 , θ_2 , θ_3 がそれぞれ 69.6%, 13.9%, 8.9% を占めていた。図 7(a) は、DEEF で得られた上位の正の座標の散布図である。EGF 刺激後、細胞集団の分布は、 θ_1 と θ_2 の座標空間上を移動し、ベースライン付近の領域に戻っている。 θ_1 と θ_2 は、Replicate1 と Replicate2 の間で EGF 刺激によって共通の軌道を示すが、 θ_3 は異なる軌道を示す。そこで、 θ_1 と θ_2 の 2 パラメーターを用いて、Replicate1 と Replicate2 に共通する EGF 刺激後の細胞集団のダイナミクスをパラメータ化した。

(a)



(b)

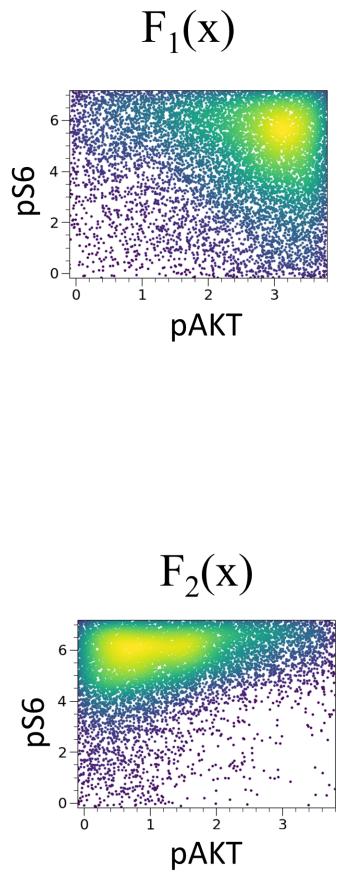


図 7 EGF stimulation data への DEEF の適用結果 細胞全体のダイナミクスを可視化し、違いを説明する支配的なパターンを構成分布として抽出した。(a) θ_1 , θ_2 , θ_3 の座標プロット。(b) F_1 と F_2 を pAKT と pS6 について描画した結果。密度プロットは、標準化された $\exp(F_i)$ からランダムにサンプリングされた 10,000 個のデータポイントから生成された。

F_1 と F_2 は、それぞれ θ_1 と θ_2 に対応しており、軌跡が表す細胞集団の変化のパターンを示している。図 7(b) は、pAKT と pS6 の F_1 と F_2 を示している。 F_1 は pAKT の発現が高く pS6 の発現も高い細胞の密度、 F_2 は pAKT の発現が低く pS6 の発現が高い細胞の密度を説明している。 F_1 の増加と θ_2 の減少が EGF 刺激後初期の応答に対応している。後半に起こる θ_2 の増加は、pAKT よりも遅れて発生した pS6 の増加に対応している。

そこで、4 次元関数である $F(x)$ を可視化するために、SPADE 解析を行い SPADE ツリー上に F_1 と F_2 を記述した。40 サンプルすべてに SPADE を適用し、10 個の細胞クラスターからなる SPADE ツリーを作成した(図 8(a))。各 SPADE クラスターは、4 種類のマーカーの発現パターンによって特徴付けることができる(図 8(b))。図 8(c) は、 F_1 と F_2 の値を持つ SPADE ツリーを示している。各クラスターには、クラスターの代表的な位置が属するグリッドの F_1 と F_2 の値が割り当てられている。SPADE ツリー上の F_1 では、クラスター 9 が最も大きい正の F_1 値を持っていた。クラスター 9 は 4 つのマーカーすべてに高い発現を示したので、この結果は妥当である。この結果は、すべてのマーカーの発現が EGF 刺激後に増加するという事実に対応する。クラスター 8 は、最も高い負の F_1 値を有し、これは、このクラスターが 4 つのすべてのマーカーについて低い発現を示したので、妥当である。一方で、 F_2 は、 F_1 とは異なる軌跡パターンに対応しており、SPADE ツリー上では異なるパターンを示している。 F_2 の正の値が最も高いクラスター 3 では、pS6 と pPLC γ 2 の発現が高く、この 2 つのマーカーは、pAKT と pERK よりも遅れて発現している。興味深いことに、pERK と pS6 について低い発現を示したクラスター 2 は、最も高い負の F_2 値を有している。各クラスターの代表値の表を用いると、このサブセットは刺激の 6 分後に得られたサンプルの密度プロットで確認することができる(図 8(d))。DEEF を用いることで、従来の方法では検出が困難なパターンについての情報を得ることができる。

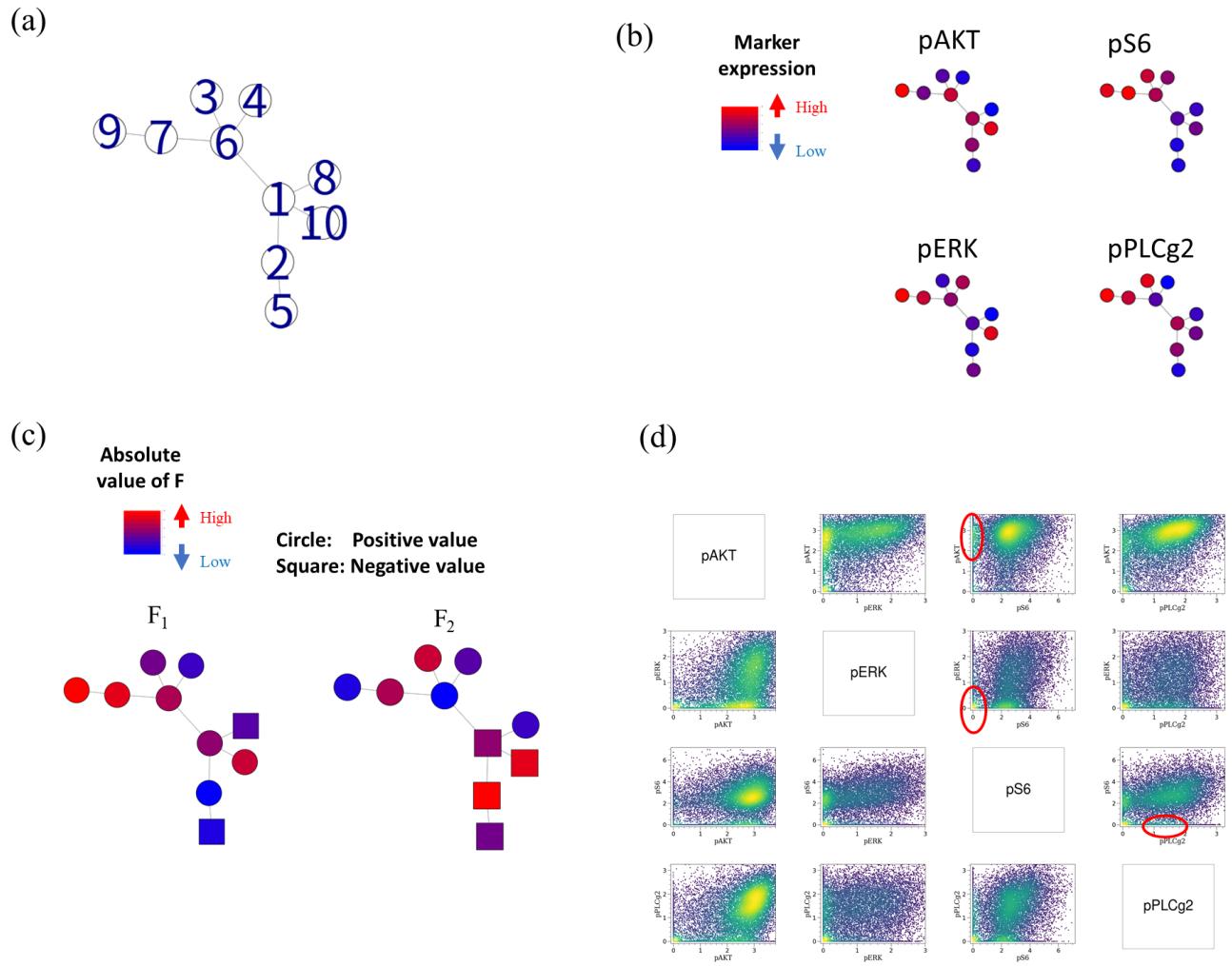


図 8 SPADE ツリー上での F_2 および F_3 の描画 (a) SPADE ツリーのラベル。 (b) 4 つのマーカー発現値で色付けした SPADE ツリー。(c) F_1 と F_2 の値で色付けした SPADE ツリー。(d) SPADE ツリーのクラスター 2 に対応する領域。クラスター 2 の対応する領域は、Replicate1 の EGF 刺激 6 分後に得られた 4 つのマーカーの密度プロット上の赤丸で示されている。

2.2.2 次元削減と時系列の再現

DEEF は、絶対固有値の上位の座標のみを用いて分布を再構成することができる。細胞集団プロファイルの次元削減として、上位 3 つの絶対固有値 ($K=3$) に対応する、 θ_1 , θ_2 , θ_{last} による合成和のみを用いて細胞集団プロファイルを表現した。図 9(a) の 1 列目のパネルは、4 つのマーカーの時間経過に沿った生データにおけるマーカー強度の中央値の変化を示している。pAkt と pERK の発現量が最初に増加し、次いで pS6 と pPLC γ 2 の発現量が増加しており、これは原論文で得られた結果と一致している。図 9(a) の 2 列目のパネルは、再構成された分布から計算されたマーカー強度の中央値の変化を θ_1 , θ_2 , θ_{last} を用いて示している。これらの結果は、主要なパターンのみを用いて再構成された細胞集団プロファイルが、元のデータのダイナミクスの特徴をよく捉えていることを示唆している。DEEF が情報を意味のあるデータとノイズに分解できると考えると、主構成分布のみを用いて再現することで、データをノイズを除去することができると考えられる。

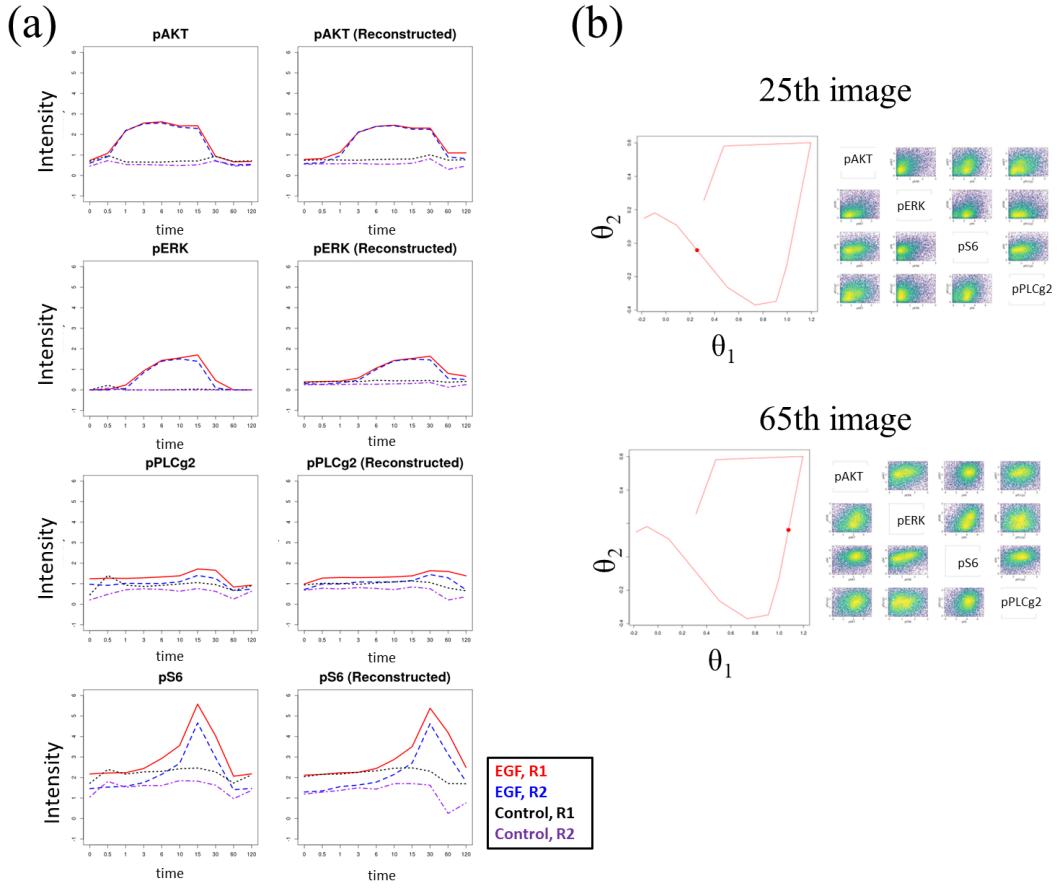


図 9 DEEF を用いた細胞集団プロファイルの次元削減の結果. (a) 各マーカー (pAKT、pERK、pPLC γ 2、および pS6) について、時間経過に沿ったマーカー発現の変化を表示している。1列目のパネルは、各マーカー発現の中央値である。2列目のパネルは、 θ 座標の上位 3つ、すなわち、負の固有値が最も高い θ_{last} と、正の固有値が最も高い θ_1 と θ_2 を用いて再現した分布の中央値である。(b) EGF 刺激後の Replicate1 の測定値間の推定細胞集団プロファイルの例として、91 枚の画像の 25 番目と 65 番目の画像を示している。 θ 座標空間の対応する点を赤い点で示している。

次に、このスキームを用いて、元のタイムコースが 10 個のタイムポイントからなる Replicate1 の EGF 刺激データのタイムコース再構成を行った。各タイムポイント間の θ 座標の値を線形補間して 10 等分し、合計 91 枚の画像で θ 座標を再構成することで、各タイムポイントにおける θ 座標の値を推定した。図 9(b) は、測定間の細胞集団プロファイルの推定値の一例として、91 枚の画像のうち、25 枚目と 65 枚目の画像を示している。推定値をもとに、K=3 での分布を再現している。未測定の時点を含む細胞集団の一連の動態はアニメーションで表示した (https://github.com/DaigoOkada/DEEF_press_gif/issues)。

3 Appendix

Theorem 1 $P(x, \boldsymbol{\theta}^P)$ and $Q(x, \boldsymbol{\theta}^Q)$ が指数型分布族の場合、以下の関係が成り立つ。

$$\langle P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) \rangle = \frac{e^{\psi(\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}}$$

Proof: 指数型分布族の定義より、確率密度関数は以下である。

$$\begin{aligned} \log P(x, \boldsymbol{\theta}^P) &= \sum_{i=0} F_i(x) \theta_i^P - \psi(\boldsymbol{\theta}^P) \\ \log Q(x, \boldsymbol{\theta}^Q) &= \sum_{i=0} F_i(x) \theta_i^Q - \psi(\boldsymbol{\theta}^Q) \\ F_0(x) &= C(x), \theta_0 = Const \end{aligned}$$

ここで、両者の内積を取ると、

$$\begin{aligned} \langle P(x, \boldsymbol{\theta}^P), P(x, \boldsymbol{\theta}^Q) \rangle &= \int P(x, \boldsymbol{\theta}^P) P(x, \boldsymbol{\theta}^Q) dx \\ &= \int e^{\sum_{i=0} F_i(x) \theta_i^P - \psi(\boldsymbol{\theta}^P)} e^{\sum_{i=0} F_i(x) \theta_i^Q - \psi(\boldsymbol{\theta}^Q)} dx \\ &= \int e^{\sum_{i=0} F_i(x) (\theta_i^P + \theta_i^Q) - (\psi(\boldsymbol{\theta}^P) + \psi(\boldsymbol{\theta}^Q))} dx \\ &= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) (\theta_i^P + \theta_i^Q)} dx \end{aligned}$$

$\boldsymbol{\theta}^{P+Q}$ を $\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q$ とおくと、さらに以下のように式変形できる。

$$\langle P(x, \boldsymbol{\theta}^P), P(x, \boldsymbol{\theta}^Q) \rangle = \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) \boldsymbol{\theta}^{P+Q}} dx$$

$\int P(x, \boldsymbol{\theta}^{P+Q}) dx = \int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} dx = 1$ より以下の式が成り立つ。

$$\int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q}} dx = e^{\psi(\boldsymbol{\theta}^{P+Q})}$$

したがって、指指数型分布族同士の内積は以下のように表現できる。

$$\begin{aligned} < P(x, \boldsymbol{\theta}^P), P(x, \boldsymbol{\theta}^Q) > &= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q}} dx \\ &= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} e^{\psi(\boldsymbol{\theta}^{P+Q})} dx \\ &= \frac{e^{\psi(\boldsymbol{\theta}^{P+Q})}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} dx \\ &= \frac{e^{\psi(\boldsymbol{\theta}^{P+Q})}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \\ &= \frac{e^{\psi(\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \end{aligned}$$

Theorem 2 $P(x, \boldsymbol{\theta}^P)$ と $Q(x, \boldsymbol{\theta}^Q)$ が拡大指指数型分布族であるとき、以下の関係式が成り立つ。

$$\frac{1}{2} \log < P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) > = \sum_{k=1} h_k \theta_k^P \theta_k^Q$$

Proof: P と Q が指指数型分布族ではないが拡大指指数型分布族であった場合、ポテンシャル関数 $\psi(\theta)$ は $\psi'(\theta) = \sum_{k=1} h_k \theta_k^2$ と表せる。

Theorem 1 は拡大指指数型分布族においてもなりたつため、

$$\begin{aligned} \frac{1}{2} \log < P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) > &= \frac{1}{2} (\psi(\boldsymbol{\theta}^{P+Q}) - \psi(\boldsymbol{\theta}^P) - \psi(\boldsymbol{\theta}^Q)) \\ &= \frac{1}{2} \left(\sum_{k=1} h_k (\theta_k^P + \theta_k^Q)^2 - \sum_{k=1} h_k (\theta_k^P)^2 - \sum_{k=1} h_k (\theta_k^Q)^2 \right) \\ &= \sum_{k=1} h_k \theta_k^P \theta_k^Q \end{aligned}$$

参考文献

- [1] Daigo Okada and Ryo Yamada. Decomposition of a set of distributions in extended exponential family form for distinguishing multiple oligo-dimensional marker expression profiles of single-cell populations and visualizing their dynamics. *PLoS one*, Vol. 15, No. 4, p. e0231250, 2020.
- [2] David JHF Knapp, Nagarajan Kannan, Davide Pellacani, and Connie J Eaves. Mass cytometric analysis reveals viable activated caspase-3+ luminal progenitors in the normal adult human mammary gland. *Cell reports*, Vol. 21, No. 4, pp. 1116–1126, 2017.
- [3] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, Vol. 29, No. 10, p. 886, 2011.

総合考察

本研究では、第 1 章ではシングルセルデータに拡張した遺伝疫学モデルの観点から、細胞集団プロファイルがオミックス状態空間の分布であることおよびそこからの特微量抽出が重要な課題であることを示した。第 2 章では、代表的なシングルセル発現データであるサイトメトリーデータを取り上げ、そこから特微量抽出を行う既存の手法について代表的な手法を紹介し、その課題を提示した。第 3 章では、確率分布同士の異同やその幾何学的構造を調べる数学の分野である情報幾何について、その基本的な概念を紹介した。第 4 章では、情報幾何の知見を用いた細胞集団プロファイルからの新規の特微量抽出手法である DEEF を提案した。

大部分のシングルセル発現データ解析手法は、1 つのサンプルでのシングルセル同士の比較や分類を目的としているのに対して、DEEF は複数のサンプルについて細胞集団プロファイル全体の異同を定量し特微量抽出を行う。この手続きは、それぞれのサンプルごとにノンパラメトリックに母分布を推定し、複数サンプルの推定された母分布のセットから分布間の内積行列を計算し、行列演算ベースのアルゴリズムを用いて分解する。ここで取り出される θ 座標は細胞集団プロファイルのデータドリブンな特微量となっている。加えて、MDS ベースの方法とは異なり、 θ 座標が元の空間をどのように説明するかの対応関係が $F(x)$ 関数によって明瞭に計算される。上記の特性から、DEEF によって情報のロスが少なく解釈性の高い特微量抽出が実施できる。

DEEF は分布を対象とした次元削減手法として位置づけられる。サンプル x 特微量の行列を対象とした次元削減手法は数多く存在する。例えば、主成分分析ではサンプル x 特微量の行列から特微量 x 特微量の分散共分散行列または相関行列をつくり、これを固有値分解することを通じて、各サンプルに主成分スコアを与える。そのため、特微量の数が巨大な行列の場合は計算が難しくなる。一方で、DEEF はサンプル \times サンプルの内積行列の固有値分解を解くので特微量数が大きくても対応可能である。また、内積計算および、一般逆行列で F を求めるステップは部分ごとに分けて計算可能なので、特微量数が巨大であっても並列計算がしやすいというメリットがある。分布はその次元が無限大であり、その情報をノンパラメトリックに表現するためには非常に多くの次元が必要になる。したがって、複数の分布を対象として分布そのものの次元削減を行う場合、DEEF はよい選択肢になると考えられる。

DEEF は、確率分布の幾何学的な性質に関する示唆を含んでいる。DEEF のアイデアのベースにある情報幾何学では、確率分布空間の幾何学的特性が広く研究されてきた。特に、指數型分布族は、双対平坦性を備えた特別な多様体に埋め込むことができる事が知られており、これらの座標系の幾何学的特性はよく調べられている。EEF のポテンシャル関数は

凸ではなく、虚軸を持つ。特に、内積行列が確率質量関数から計算される場合は、負の固有値が現れる。そのため、EEF 空間の解釈は、通常の指数型分布族の場合での情報幾何学空間の解釈よりも複雑になる。EEF 空間の特徴の 1 つは、分布が割り当てられていない部分空間があることである。内積が不確定な空間の一例は、特殊相対性理論と深い関係があるミンコフスキ空空間がある。これらの特徴によって定義される EEF がつくる多様体の特性は、今後さらなる検討を要する。理論的幾何学の観点から EEF の性質を検討することは今後の課題である。

現在の DEEF の大きな課題の 1 つは、DEEF が数マーカー程度までのデータしか扱えないことである。マーカー数の次元が上がると考慮すべきグリッド数が指数関数的に増加し、kNN による密度推定の計算コストが高くなるため、ゲノムワイドな遺伝子発現を扱うことができない。この課題の解決法としては、カーネル法などを活用し、データ点から内積行列を直接計算するように手法を改良することなどが考えられる。そのような改良が実現すれば、DEEF は数マーカーのサイトメトリーデータセットだけでなく、数千-数万の遺伝子発現を測定する single cell RNA-seq データセットにも適用可能な手法になる。

DEEF は主としてサイトメトリーデータを対象に細胞集団プロファイルの特微量抽出に用いたが、この他にも分布として定義することで定量的な解釈がしやすくなる形質が数多くある。例えば、エピゲノムプロファイルは次世代シーケンサーから出力されるショートリードのゲノム上の 1 次元分布として表現される。これは、分布として解釈することが可能である。また、時系列イメージングデータにおける細胞の動きも、多数の細胞のベクトルの分布として特徴づけることができる。本研究で用いたアプローチは、1 細胞発現データのみならず、生命科学における分布型データに対しての広い適用可能性があるノンパラメトリック解析手法である。

このような分布型の形質の他にも、生命科学で取り扱う生命現象・病理現象には単なる数値ベクトルとしてでは表現できない形質は色々あり、そのようなデータからの特微量抽出は計算生物学の主要課題の 1 つである。例えば、画像データからの特微量抽出はバイオイメージング分野で様々な手法が提案されている。他の例では、細胞の 3 次元形データは球化処理および球面調和関数展開で特微量抽出することができる。実験技術の進歩で単なる数値ベクトルとしてそのまま解釈できない様々な大規模データが産生されており、生命科学データからの特微量抽出手法の開発は今後の計算生物学分野の重要なトピックである。分布型データからの特微量抽出も、このような近年の大規模化・複雑化している計算生物学研究の潮流の中に位置づけられる。