博士論文

統計的機械学習と
マイクロアレイデータ解析への応用

2011年

石川 勇太

目次

図目次

表目次

第 1 章

はじめに

本稿では, 統計的機械学習の 1 分野である密度推定問題, および, 統計的機械学習の応用例として, マイクロアレイデータ解析を取り扱う. 密度推定問題とは, 観測されたデータから背後の確率的構造, すなわち, 確率密度関数を推定する問題である (図 1.1). 確率密度分布の推定には, 大きく分けて 3 つのアプローチが存在する. 1 つ目は, 確率密度関数をパラメータを用いてモデル化し, データに最もよく合うパラメータを推定する方法で, パラメトリックモデルと呼ばれる. パラメトリックモデルの例として, 最尤法, ベイズ推定などが挙げられる. 2 つ目は, 確率密度関数をパラメトリックにモデル化せず, データに依存して分布形を求める方法で, ノンパラメトリックモデルと呼ばれる. ノンパラメトリックモデルの例としては, カーネル密度推定法 [1], k-最近傍密度推定法 [2] などが挙げられる. また, 3 つ目はこれらの中間的な方法で, セミパラメトリックと呼ばれるものである. この方法は, 分布を表現するために必要なパラメータ数をコントロールできるようにした方法であり, 混合分布モデル (mixture model) [3] はその一例である. 本稿前半では, 統計的機械

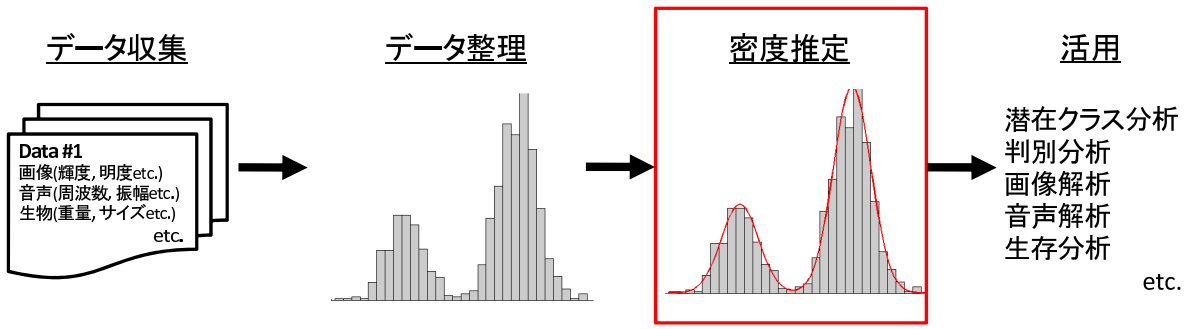


図 1.1: 密度推定

学習における密度推定問題の理論的研究として, EM アルゴリズム, 変分ベイズ法による混合正規分布推定の局所最適性の問題を取り上げる.

また, 本稿の後半では, 統計的機械学習の応用例としてバイオインフォマティクス, 特にマイクロアレイデータ解析を取り扱う. 近年のマイクロアレイ技術の発達により, 膨大な数の遺伝子データ (マイクロアレイデータ) が得られるようになった. マイクロアレイデータの特徴として, 得られるデータ数が少なく, データの次元数 (遺伝子数など) がデータ数に対して非常に大きいことが挙げられる. このようなデータから有用な知見を得るためには, 伝統的な統計的手法では不十分になりつつある. そこで, パターン認識などで用いられる統計的機械学習の方法によるマイクロアレイデータ解析が活発に行われている. 本稿では, いくつかのマイクロアレイデータ解析タスクの中でも, 遺伝子群解析と array CGH データによるゲノム異常領域同定問題を取り扱う.

本章では, 1.1 節で密度推定問題の概略, 1.2 節でマイクロアレイデータ解析についての概略を見る.

1.1 混合正規分布推定

観測されたデータからその背後の確率的構造, すなわち確率密度分布を推定することは, データの特徴を捉え, 得られた知見を活用する上で重要である. 本稿では, 密度推定問題の中でも, 特に混合正規分布推定問題に着目する. 混合正規分布モデル (Gaussian mixture model: GMM) [3] とは, いくつかの正規分布の重み付け和として表現されるモデルであり, 様々な分布形を表現することができ, 広く用いられている. 観測されたデータに基づき, 確率密度分布を推定する方法としては最尤法 [4] やベイズ推定 [5] が挙げられる. しかしながら, 一般にデータが高次元であることや混合分布推定では各データの所属するクラスが未知, すなわち, 各データが, 混合分布を構成する複数の正規分布のどの混合要素から生成されたものであるかが未知であるため, これらの手法では解析解が得られない, 計算が非常に困難となるなどの問題が生じる. このように, 非観測データ (潜在変数) を含むデータから密度推定を行う代表的な手法として EM アルゴリズム (expectation-maximization algorithm) [6, 7], 変分ベイズ法 (variational Bayes method: VB 法) [8, 9], マルコフ連鎖モンテカルロ法 (Markov chain Monte Carlo: MCMC) [10] などが挙げられる. 本稿では, これらの中でも, 最も広く用いられていると思われる EM アルゴリズム, また近年活発に研究が行われている変分ベイズ法について考察する. EM アルゴリズ

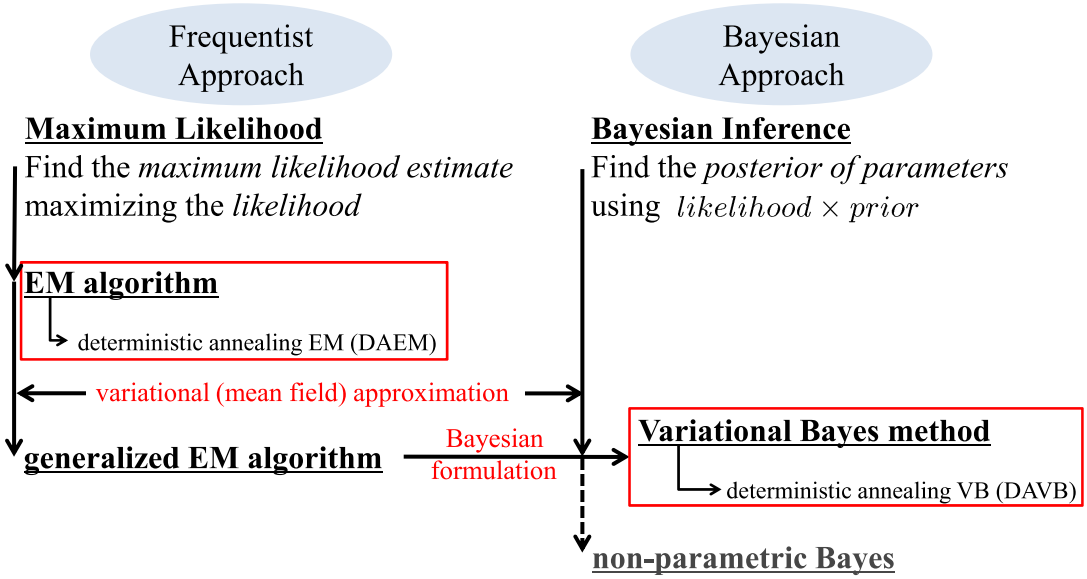


図 1.2: EM アルゴリズム, 変分ベイズ法の背景

ム, 変分ベイズ法はともに勾配法による反復計算によってパラメータを求める手法であるが, それらの背景は大きく異なる (図 1.2).

EM アルゴリズムは頻度論的なアプローチであり, 混合正規分布のモデルパラメータ (平均ベクトル, 分散共分散行列, 混合比) の点推定を行う手法である. その基礎は最尤法であり, 観測されたデータに対して尤もらしいパラメータを求める (尤度を最大化する) 方法である. 一方, 変分ベイズ法はベイズ推定における計算の複雑さを近似した方法である. ベイズ推定の枠組みでは, パラメータの点推定の行うのではなく, パラメータを確率的に変動する値, すなわち確率変数として捉え, パラメータの分布のパラメータ, すなわちハイパーパラメータを推定する. ベイズ推定では, 事前にパラメータに確率分布 (事前分布) を与え, データが得られた後, その分布がどのように変化するか (事後分布) を, ベイズの定理に基づいて推定する.

本稿では, EM アルゴリズム, 変分ベイズ法のどちらがより良いかという議論は避け, 両者が持つ局所最適性の問題に取り組む. 前述した通り, いずれの手法も勾配法による反復計算によりパラメータ (変分ベイズ法ではハイパーパラメータ) を推定する方法であり, 局所解へ収束する問題, 局所最適性の問題を有する. この問題を改善するために, 多くのアルゴリズムが提案されている [3, 7, 11, 12, 13, 14, 15, 16] が, 本稿では統計力学のアナロジーを用いた, 確定的アニーリング (deterministic

annealing: DA) 法 [11, 13] による方法に注目する. DA 法では, EM アルゴリズムや変分ベイズ法の枠組みと統計力学の枠組みを対応付けることで目的関数に温度パラメータを導入し, 単純な形状の目的関数から徐々に元の目的関数に変化させる各段階においてパラメータを逐次最適化する. その背景には, 高温状態での目的関数は, 元の目的関数の大域的構造を近似しているというアイディアがあり, 高温状態で得られたパラメータ付近に, 元の目的関数における良いパラメータが存在するであろうという考えがある. DA 法を用いた手法, 確定的アニーリング EM(deterministic annealing EM: DAEM) アルゴリズム, 確定的アニーリング変分ベイズ (deteministic annealing VB: DAVB) 法を混合正規分布推定に適用することで, 局所最適性が改善することが可能である [11, 13]. 本稿では, DA 法のアイディアを利用した多方向探索アプローチを提案する. 提案法では, DA 法における高温での解を原始初期点と定義し, この点が持ついくつかの特徴を利用することで, 原始初期点を起点とした探索方向を解析的に生成する. 探索方向は一般に 1 つとは限らず複数の探索方向が存在しうるが, 本手法ではこれらの方向を原始初期点における目的関数の Hesse 行列の固有値に基づいて自動的に生成することが可能であり, 計算機実験により多方向探索法が良い性能を示すことを示す.

1.2 統計的機械学習を用いたマイクロアレイデータ解析

本稿の後半では, 統計的機械学習の応用例としてマイクロアレイデータ解析を行う. 近年, マイクロアレイ技術の発達により, 非常に多くの遺伝子データを得ることが可能となった [17]. これにより, 生物学, 特に分子生物学に関連する様々な分野の研究が, 情報工学の技術を用いてなされるようになった (図 1.3 [18]). それらの分野を総称してバイオインフォマティクス (Bioinformatics) と呼び, 活発に研究が行われている.

本稿では特に, マイクロアレイデータ解析を行う. マイクロアレイデータは, 遺伝子数, すなわちデータの次元数 p が非常に大きく, 一方データ数 n が比較的小さい ($n << p$) ことが特徴である. このようなデータから有用な情報を得るには, 伝統的な統計的手法, 例えば t-検定などの簡単な検定, 回帰分析, 分割表を用いた解析, 分散分析など, では不十分となりつつある. そのような中で, 統計的機械学習の手法を用いたマイクロアレイデータ解析が注目を集め, 研究が行われている [18]. 変数選択法, 距離学習などによる重要遺伝子の同定 [19, 20, 21, 22] や, 混合正規分布モデルを用いたマイクロアレイの画像データ解析やクラスタリング [23, 24, 25], サ

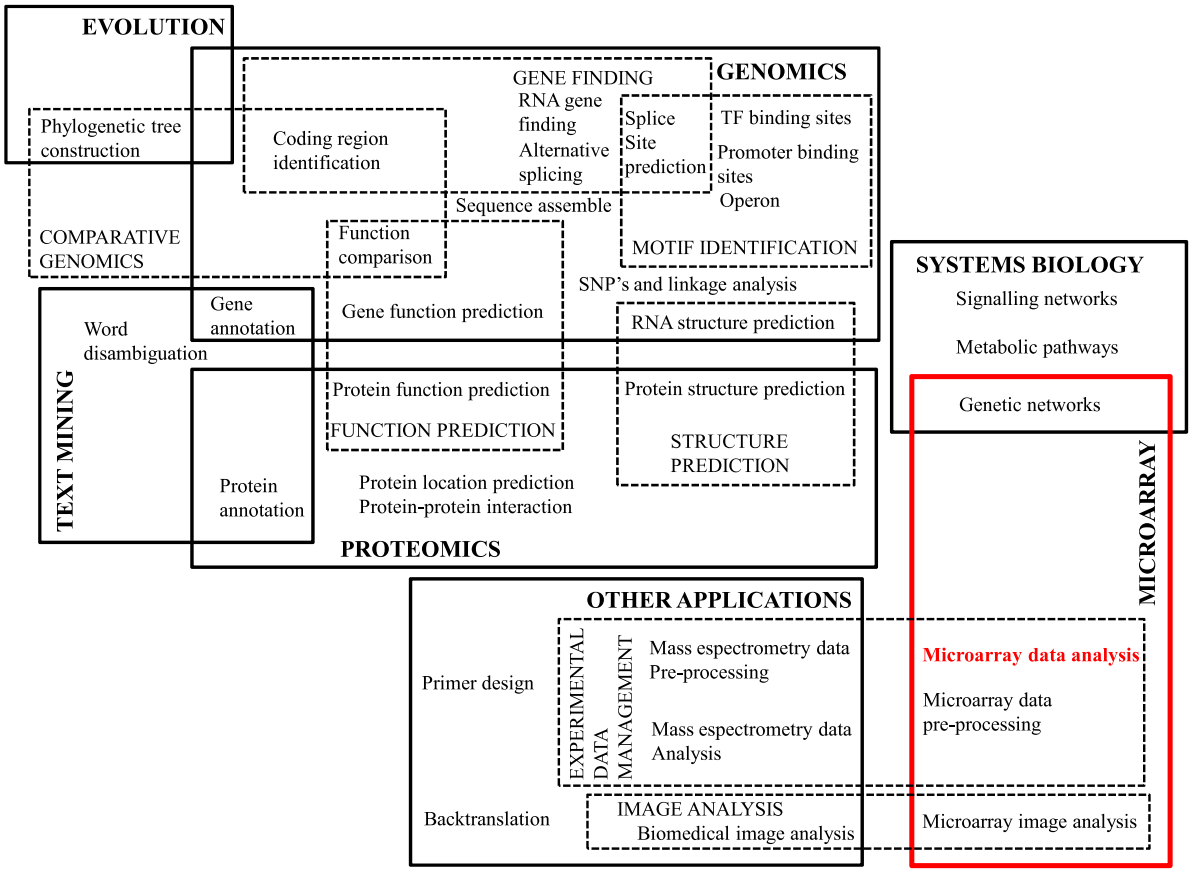


図 1.3: バイオインフォマティクスの研究領域 [18]:本研究で扱う分野を赤枠および赤字で示す.

ポートベクトルマシンなどの分類器による疾患判別 [26, 27] など, 多くのタスクが様々な手法を用いて行われている.

マイクロアレイデータ解析の重要なタスクの 1 つに, 異なる疾患において遺伝子発現量の異なる遺伝子を同定する問題がある. この問題に対しては, しばしば変数選択法や階層型クラスタリングなどの手法が用いられる. 一方で, 遺伝子はしばしば複数の遺伝子が互いに影響を及ぼしながら発現する. このような場合, 個々の遺伝子を個別に調べるのではなく, 関連性のある遺伝子をグループ化したもの (遺伝子群と呼ばれる) により多次元的に解析を行うことが重要である. 遺伝子発現量データを遺伝子群という単位で解析を行うタスクを遺伝子群解析 (gene set analysis: GSA) と呼ぶ. 遺伝子群解析を行う代表的な手法として, GSEA(gene set enrichment analysis) [28] や SAFE(significance analysis of function and expression) [29] がある (SAFE は GSEA よりも一般的な解析の枠組みを与えると解釈できる). しか

しながら, これらは個々の遺伝子に関する単変量統計量を組み合わせた手法であり, 遺伝子群に含まれている複数遺伝子に対して多次元的に解析を行っているわけではない. また, GSEA に批判的な文献 [30] も見られる. そこで, 本稿では, サポートベクトルマシン (support vector machine: SVM) [31] を用いた遺伝子群解析の枠組みを導入する. 遺伝子群解析は, 多変量 2 標本検定として定式化されるが, SVM の分類誤差を検定統計量として用いることで, 遺伝子群データを多次元データとして取り扱うことができ, また, 異なるサイズの遺伝子群との統計量の比較などを行うことが可能となる.

マイクロアレイデータには, array comparative genomic hybridization (array CGH) データと呼ばれる, 遺伝子発現量データとは異なるデータがある [32]. array CGH データは, DNA コピー数を定量化したデータであり, DNA コピー数異常に起因する疾患, 例えば癌など, の解析を行う際に有用である. 本稿では, array CGH データにおける, 2 標本で特徴の異なる異常領域を同定する問題を取り扱う. これにより, 癌のサブタイプの発見や, 同定された異常領域を用いた疾患判別などを行うことが可能となる. CGH マイクロアレイは, 遺伝子発現量データとは異なり, DNA 全体をプローブ (遺伝子断片) により表現したものであり, 隣り合うプローブは系列的なつながりを持っている. したがって, array CGH 解析においては, 隣接するプローブの相関を考慮した解析法が望まれる. このような系列データにおける異常領域を同定する手法として, ADM(aberrant detection method) [33], 混合正規分布を用いた隠れマルコフモデル (hidden markov model: HMM) による方法 [34, 35], コピー数が一定とみなせる領域に区分する方法 [36, 37] などが提案されている. このような手法は, 1 種の疾患に特異的な異常領域の同定は可能であるが, 2 種の疾患で特徴の異なる領域を同定することはできない. 本稿では, 最近傍多変量検定を用いたゲノム異常領域同定を行う. ゲノム異常領域同定問題も, 遺伝子群解析と同様, 多変量 2 標本検定として定式化できる. 最近傍分類器の分類誤差を検定統計量として用いることで, 2 種の疾患の多次元的な特徴の違いを定量化することができ, また, ゲノム異常領域同定におけるいくつかの重要な問題に対処することが可能となる.

1.3 本論文の構成

本稿の構成は次の通りである. 前半の第 2 章, 第 3 章, 第 4 章では密度推定問題を取り扱う. まず, 第 2 章において, 本稿で取り扱う混合正規分布についての説明を行う. 第 3 章では, 混合正規分布推定問題に対する頻度論的アプローチである EM

アルゴリズムとその拡張について説明する. ここでは, EMアルゴリズムの局所最適性問題の改善法の1つであるDAEMアルゴリズムについて説明を行い, DAEMアルゴリズムのアイディアを活用した別のアプローチである多方向探索EMアルゴリズムを提案する. 第4章では, ベイズ的アプローチである変分ベイズ法による混合正規分布推定を見る. DAEMアルゴリズムと同様の枠組みをベイズ推定へ適用したDAVB法について説明を行い, EMアルゴリズムで定式化を行った多方向探索アルゴリズムを, ベイズ推定の枠組みへも適用する.

後半の第5章, 第6章, 第7章では, 統計的機械学習の応用例としてマイクロアレイデータ解析を取り扱う. 第5章において, マイクロアレイデータ解析の概要と2種のタスク, 遺伝子群解析とarray CGHデータを用いたゲノム異常領域同定問題について触れる. 第6章では, サポートベクトルマシンの分類誤差を検定統計量として用いた遺伝子群解析の枠組みを導入する. ここでは, 最小全域木(minimum spanning tree: MST)とパス追跡を用いたSVM学習により, 効率的な多重多変量2標本検定を行うことができることを見る. 第7章では, array CGHデータのゲノム異常領域同定問題を取り扱う. ここでは, 最近傍多変量検定により, ゲノム異常領域同定におけるいくつかの問題点に対処し, 効率的な領域同定を行うことができることを見る. 第8章で本稿をまとめる.

第 2 章

混合正規分布

与えられたデータからそのデータを生成した分布を推定することは, そのデータの特徴を解析する上で重要である. 混合分布は単一の分布に比べ表現力も強く, 確率ニューラルネット, パターン認識, バイオインフォマティクスなど応用範囲も幅広い. 本章では, 特に応用上重要な正規分布およびその重ね合わせである混合正規分布について説明を行う.

2.1　正規分布

データ $\boldsymbol{x} \in \mathbb{R}^p$ に対する, p 変量正規分布 (multivariate normal distribution, multivariate Gaussian distribution) は以下のような式で表される:

$$p(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right\}.$$

ただし, $\mathbb{R}^p$ は p 次元の実数空間を表し, $\boldsymbol{m} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ はそれぞれ平均ベクトル, 分散共分散行列である. また, $\boldsymbol{\Sigma}^{-1}$ は精度行列と呼ばれる. 正規分布は確率密度関数であるので, 以下の確率の要件を満たす:

$$\int_{-\infty}^{\infty} p(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{x} = 1,$$
$$0 \leq p(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma}) \leq 1.$$

なお, 確率変数 X が平均ベクトル $\boldsymbol{m}$, 分散共分散行列 $\boldsymbol{\Sigma}$ の正規分布に従うとき, しばしば,

$$X \sim \mathcal{N}(X|\boldsymbol{m}, \boldsymbol{\Sigma}),$$

と表記される. 図 2.1 に平均 0, 分散 1 の単変量正規分布および平均ベクトル $\boldsymbol{m} = [0, 0]^\top$, 分散共分散行列 $\boldsymbol{\Sigma} = I$(ただし, I は単位行列) の 2 変量正規分布の例を

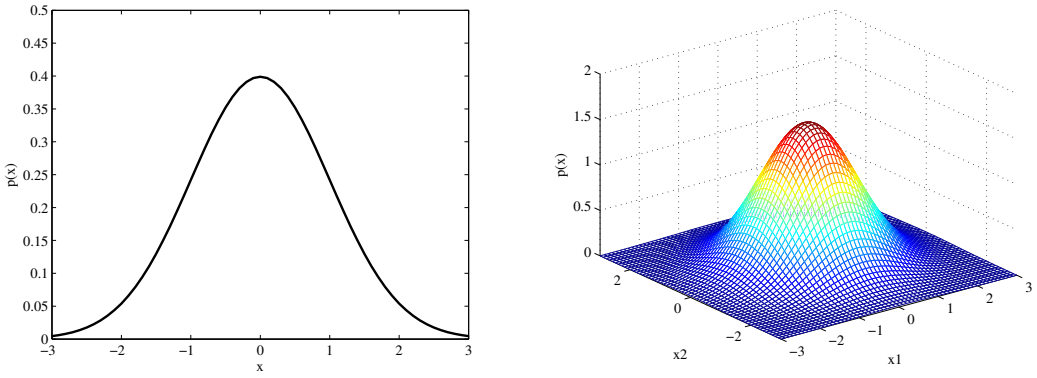


図 2.1: 正規分布の例: 単変量正規分布 (左) と 2 変量正規分布 (右)

示す. 図のように, 正規分布は平均ベクトルを中心とした “釣鐘状” の形状をしている.

n 個のデータ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$[1] , $\boldsymbol{x}_i \in \mathbb{R}^p$ が観測されたとき, それらのデータが生成される同時確率は以下で与えられる:

$$\begin{aligned} p\big(\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}|\boldsymbol{m},\boldsymbol{\Sigma}\big) &= \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{m},\boldsymbol{\Sigma}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{m})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{m})\right\}. \quad (2.1) \end{aligned}$$

2.2 混合正規分布

次に, 正規分布の重ね合わせとして表される混合正規分布 (Gaussian mixture distribution) [3] について説明する. 混合正規分布は, 正規分布に比べ様々な形状の分布を表現することができ, 多くのアプリケーション分野で用いられている.

$\boldsymbol{x} \in \mathbb{R}^p$ に対する K 混合 p 変量正規分布は以下で与えられる:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k,\boldsymbol{\Sigma}_k) \quad (2.2)$$

正規分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k,\boldsymbol{\Sigma}_k)$ は混合要素と呼ばれ, 平均ベクトル $\boldsymbol{m}_k \in \mathbb{R}^p$, 分散共分散行列 $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ を持つ. また, π_k は混合要素 k の混合比である. $p(\boldsymbol{x})$ は確率密度関数

[1]本稿では, ある自然数 n に対して, $i = 1,\ldots,n$ の略記法として, $\mathbb{N}_n$ を用いる. ただし, $\mathbb{N}$ は自然数の集合.

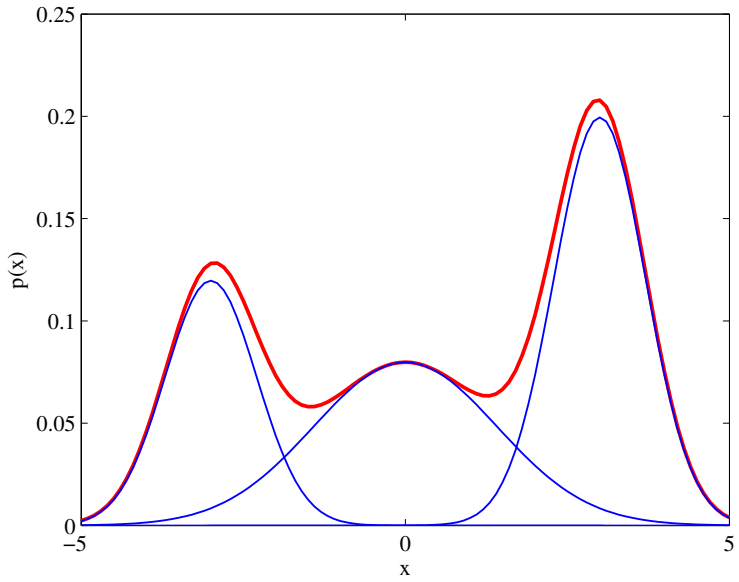


図 2.2: 3 混合単変量正規分布の例: 赤線は混合分布, 青線は混合要素

であるため, $\boldsymbol{x}$ に関する積分が 1 とならなければならない. すなわち,

$$\int_{-\infty}^{\infty} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$$

ここから, π_k に関して以下のような制約が導かれる:

$$\begin{aligned}\int_{-\infty}^{\infty} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{k=1}^{K} \pi_k \int_{-\infty}^{\infty} \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)\mathrm{d}\boldsymbol{x} \\ &= \sum_{k=1}^{K} \pi_k \\ &= 1. \end{aligned} \tag{2.3}$$

ここで,

$$\int_{-\infty}^{\infty} \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)\mathrm{d}\boldsymbol{x} = 1,$$

を用いた. 更に, $p(\boldsymbol{x}) \geq 0$ と $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k) \geq 0$ であることから,

$$0 \leq \pi_k \leq 1 \quad k = 1, \ldots, K, \tag{2.4}$$

も導かれる. 式 (2.3), (2.4) より, 混合比 π_k は確率の要件を満たすことがわかる.

確率の和および積の法則により, $p(\boldsymbol{x})$ は

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}|k), \tag{2.5}$$

と表される. 式 (2.5) は式 (2.2) と等価であるため, $\pi_k = p(k)$ と π_k を混合要素 k の事前確率 (prior probability) とみなすことも可能である. 更に, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k) = p(\boldsymbol{x}|k)$ は, k によって条件付けされた混合要素 k の $\boldsymbol{x}$ に関する確率密度と解釈できる. これらの事実は次章以降で説明する EM アルゴリズム, 変分ベイズ法において重要である.

混合正規分布はパラメータ $\{\pi_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k\}_{k\in\mathbb{N}_K}$ によって完全に特徴付けられるため, 以下のようにパラメトリックに表現することが可能である:

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k). \tag{2.6}$$

ここで, $\boldsymbol{\pi} = \{\pi_k\}_{k\in\mathbb{N}_K}$, $\boldsymbol{m} = \{\boldsymbol{m}_k\}_{k\in\mathbb{N}_K}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}_{k\in\mathbb{N}_K}$ とした. また, データ $X = \{\boldsymbol{x}_i\}_{i=1}^{n}$ が観測されたときの混合正規分布の対数尤度関数は,

$$\ln p(X|\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k) \right\},$$

で与えられる.

図 2.2 に以下で与えられる 3 混合 1 変量正規分布の例を示す:

$$p\big(\boldsymbol{x}|\{\boldsymbol{\pi}_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k\}_{k\in\mathbb{N}_3}\big) = 0.3\mathcal{N}(\boldsymbol{x}|-3, 2) + 0.2\mathcal{N}(\boldsymbol{x}|0, 0.5) + 0.5\mathcal{N}(\boldsymbol{x}|3, 2).$$

第 3 章

EMアルゴリズムと混合正規分布推定

本章では, EMアルゴリズムとEMアルゴリズムを拡張した手法を用いた混合正規分布推定問題を取り扱う. 正規分布および混合正規分布推定問題は古くから研究されており, 今なお研究が進んでいる. 正規分布のパラメータは最尤法により推定可能であるが, 混合正規分布については, 一般に各観測データが所属する混合要素が未知であるため, パラメータの推定が困難となる. 混合正規分布のパラメータはEMアルゴリズム (Expectation-Maximization algorithm) と呼ばれる繰り返しアルゴリズムによって, 逐次的に尤度関数を最大化することで推定できる. しかし, 勾配型のアルゴリズムであるEMアルゴリズムは, その目的関数が多峰であるが故に局所最適性の問題を有する. 本章では, この問題を改善するため, 原始初期点 (primitive initial point: PIP) を用いた多方向探索型のアプローチ [38] を提案し, 計算機実験によりその有効性を検証する.

3.1 最尤法

本節では, 密度推定問題に対する基本的な解法である最尤法 (maximum likelihood: ML) について説明を行う. 最尤法では, 得られたデータを生成する最も尤もらしい分布のパラメータ (最尤推定値, maximum likelihood estimate: MLE) を, 尤度関数を最大化することで求める手法である. 最尤法について説明を行った後, 最尤法による正規分布推定のシミュレーションを行う. そこでは, データが正規分布に属さない, 混合正規分布から生成されたデータに対しては, 最尤法によるパラメータ推定では不適切であることをみる.

3.1.1 最尤法

いま, n 個の観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ が得られたとすると, 観測データの同時確率密度は以下の式で表すことができる:

$$p(\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}|\Theta) \;=\; \prod_{i=1}^{n} p(\boldsymbol{x}_i|\Theta). \tag{3.1}$$

式 (3.1) はパラメータ Θ が与えられたもとで観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ が得られたときの同時確率密度を考えている. ここで, 逆に観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$ が固定されており, パラメータ Θ が未知であると考えてみる. すると, 式 (3.1) はパラメータ Θ の関数と見なすことができ, これを尤度関数 (likelihood-function) と呼ぶ:

$$L(\Theta) \;=\; \prod_{i=1}^{n} p(\boldsymbol{x}_i|\Theta).$$

最尤法ではこの尤度関数 $L(\Theta)$ を最大にするようなパラメータ Θ をその推定値とする. つまり, 可能なパラメータのうち, 観測データの同時確率を最大にするものを最尤推定値とする.

多くの確率モデルでは尤度関数 $L(\Theta)$ を直接最大化するよりも, 以下のような対数尤度関数 $\mathcal{L}$ を最大化するほうが計算が簡単になる:

$$\begin{aligned}\mathcal{L}(\Theta) \;&=\; \ln \prod_{i}^{n} p(\boldsymbol{x}_i|\Theta) \\ &=\; \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\Theta).\end{aligned}$$

対数尤度関数 $\mathcal{L}$ を最大にする Θ は以下の式を解くことによって求まる:

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \Theta} \;=\; 0. \tag{3.2}$$

3.1.2 最尤法による正規分布推定

本節では, 前節で述べた最尤法により, $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ が与えられたときの多変量正規分布のパラメータ $\Theta = \{\boldsymbol{\mu},\ \boldsymbol{\Sigma}\}$ を推定する. ここで, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ はそれぞれ平均ベクトル, 分散共分散行列である.

多変量正規分布の対数尤度関数は式 (2.1) の対数を取ることにより, 以下のように表される:

$$\begin{aligned}\ln p(\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \;&=\; \ln \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right\} \\ &=\; \sum_{i=1}^{n} \ln \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right\}\end{aligned} \tag{3.3}$$

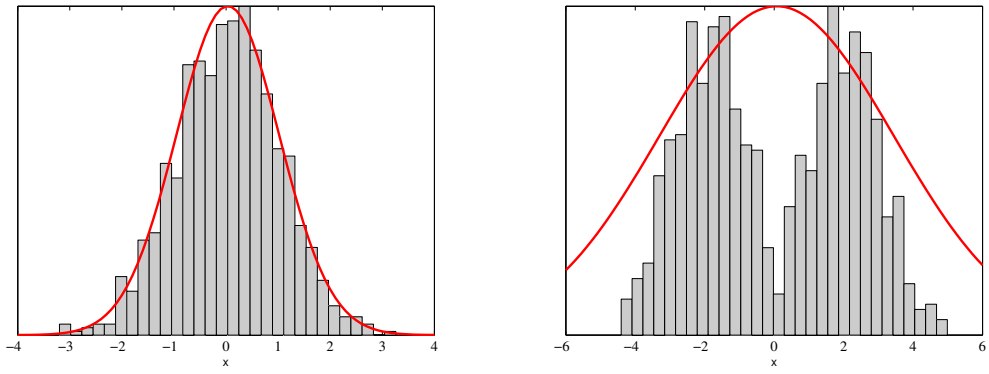


図 3.1: 最尤法の例: 正規分布 (左) と混合正規分布 (右)

式 (3.3) をパラメータ $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ に関して偏微分をして 0 とおくことで, 次のようにパラメータの最尤推定値 $\boldsymbol{\mu}_{\mathrm{ML}}, \boldsymbol{\Sigma}_{\mathrm{ML}}$ が得られる:

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \tag{3.4}$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{ML}})^{\top}. \tag{3.5}$$

式 (3.4), (3.5) を見てもわかるように, 正規分布の最尤推定値はそれぞれ標本平均, 標本分散となっていることが見て取れる.

最後に, 人工データを用いて, 最尤法による正規分布推定のシミュレーションを行う. 用いる人工データは $X \sim \mathcal{N}(0,1)$ および $X \sim 0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$ からそれぞれ 1000 個ずつ生成した 2 通りのものを用いる.

結果を図 3.1 に示す. 図中のヒストグラムは生成されたデータを表し, 赤線は最尤法により推定されたパラメータを持つ正規分布を表している. 前者のデータは単一の正規分布から生成されたものであり, 正規分布を仮定した最尤法により上手く推定が行えている. 一方で, 後者のデータは混合正規分布から生成されたデータであり, 最尤法での推定結果はデータにフィットしていないことがわかる.

これらの結果から, データの分布が複雑な場合 (今回は混合正規分布とした), 単純な最尤法ではデータにフィットしたパラメータが推定できないことがわかる. このような場合, 次節で説明する EM アルゴリズムを用いることで, パラメータを推定することが可能である.

3.2 EMアルゴリズム

本節では, 最尤推定値を求める反復解法であるEMアルゴリズムについて説明する.

3.2.1 EMアルゴリズム

最尤法において, 式(3.2)の解が解析的に求まればよいが, 非線形となり解析解が得られない場合がある. このような問題に対する数値的な解法としてEMアルゴリズムがある. EMアルゴリズムでは, 観測データを不完全であるとみなし, 問題を解決できるレベルまで完全化し, そのフレームワークで尤度最大化を逐次的に行う.

いま, n 個のデータ点 $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{z}_i \in \mathcal{Z}$ が確率密度関数 $p(X, Z|\Theta)$ によって生成されたとする. ただし, $\mathcal{Z}$ は潜在変数の空間, X, Z は確率変数を表す. $\boldsymbol{x}_i \in \mathbb{R}^p$ のみが観測され, $\boldsymbol{z}_i \in \mathcal{Z}$ は非観測データ (潜在変数) とする. ここで, Θ はパラメータ集合を表し, データ点 $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$ を生成する確率密度関数を $p(X|\Theta)$ とする. EMアルゴリズムでは, $\{(\boldsymbol{x}_i,\ \boldsymbol{z}_i)\}_{i\in\mathbb{N}_n}$ を**完全データ**と呼び, $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$ を**不完全データ**と呼ぶ. このように観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$ だけでは, 最尤推定が困難な問題に対し, 非観測データ (ここでは潜在変数 $\boldsymbol{z}_i$) が含まれていると考え, データを完全化する.

このような完全データのフレームワークを用いて, 不完全データの対数尤度は以下のように表すことができる:

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\Theta) \\ &= \sum_{i=1}^{n} \ln \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{x}_i, \boldsymbol{z}|\Theta). \end{aligned} \tag{3.6}$$

また, 完全データの対数尤度も以下のように定義できる:

$$\mathcal{L}_{\mathrm{cmp}}(\Theta) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i, \boldsymbol{z}_i|\Theta). \tag{3.7}$$

EMアルゴリズムでは, 不完全データの対数尤度を直接最大化するのではなく, 反復により対数尤度 $\mathcal{L}(\Theta)$ を逐次増加させる. つまり, すでに求められたパラメータの値 $\Theta^{(t)}$ から, 対数尤度を増加させるような新しいパラメータ $\Theta^{(t+1)}$ を推定する操作を繰り返すことにより, 対数尤度を最大化するパラメータを求める.

ここで, 観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$ が得られたもとで, パラメータを $\Theta^{(t)}$ から $\Theta^{(t+1)}$ に

更新したときの対数尤度の差を求めてみる:

$$\begin{aligned}
\mathcal{L}(\Theta^{(t+1)}) - \mathcal{L}(\Theta^{(t)}) &= \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\Theta^{(t+1)}) - \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\Theta^{(t)}) \\
&= \sum_{i=1}^{n} \ln \frac{p(\boldsymbol{x}_i|\Theta^{(t+1)})}{p(\boldsymbol{x}_i|\Theta^{(t)})} \\
&= \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln \frac{p(\boldsymbol{x}_i|\Theta^{(t+1)})}{p(\boldsymbol{x}_i|\Theta^{(t)})} \\
&= \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln \frac{p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta^{(t+1)})}{P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t+1)})} \frac{P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})}{p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta^{(t)})} \\
&= \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln \frac{p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta^{(t+1)})}{p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta^{(t)})} \\
&\quad + \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln \frac{P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})}{P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t+1)})}. \qquad (3.8)
\end{aligned}$$

ここで, $P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})$ は以下のような事後確率である:

$$P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) = \frac{p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta^{(t)})}{\sum_{\boldsymbol{z}'\in\mathcal{Z}} p(\boldsymbol{x}_i,\boldsymbol{z}'|\Theta^{(t)})}. \qquad (3.9)$$

また, Q 関数, H 関数を,

$$\begin{aligned}
Q(\Theta^{(t)},\Theta) &= \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta), \qquad (3.10) \\
H(\Theta^{(t)},\Theta) &= \sum_{i=1}^{n} \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)}) \ln P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta),
\end{aligned}$$

と定義すると, 式 (3.8) は以下のようになる:

$$\begin{aligned}
\mathcal{L}(\Theta^{(t+1)}) - \mathcal{L}(\Theta^{(t)}) &= Q(\Theta^{(t)},\Theta^{(t+1)}) - Q(\Theta^{(t)},\Theta^{(t)}) \\
&\quad + H(\Theta^{(t)},\Theta^{(t)}) - H(\Theta^{(t)},\Theta^{(t+1)}).
\end{aligned}$$

ここで, ジェンセンの不等式より次式が成り立つ:

$$H(\Theta^{(t)},\Theta^{(t+1)}) - H(\Theta^{(t)},\Theta^{(t)}) \leq 0.$$

従って, パラメータの更新に関する以下のような不等式が導出できる:

$$\mathcal{L}(\Theta^{(t+1)}) - \mathcal{L}(\Theta^{(t)}) \geq Q(\Theta^{(t)},\Theta^{(t+1)}) - Q(\Theta^{(t)},\Theta^{(t)}). \qquad (3.11)$$

式(3.11)より, 以下のような条件を満たす$\Theta^{(t+1)}$を得ることにより対数尤度を増加させることができることがわかる:

$$Q(\Theta^{(t)}, \Theta^{(t+1)}) \geq Q(\Theta^{(t)}, \Theta^{(t)}).$$

導出された式(3.10)で表されるQ関数は現在のパラメータ$\Theta^{(t)}$が与えられた事後確率$P(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$のもとでの完全データの対数尤度$\mathcal{L}_{\mathrm{cmp}}$の条件付き期待値である. EMアルゴリズムは$Q$関数を最大化することで対数尤度を増加させる. これを収束するまで反復することで間接的に最尤推定を実現する.

具体的には$t+1$回目の反復において, まず, Eステップ(expectation-step)で$Q(\Theta^{(t)}, \Theta)$を計算し, Mステップ(maximization-step)でその$Q(\Theta^{(t)}, \Theta)$を最大にするΘを求め$\Theta^{(t+1)}$とする. これを収束するまで繰り返す.

3.2.2 EMアルゴリズムの流れ

EMアルゴリズムの簡単な流れを以下に示す.

[EMアルゴリズム]

Step 1. 初期値$\Theta^{(0)}$を設定し, $t \leftarrow 0$とする.

Step 2. 以下のステップを収束するまで繰り返す.

E-step: 事後確率$P(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$を用いて$Q(\Theta^{(t)}, \Theta)$を計算する.

M-step: $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta^{(t)}, \Theta)$とし, $t \leftarrow t+1$とする.

上記のEMアルゴリズムに関し, 次の定理が証明されている.

定理 1 *EMアルゴリズムの各反復において, $\mathcal{L}_{\mathrm{cmp}}(\Theta)$は単調増加, すなわち, $\mathcal{L}_{\mathrm{cmp}}(\Theta^{(t+1)}) \geq \mathcal{L}_{\mathrm{cmp}}(\Theta^{(t)})$が成り立ち, その等号は$Q(\Theta^{(t+1)}, \Theta^{(t)}) = Q(\Theta^{(t)}, \Theta^{(t)})$のときかつそのときに限る.*

つまり, ある初期値$\Theta^{(0)}$からEMステップを実行することにより, $\mathcal{L}(\Theta)$の極大値へ収束することを示している. ただし, t回目の反復で得られる推定値$\Theta^{(t+1)}$は事後確率$P(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$に強く依存する. 推定の初期段階では推定値$\Theta^{(t)}$は信頼性が低く, 事後確率$P(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$も信頼性が低い. つまり, EMアルゴリズムによって得られる解は, 結局, 初期値$\Theta^{(0)}$に依存することになる.

次節ではEMアルゴリズムにおけるこのような局所最適性の問題に対する一つの解決策として提案されたDAEMアルゴリズムについて説明する.

3.3 確定的アニーリングEMアルゴリズム

EMアルゴリズムは, 得られる解が$\Theta^{(0)}$, すなわち初期値に大きく依存する. ここでは, EMアルゴリズムの局所最適性問題の改善法の1つであるDAEMアルゴリズムについて説明する.

3.3.1 確定的アニーリングEMアルゴリズム

確定的アニーリングEM(deterministic annealing EM: DAEM)アルゴリズムは最大エントロピー原理と統計力学のアナロジーを利用する[11]. いま, n個の非観測データ$\{\boldsymbol{z}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{z}_i \in \mathcal{Z}$を状態とする系を考え, 各状態は$\{\boldsymbol{z}_i\}_{i\in\mathbb{N}_n}$に対応して以下のようなエネルギーを持つとする:

$$E(\{\boldsymbol{z}_i|\Theta\}_{i\in\mathbb{N}_n}) \;=\; -\mathcal{L}_{\mathrm{cmp}}(\Theta).$$

非観測データがより好ましい状態であればエネルギーもより小さくなることがわかる. このような系の平衡状態では, 各状態の存在確率は以下のボルツマン分布に従う:

$$\begin{aligned} P(\{\boldsymbol{z}_i\}_{i\in\mathbb{N}_n}|\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n},\Theta) &= \prod_{i=1}^{n}\frac{p(\boldsymbol{x}_i,\boldsymbol{z}_i|\Theta)^\beta}{\sum_{\boldsymbol{z}\in\mathcal{Z}}p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta)^\beta} \\ &= \prod_{i=1}^{n} q(\boldsymbol{z}_i|\boldsymbol{x}_i,\Theta). \end{aligned}$$

ここで, $\beta(>0)$は統計力学における温度パラメータであり, $1/\beta$が温度に相当する. $q(\boldsymbol{z}_i|\boldsymbol{x}_i,\Theta)$は以下のような温度パラメータ$\beta$付きの事後確率である:

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i,\Theta) \;=\; \frac{p(\boldsymbol{x}_i,\boldsymbol{z}_i|\Theta)^\beta}{\sum_{\boldsymbol{z}\in\mathcal{Z}}p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta)^\beta}.$$

これにより, 系における分配関数$Z(\Theta)$が定義され, 統計力学アナロジーから以下の自由エネルギー$F_\beta(\Theta)$が定義できる:

$$F_\beta(\Theta) \;=\; -\frac{1}{\beta}\ln Z(\Theta). \tag{3.12}$$

ここで, 分配関数$Z(\Theta)$は以下で与えられる:

$$Z(\Theta) \;=\; \prod_{i=1}^{n}\sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta)^\beta.$$

よって, 式(3.12)は次のようになる:

$$\begin{aligned} F_\beta(\Theta) &= -\frac{1}{\beta}\ln\prod_{i=1}^{n}\sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta)^\beta \\ &= -\frac{1}{\beta}\sum_{i=1}^{n}\ln\sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta)^\beta. \end{aligned}$$

系はこの自由エネルギー$F_\beta(\Theta)$を最小とする状態に落ち着く.

EMアルゴリズムと同様に, β付きの事後確率$q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})$のもとで, $\mathcal{L}_{\rm cmp}$の条件付き期待値$Q_\beta(\Theta^{(t)},\Theta)$を求めることでDAEMアルゴリズムの枠組みを構成できる:

$$-F_\beta(\Theta) = Q_\beta(\Theta^{(t)},\Theta) - \frac{1}{\beta}H_\beta(\Theta^{(t)},\Theta).$$

ただし,

$$\begin{aligned} Q_\beta(\Theta^{(t)},\Theta) &= \sum_{i=1}^{n}\sum_{\boldsymbol{z}\in\mathcal{Z}} q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})\ln p(\boldsymbol{x}_i,\boldsymbol{z}|\Theta), \\ H_\beta(\Theta^{(t)},\Theta) &= \sum_{i=1}^{n}\sum_{\boldsymbol{z}\in\mathcal{Z}} q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})\ln q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta). \end{aligned}$$

したがって, DAEMアルゴリズムではβ付きのQ関数である$Q_\beta(\Theta^{(t)},\Theta)$を逐次最大化することで$-F_\beta(\Theta)$を間接的に最大化する.

3.3.2 温度パラメータβの効果

推定の初期段階では推定値$\Theta^{(t)}$の信頼性が低く, 事後確率$P(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})$も信頼性が低くなる. そこで, DAEMアルゴリズムは, 推定の初期段階におけるβ付きの事後確率$q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})$の影響をできるだけ弱めるように, βの値を小さく($\beta\approx 0$)設定する. このとき, $q(\boldsymbol{z}|\boldsymbol{x}_i,\Theta^{(t)})$は推定値に依らず一様分布となる. このような高温の状態では, $F_\beta(\Theta)$の極小値は唯一となり, 如何なる初期値を与えても, EMステップを繰り返すことによりその最小値を求めることができる.

次に, βの値を少し大きくする(温度を下げる)ことを考える. このように, 温度を変化させるプロセスを**アニーリング**と呼ぶ. アニーリングにより, $F_\beta(\Theta)$の形状が緩やかに変化し, それにともなって異なる位置へ極値が移動する. このとき, この温度での最小値は, 先に得られた最小値の近傍に位置していると思われる. そこで, 先に得られた推定値を初期値としてEM推定を行えば, この温度での最小値を得ることができる. このように徐々に温度を下げながらEM推定を行うことで, 各温度での極小値が逐次追跡されることになる.

また, $\beta = 1$ となったとき, β 付きの事後確率 $q(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$ はEMアルゴリズムにおける事後確率 (式 (3.9)) と一致する:

$$q(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)}) = P(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)}), \quad \text{for } \beta = 1.$$

このとき, 自由エネルギー $F_\beta(\Theta)$ は対数尤度と一致する (符号は逆):

$$-F_1(\Theta) = \mathcal{L}(\Theta), \quad \text{for } \beta = 1.$$

つまり, $\beta = 1$ における $F_\beta(\Theta)$ の最小値を与えるパラメータは最尤推定値となる.

このことから, DAEMアルゴリズムのアニーリング制御では $\beta(> 0)$ を小さな値 (高温) から始め, 徐々に β を増加させ, 最終的には $\beta = 1$ とすればよいことがわかる.

3.3.3 DAEMアルゴリズムの流れ

DAEMアルゴリズムの流れを以下に示す.

[DAEMアルゴリズム]

Step 1. $\beta \leftarrow \beta_{\min}(\approx 0)$ とする.

Step 2. 初期値 $\Theta^{(0)}$ を設定し, $t \leftarrow 0$ とする.

Step 3. 以下のステップを収束するまで繰り返す.

E-step： 事後確率 $q(\boldsymbol{z}|\boldsymbol{x}_i, \Theta^{(t)})$ を用いて $Q_\beta(\Theta^{(t)}, \Theta)$ を計算する.

M-step：$\Theta^{(t+1)} = \arg\max_{\Theta} Q_\beta(\Theta^{(t)}, \Theta)$ とし, $t \leftarrow t + 1$ とする.

Step 4. $\beta < 1$ なら $\beta \leftarrow \min(\beta \times \text{const}, 1)$ とし, Step 3 へ. $\beta \geq 1$ であれば, 終了.

上記のDAEMアルゴリズムに関し, 次の定理が証明されている.

定理 2 *DAEMアルゴリズムの EMステップの各反復において, $F_\beta(\Theta)$ は単調減少, すなわち, $F_\beta(\Theta^{(t+1)}) \leq F_\beta(\Theta^{(t)}$ が成り立ち, その等号は $Q_\beta(\Theta^{(t+1)}, \Theta^{(t)}) = Q_\beta(\Theta^{(t)}, \Theta^{(t)})$ のときかつそのときに限る.*

3.4 混合正規分布推定

ここでは, EMアルゴリズム, DAEMアルゴリズムを用いた混合正規分布推定を定式化する. まず, EM, DAEMアルゴリズムの混合正規分布におけるパラメータ更新式を示し, その後, 各アルゴリズムの解空間における挙動, 局所最適性の問題について考察する.

3.4.1 EMアルゴリズムを用いた混合正規分布推定

まず, 混合正規分布におけるEMアルゴリズムのパラメータの更新式を導出する. K 混合 p 変量正規分布における推定すべきパラメータ集合 Θ は混合比 $\boldsymbol{\pi} = \{\pi_k\}_{k\in\mathbb{N}_K}$, 平均ベクトル $\boldsymbol{m} = \{\boldsymbol{m}_k\}_{k\in\mathbb{N}_K}$ および分散共分散行列 $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}_{k\in\mathbb{N}_K}$ であり,

$$\begin{aligned}\Theta &= \{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}_K},\\ \boldsymbol{\theta}_k &= \{\pi_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k\},\end{aligned}$$

で表す. いま, n 個の観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ が与えられたとし, それらを用いて K 混合 p 変量正規分布のパラメータ Θ を推定する. EMアルゴリズムへの適用にあたり, 潜在変数 $\{z_i\}_{i=1}^n$ を導入する. つまり, $\{\boldsymbol{x}_i, z_i\}_{i\in\mathbb{N}_n}$ が得られたと考える. 混合正規分布における潜在変数 z_i は観測データ $\boldsymbol{x}_i$ がどのクラスから生成されたかを表すスカラーであり, $\boldsymbol{x}_i$ が混合要素 k により生成された場合 $z_i = k$ となる. このとき, EMアルゴリズムにおける Q 関数は次のようになる:

$$\begin{aligned}Q(\Theta^{(t)}, \Theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) \ln p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k),\\ p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k) &= \pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k),\\ P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) &= \frac{\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)}.\end{aligned} \tag{3.13}$$

EMステップに従って Q 関数を最大にするパラメータを求める. ただし, 混合比 $\{\pi_k\}_{k\in\mathbb{N}_k}$ に関する制約条件,

$$\sum_{k=1}^{K} \pi_k = 1,$$

があるため, ラグランジェ乗数 λ を含む以下の関数を最大化するパラメータを求めることになる:

$$\mathcal{Q}(\Theta^{(t)}, \Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) \ln p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k) - \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]. \tag{3.14}$$

実際に式 (3.14) を最大化すべく $\partial\mathcal{Q}/\partial\Theta = 0$ を解くと, 各パラメータの更新式が以下

のように求まる:

$$\begin{aligned}
\pi_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}), \\
\boldsymbol{m}_k^{(t+1)} &= \frac{\sum_{i=1}^{n} \boldsymbol{x}_i P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}, \\
\boldsymbol{\Sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{m}_k^{(t+1)})^\top P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} P(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}.
\end{aligned}$$

これらの更新式に従いパラメータの更新, Q 関数の計算を収束するまで繰り返すことで, 混合正規分布のパラメータの最尤推定値を得る.

3.4.2 DAEMアルゴリズムを用いた混合正規分布推定

次に, DAEMアルゴリズムによる混合正規分布推定のパラメータ更新式を導出する. DAEMアルゴリズムではEMアルゴリズムにおける Q 関数のかわりに, Q_β 関数を最大にするパラメータを求める:

$$\begin{aligned}
Q_\beta(\Theta^{(t)}, \Theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) \ln p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k), \\
p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k) &= \pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k), \\
q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) &= \frac{\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)^\beta}{\sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Sigma}_k)^\beta}.
\end{aligned} \tag{3.15}$$

ただし, 前節と同様の制約条件があるため, ラグランジェ乗数 λ を含む以下の関数を最大化するパラメータを求めることになる:

$$\mathcal{Q}_\beta(\Theta^{(t)}, \Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}) \ln p(\boldsymbol{x}_i, k|\boldsymbol{\theta}_k) - \lambda\left[\sum_{k=1}^{K} \pi_k - 1\right].$$

EMアルゴリズムと同様, $\partial\mathcal{Q}/\partial\Theta = 0$ を解くと, 各パラメータの更新式が以下のように求まる:

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)}), \tag{3.16a}$$

$$\boldsymbol{m}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}, \tag{3.16b}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{m}_k^{(t+1)})^\top q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} q(k|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}. \tag{3.16c}$$

DAEMアルゴリズムへの移行は, EMアルゴリズムにおける式(3.13)の事後確率から式(3.15)のβ付き事後確率へ修正すればよいことがわかる.

3.4.3 EMアルゴリズムの局所最適性

ここでは, EMアルゴリズムの単変量混合正規分布推定における局所最適性の問題についてシミュレーションデータにより考察する. 実験に用いるデータは以下の4混合単変量正規分布から生成された人工データを用いる:

$$p\big(\boldsymbol{x}|\Theta\big) = 0.1\mathcal{N}(-6, 2) + 0.3\mathcal{N}(-2, 1.5) + 0.4\mathcal{N}(4, 2) + 0.2\mathcal{N}(7, 1) \tag{3.17}$$

図3.2に観測データのヒストグラムおよびいくつかの初期値によるEMアルゴリズムの推定結果を示す. 図において, 白のヒストグラムが観測データ, 赤線が真の分布(式(3.17)), その他の色の曲線がEMアルゴリズムによって推定された分布である.

図からわかるように, 初期値を変えると推定結果も大きく異なることが見て取れる. 青線の分布が比較的よくデータにフィットしている一方で, 緑線の分布は左2つの分布を無視したような形状をしており, 若干オーバーフィット気味である. このように, EMアルゴリズムでは初期値を慎重に設定しなければならないことがわかる. ただし, 今回の人工データはランダムに生成されたものであり, 真の分布と多少異なる配分でデータが生成されている. このような場合, 例え人工データであっても必ずしも真のパラメータが大域的最適解になるとは限らない点に注意されたい.

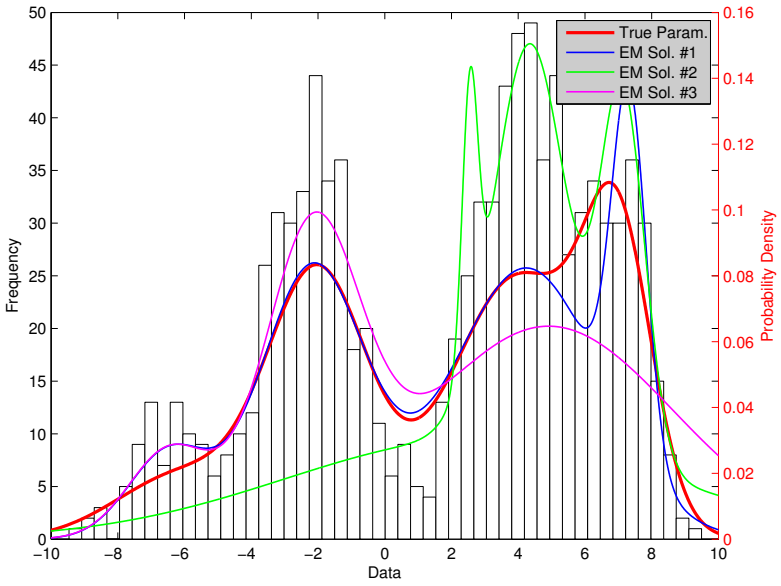


図 3.2: 局所最適性の例

3.4.4 解空間の形状と局所最適性

ここでは, 2 混合単変量正規分布のパラメータ推定問題を用いて, EM, DAEM アルゴリズムの尤度空間におけるパラメータの挙動を可視化する. 推定するパラメータは 2 つの平均のみとし, 混合比, 分散は既知であるとする:

$$p(x|\Theta) = \frac{0.3}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-m_1)^2\right\} + \frac{0.7}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-m_2)^2\right\} \tag{3.18}$$

$$\Theta = (m_1, m_2). \tag{3.19}$$

学習データとして 100 個のサンプルを人工的に生成した. データ生成に用いたパラメータは $(m_1, m_2) = (-2, 4)$ である.

この推定問題における対数尤度関数 $\mathcal{L}$ を図 3.3 に示す. 図より大域的最適解 $(m_1, m_2) = (-2, 4)$ と局所最適解 $(m_1, m_2) = (4, -2)$ が存在していることがわかる.

EM アルゴリズムにおける局所最適性の問題を明確にするために, 2 つの初期値を与え, 推定を行った. 図 3.4(左) は初期値 $(m_1^{(0)}, m_2^{(0)}) = (-4, -2)$ を与えた結果を示しており, 大域的最適解へ収束している. 図 3.4(右) は初期値 $(m_1^{(0)}, m_2^{(0)}) = (-2, -4)$ を与えた結果を示しており, 局所最適解へ収束している. 図より, EM アルゴリズムは初期値の与え方によって, 得られる解の性能が大きく左右されることが確認できる. 図 3.4(右) のように局所最適解付近の初期値を与えた場合, 大域的最適解を得

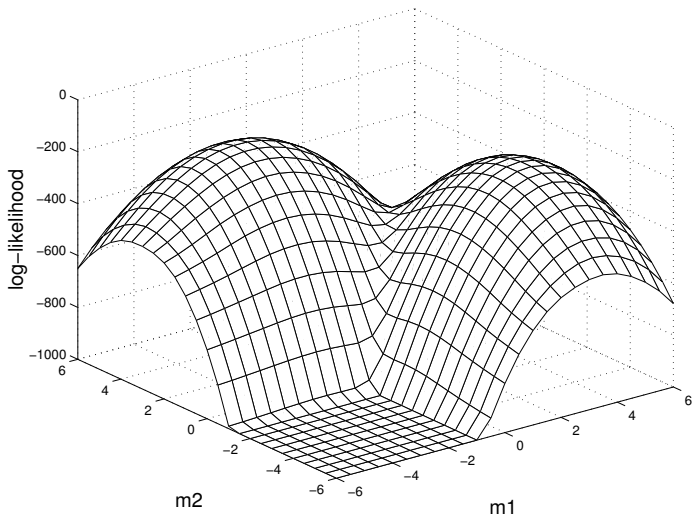


図 3.3: 対数尤度関数 $\mathcal{L}$

ることはできず, 初期値の周辺に存在する局所最適解へ収束する. つまり, EM アルゴリズムの解品質は初期値に大きく左右される. 良解を得るためには異なる初期値によって複数回推定を行わなければならない. いずれにせよ, EM アルゴリズムが局所最適性を有していることは明らかである.

次に EM アルゴリズムと同様, DAEM アルゴリズムの挙動を可視化する. 図 3.5 に初期値 $(m_1^{(0)}, m_2^{(0)}) = (-2, -4)$ を与えた DAEM アルゴリズムの推定結果を示す. これは, DAEM アルゴリズムがいかにして EM アルゴリズムにおける局所最適性の問題を解決しているかを明確にするためであり, EM アルゴリズムでは大域的最適解へ到達できない初期値を与えた. 8 回のアニーリングステップのうち (a)〜(d) はそれぞれ $\beta = 0.1, 0.14, 0.753, 1$ における探索点および尤度関数の遷移を示してある. また, DAEM アルゴリズムにおける解空間は, 自由エネルギー $F_\beta(\Theta)$ であるが, EM アルゴリズムとの比較のため, 図における縦軸は $-F_\beta(\Theta)$ としてある.

$F_\beta(\Theta)$ は β の値が十分に小さい (高温) ときには, 唯一の極値を持つ. このような解空間ではいかなる初期値を与えても, EM ステップを繰り返すことでその唯一の極値へ収束することができる (図 3.5(a)). その後, 前回の温度での推定値を初期値に用いて徐々に β の値を増加させ (温度を下げ) ながら推定を行なっていく (図 3.5(b)(c)). 最終的には $\beta = 1$ とし解空間 $F_\beta(\Theta)$ を対数尤度 $\mathcal{L}(\Theta)$ と一致させ, 推定を行なうことで最尤推定値を得る (図 3.5(d)).

今回の例では, EM アルゴリズムでは局所最適解へ収束してしまうような初期値

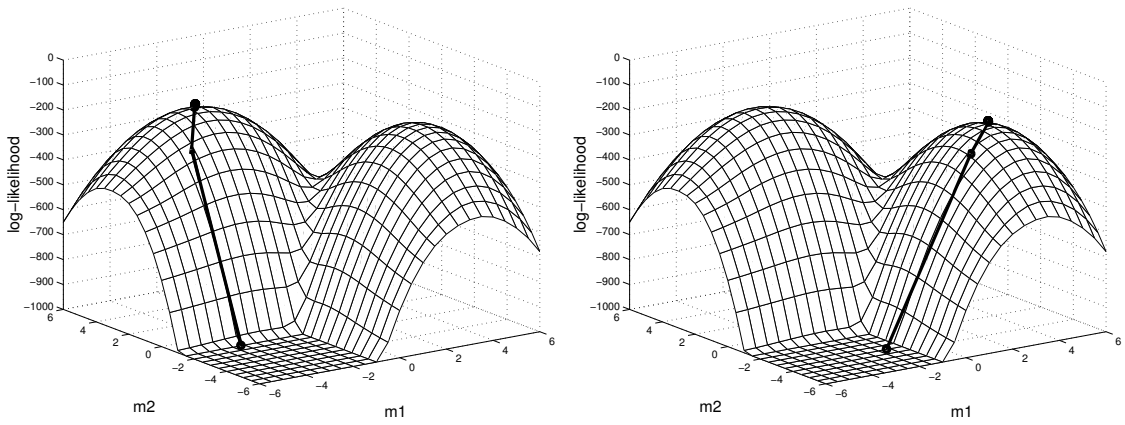


図 3.4: EMアルゴリズムの推定結果の例: 初期値 $(m_1^{(0)}, m_2^{(0)}) = (-4, -2)$(左) と初期値 $(m_1^{(0)}, m_2^{(0)}) = (-2, -4)$(右)

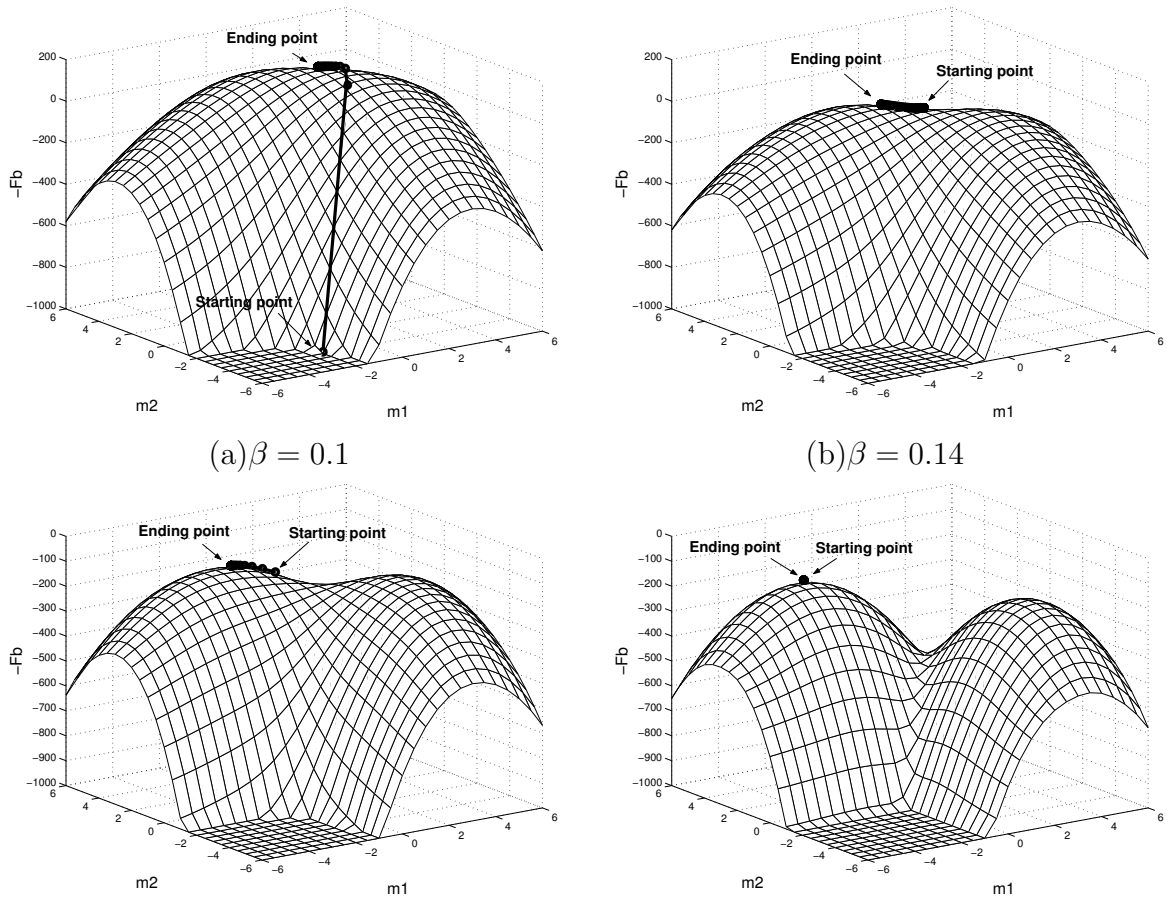
を与えても, DAEMアルゴリズムでは大域的最適解を得ることができた. 初期値に依存しないという面で, EMアルゴリズムに比べ良解が得られる. ただし, DAEMアルゴリズムが常に大域的最適解へ到達する保証はない. ここで行なったシミュレーションでは大域的最適解へ到達する場合を示してあるが, 実際には局所最適解へ陥ってしまうこともある. これは探索点が図3.5(b) のように鞍点に存在しているとき, 大域的最適解と局所最適解へ向かう2つの経路が存在するが, このときDAEMアルゴリズムがランダムに経路の選択を行なうためである. このような分岐点において常に正しい経路を選択できるとは限らない. また, 1度でも経路の選択を誤れば, その後の探索では大域的最適解へ到達することができない可能性がある.

3.5 多方向探索EMアルゴリズム

前節で見たように, DAEMアルゴリズムは鞍点でのランダム性により, 初期値の設定法とは異なる局所最適性を持つ. DAEMアルゴリズムでより良い解を求めるには, 複数回の推定を必要とし, 計算コストも大きくなる. 本節では, DAEMアルゴリズムの利点を利用した多方向探索型のアルゴリズムを導入することによりEMアルゴリズムの局所最適性の問題の改善に取り組む.

3.5.1 多方向探索EMアルゴリズム

前節で見たように, EMアルゴリズムは推定結果が初期値に大きく依存するという局所最適性の問題がある. また, DAEMアルゴリズムにおいても, 探索点が鞍



(a)$\beta = 0.1$ (b)$\beta = 0.14$

(c)$\beta = 0.753$ (d)$\beta = 1.0$

図 3.5: DAEMアルゴリズムの推定結果の例: 初期値 $(m_1^{(0)}, m_2^{(0)})$

点 (saddle point) に停留してしまう場合, 探索方向をランダムに決定するため, 探索方向を誤ると良い解を見逃してしまう危険性があることを指摘した. 従って, より良い解を見つけるためにDAEMアルゴリズムを複数回繰り返すといった対処が取られる. しかしながら, DAEMアルゴリズムはアニーリングプロセスも行うため, 全体として計算コストが高くなる傾向にある.

そこで, 目的関数, つまり対数尤度関数のHesse行列を調べることで, 鞍点における探索方向を解析的に求め, そのような点において探索点をいくつか生成し, 多方向探索を行う手法を提案する.

3.5.2 原始初期点

ここでは, 多方向探索の初期点となる**原始初期点 (primitive initial point: PIP)** について説明を行う.

原始初期点はDAEMアルゴリズムの高温 $\beta \approx 0$ における解として定義される.

ここで, DAEMアルゴリズムの事後分布式 (3.15), 更新式 (3.16a), (3.16b), (3.16c) を再掲する:

$$
\begin{aligned}
q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)}) &= \frac{\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k,\boldsymbol{\Sigma}_k)^\beta}{\sum_{k=1}^{K}\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k,\boldsymbol{\Sigma}_k)^\beta},\\
\pi_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)}),\\
\boldsymbol{m}_k^{(t+1)} &= \frac{\sum_{i=1}^{n}\boldsymbol{x}_i q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})},\\
\boldsymbol{\Sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{m}_k^{(t+1)})(\boldsymbol{x}_i-\boldsymbol{m}_k^{(t+1)})^\top q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})}{\sum_{i=1}^{n} q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})}.
\end{aligned}
$$

これらの式において $\beta = 0$ を代入することで以下を得る:

$$
\begin{aligned}
q(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)}) &= \frac{1}{K}, && \text{(3.20a)}\\
\pi_k^{(t+1)} &= \frac{1}{K}, && \text{(3.20b)}\\
\boldsymbol{m}_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, && \text{(3.20c)}\\
\boldsymbol{\Sigma}_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{m}_k^{(t+1)})(\boldsymbol{x}_i-\boldsymbol{m}_k^{(t+1)})^\top. && \text{(3.20d)}
\end{aligned}
$$

式 (3.20a) は全ての混合要素で事後分布が等しく, 一様分布となることを意味する. 故に, 式 (3.20b) の混合比も同様に全ての混合要素で等しくなる. 更に, 式 (3.20c), (3.20d) は, 全ての混合要素で標本平均, 標本分散, すなわち最尤推定値となっている. これらのことから, DAEMアルゴリズムは $\beta \approx 0$ において唯一つの解を持つことがわかる. また, 図 3.5(a) からもこの事実が確認できる. この解を原始初期点 (primitive initial point: PIP) と呼び, $\boldsymbol{\theta}^{\mathrm{PIP}} = \{\pi^{\mathrm{PIP}}, \boldsymbol{m}^{\mathrm{PIP}}, \boldsymbol{\Sigma}^{\mathrm{PIP}}\}$ と表記する. ただし,

$$
\begin{aligned}
\pi^{\mathrm{PIP}} &= \frac{1}{K},\\
\boldsymbol{m}^{\mathrm{PIP}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,\\
\boldsymbol{\Sigma}^{\mathrm{PIP}} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{m}^{\mathrm{PIP}})(\boldsymbol{x}_i-\boldsymbol{m}^{\mathrm{PIP}})^\top,
\end{aligned}
$$

である.

原始初期点は温度を下げて(βを大きくして)いくと, 鞍点となることが実験的に確認されている. 図3.3などにおける点$\{m_1, m_2\} = \{1, 1\}$はこの一例である. この事実は, DAEMアルゴリズムでは, βを大きくしたときに**必ず**ランダムな摂動が必要となることを意味しており, 結果DAEMアルゴリズムの解品質は, 原始初期点における最初の摂動に依存する. しかしながら, DAEMアルゴリズムが良い性能を示すことが [11] で報告されているため, 原始初期点を上手く利用することでより良い解が得られると考えられる.

次節では, 原始初期点におけるHesse行列を利用し慎重に探索方向を生成する多方向探索EMアルゴリズムについて説明を行う.

3.5.3 多方向探索EMアルゴリズム

本節では, 原始初期点と, 原始初期点における目的関数のHesse行列(付録B)を利用した多方向探索EMアルゴリズムについて説明を行う.

前節で述べたように, DAEMアルゴリズムは良い解を得るためにアニーリングプロセスに加え, 複数回の推定を必要とするため計算コストが高くなる. そこで, アニーリングプロセスを行う代わりに, 原始初期点でのHesse行列を利用することで鞍点での探索方向を解析的に決定し, 原始初期点を初期値として多方向に探索点を割り振り推定を行う, 多方向探索アルゴリズムを考える. 図3.6に鞍点における探索方向のアイディアを示す. EMアルゴリズムの枠組みでは, 対数尤度関数を最大化することが目的であるため, 図のような鞍点においては赤色の矢印の方向に探索点を割り振るべきである. このような方向は, 鞍点での目的関数のHesse行列の固有値および固有ベクトルを調べることで求めることが可能である. 従って, 原始初期点におけるEMアルゴリズムの目的関数, 式(3.6)のHesse行列を計算し, その固有値, 固有ベクトルを計算する. 得られた固有ベクトルの中で正の固有値に対応する固有ベクトル方向に探索を行うことで, 対数尤度を増加させるパラメータ方向へ推定を行うことが可能である.

しかし, データの次元数p, 混合数Kが大きい場合, 正の固有値の数も膨大となり, その全ての方向に探索点を割り当てるのは効率的でない. 特に小さな正の固有値方向に探索を行っても対数尤度を大きく増加させることはできない. 更に, 複数の正の固有値に対応する固有ベクトルによって張られる空間内であれば, どちらの方向に探索を行っても対数尤度を増加させることができる. これらの事実から, なるべく大きな正の固有値を用い, かつ, 選択された固有値に対応する固有ベクトル

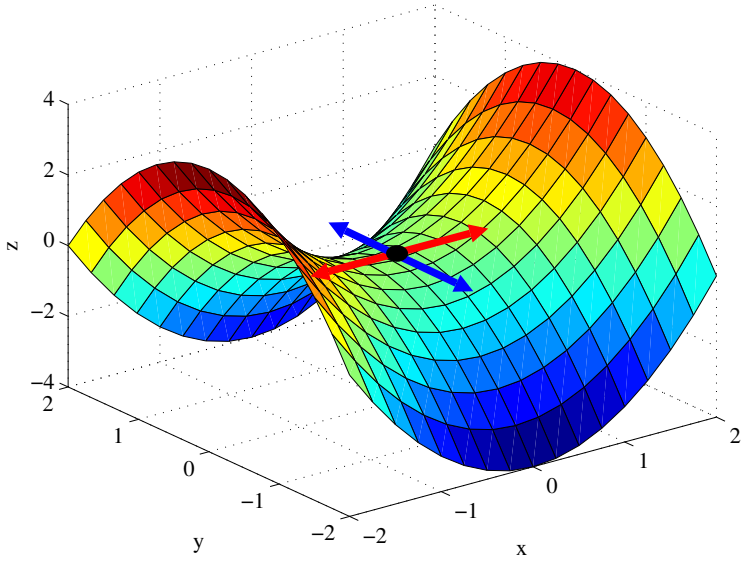


図 3.6: 鞍点と探索すべき方向: 目的関数の最大化を行う場合, 図中の赤色の矢印の方向に探索点を割り振る.

によって張られる空間内において効率の良い探索方向を生成するため, 以下で説明する方法によって探索方向を生成する.

まず, 原始初期点において目的関数 (式 (3.6)) の Hesse 行列がいくつかのの正の固有値を持つとする. 大きな正の固有値に対する固有ベクトル方向は目的関数の曲率が高いため, ヒューリスティックとして, 固有値の値の大きいものから順に累積寄与率が 80%となる固有値までを選択する. 今, 選択された固有値の個数を R とし, それらの固有値に対応する固有ベクトルの正規直交系を $U = \{\boldsymbol{u}_r\}_{r\in\mathbb{N}_R}$ とする. このとき, 以下で与えられるベクトル集合 V を探索方向に加える:

$$\begin{aligned} V &= \left\{ \sum_{r=1}^{R} \{-,+\} \boldsymbol{u}_r \right\} \\ &= \{(\boldsymbol{u}_1 + \cdots + \boldsymbol{u}_R), \ldots, (-\boldsymbol{u}_1 - \cdots - \boldsymbol{u}_R)\} \\ &= \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{2^R}\}. \end{aligned}$$

ただし, 上式においては $\{-,+\}$ により, R 個の固有ベクトルの全ての正負の組み合わせを意味するものとする. 従って, 実際に探索点を割り当てる方向は

$$\pm U \cup V = \{\pm\boldsymbol{u}_1, \ldots, \pm\boldsymbol{u}_R, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{2^R}\},$$

の $2R + 2^R$ 方向となる. 例として, 図 3.7 に重複度 $R = 2$ の場合の二次元固有空間内

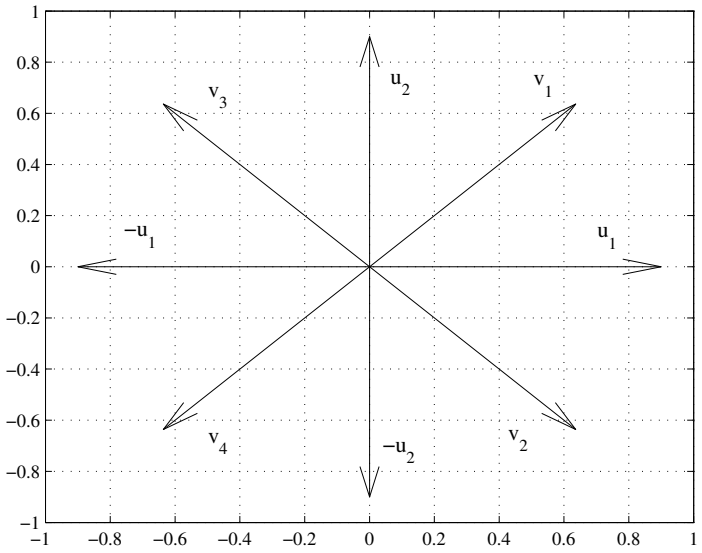


図 3.7: 2次元固有空間内で生成されるベクトル群

の分岐方向を示す. 固有空間が2次元の場合, $2 \times 2 + 2^2 = 8$ より計8つのベクトルが生成される.

上記のようなベクトル集合Vを考える根拠は次のようである. 分岐すべき方向を固有ベクトルによって張られる空間内で一様に生成するために, R次元超球を考える. $\boldsymbol{x}$を単位ベクトル, Sを$\cos\theta_r$のrに関する和とする. ただし, $\cos\theta_r$は$\boldsymbol{x}$と固有ベクトル$\boldsymbol{u}_r$のなす角とする:

$$S = \sum_{r=1}^{R} \cos\theta_r = \sum_{r=1}^{R} \boldsymbol{x}^\top \boldsymbol{u}_r.$$

θ_rの範囲を$0 \leq \theta_r \leq \pi/2$とすれば, θ_rが大きくなるほど$\cos\theta_r$は小さくなる. 従って, $||\boldsymbol{x}|| = 1$のもとでSを最小化することを考える. すなわち, 各固有ベクトルからの角度の和が最大となる単位ベクトル$\boldsymbol{x}$を考えることにする. なぜなら, 出来るだけ各固有ベクトルからの角度の離れた方向を探索する方が効率的であると思われるからである. 制約条件$||\boldsymbol{x}|| = 1$があるため, ラグランジェ乗数λを含む以下の式を最小化することになる:

$$J = \sum_{r=1}^{R} \boldsymbol{x}^\top \boldsymbol{u}_r - \lambda(\boldsymbol{x}^\top \boldsymbol{x} - 1). \tag{3.21}$$

実際に式 (3.21) を最小化すべく $\partial J/\partial \boldsymbol{x} = 0$ を解くと,

$$\boldsymbol{x}^* = \frac{1}{\sqrt{R}} \sum_{r=1}^{R} \boldsymbol{u}_r, \tag{3.22}$$

となる. 式 (3.22) は任意の固有ベクトル $\boldsymbol{u}_r$ と等しい角をなす.

3.5.4 多方向探索 EM アルゴリズムの流れ

以下に多方向探索 EM アルゴリズムの流れを示す:

[多方向探索 EM アルゴリズム]

Step 1. 原始初期点 $\boldsymbol{\theta}^{\mathrm{PIP}} = \{\pi^{\mathrm{PIP}}, \boldsymbol{m}^{\mathrm{PIP}}, \boldsymbol{\Sigma}^{\mathrm{PIP}}\}$ を計算.

Step 2. $\boldsymbol{\theta}^{\mathrm{PIP}}$ における対数尤度関数の Hesse 行列を計算.

Step 3. 累積寄与率に応じて R 個の正の固有値を選択.

Step 4. 初期値の集合 $\{\boldsymbol{\theta}_r^{\mathrm{init}}\}_{r\in\mathbb{N}_{2R+2^R}} = \{\boldsymbol{\theta}^{\mathrm{PIP}} + \Delta\boldsymbol{\theta}_r\}_{r\in\mathbb{N}_{2R+2^R}}$ を生成.

Step 5. Step 4 により生成された各初期値を用いて EM アルゴリズムによりパラメータを推定.

上記のアルゴリズムにより得られたパラメータの中で, 対数尤度の最も高いパラメータを最終的な解とする.

3.6 計算機実験

最後に, いくつかの実データを用いて, 多方向探索 EM アルゴリズムの性能を評価する. 実験の目的は, k-mean 法によって初期値を生成した場合の EM アルゴリズムで得られた最良解と, 多方向探索 EM アルゴリズムで得られた最良解の比較である. 実験に用いるデータを表 3.1 にまとめる. 結果を表 3.2 に示す.

表からわかるように, 10 個のデータセット中 6 個で多方向探索 EM アルゴリズムが良い解を得ており, 1 個のデータセットで k-means 法による EM アルゴリズムと同じ解を得ている. すなわち, 10 個のデータセット中 7 個で多方向探索 EM アルゴリズムが良い解を得ており, また, 各データセットにおいて得られた解のばらつきも比較的小さいと言える. また計算時間も k-means 法による初期値生成と同程度であり, Hesse 行列を用いた初期値生成は, EM アルゴリズムの初期値生成手段として有効であると言える.

表 3.1: Data Set

Data	次元数 p	データ数 n	Source
australian(#1)	11	689	UCI[39]
bodyfat(#2)	14	252	StatLib[40]
breast-cancer(#3)	10	683	UCI
diabetes(#4)	7	768	UCI
heart(#5)	11	270	UCI
iris(#6)	4	137	UCI
liver-disorders(#7)	5	345	UCI
mpg(#8)	6	383	UCI
space-ga(#9)	6	3107	StatLib
wine(#10)	13	175	UCI

3.7 まとめ

本章では, EMアルゴリズムによる混合正規分布推定を取り扱った. まず, 最尤法による正規分布推定を説明し, シミュレーションにより, データが正規分布に従わない場合に適切なパラメータを推定することができないことを見た. 次にEMアルゴリズム, DAEMアルゴリズムの説明を行い, これらのアルゴリズムによる混合正規分布推定の定式化を行った. 人工データによるシミュレーションにより, EMアルゴリズムの局所最適性の問題およびDAEMアルゴリズムの挙動を可視化した. このような問題を改善するために, DAEMアルゴリズムの枠組みから得られる原始初期点を定義し, 原始初期点とそこでのHesse行列を用いた多方向探索アルゴリズムを提案し, 実データによる計算機実験によりその有効性を示した. なお, 本文では割愛したが, Sampling Subsampling法 [41] を利用した手法 [42] や一部で温度パラメータを用いることで不要な探索点を枝刈りする手法 [43] などによる多方向探索EMアルゴリズムの拡張を行い, 解品質の向上, 計算コスト削減を行うことも可能である.

表 3.2: EM(kmeans) vs. EM(PIP)

Data	Method	Best	Average ± S.E.	Computational Time			Initial Points
				Init.	Est.	Total	
#1	EM(kmeans)	**-783**	-1971 ±1188.49	0.02	0.03	0.05	2
	EM-PIP	-1683	-1854 ±171.23	0.09	0.07	0.16	2
#2	EM(kmeans)	1188	1155 ±9.99	0.07	1.74	1.81	14
	EM-PIP	**1209**	1188 ±4.20	0.07	1.94	2.01	14
#3	EM(kmeans)	**9471**	4285 ±515.74	0.19	0.29	0.48	24
	EM-PIP	5721	3045 ±262.80	0.07	0.57	0.64	24
#4	EM(kmeans)	2453	1279 ±216.12	0.11	0.14	0.25	14
	EM-PIP	**4849**	2028 ±312.38	0.04	0.46	0.50	14
#5	EM(kmeans)	326	-544 ±184.36	0.03	0.07	0.10	8
	EM-PIP	**947**	-1168 ±329.39	0.05	0.11	0.16	8
#6	EM(kmeans)	68	45 ±0.37	5.00	34.99	39.99	1044
	EM-PIP	**94**	52 ±0.17	0.49	52.86	53.35	1044
#7	EM(kmeans)	**460**	450 ±1.98	0.07	3.14	3.21	14
	EM-PIP	**460**	456 ±0.81	0.04	3.14	3.18	14
#8	EM(kmeans)	2506	742 ±16.69	6.32	6.73	13.05	1044
	EM-PIP	**4141**	699 ±22.14	0.53	13.74	14.27	1044
#9	EM(kmeans)	6863	6863 ±0.00	0.04	5.01	5.05	2
	EM-PIP	**8677**	8677 ±0.00	0.10	3.61	3.71	2
#10	EM(kmeans)	**-147**	-277 ±0.67	5.43	67.29	72.72	1044
	EM-PIP	-241	-277 ±0.56	0.64	75.50	76.14	1044

第 4 章

変分ベイズ法と混合正規分布推定

混合正規分布推定を行う手法として, 前章で説明したEMアルゴリズムによる方法の他に, マルコフ連鎖モンテカルロ法 (MCMC) 法やベイズ推定を近似した変分ベイズ (variational bayes method: VB) 法などがある. 変分ベイズ法はパラメータの事後分布の近似を行うことができる, オーバーフィットを避けられる, 事前知識を事前分布として導入可能であるなどの利点を持つ. 本章では, 混合正規分布推定における変分ベイズ法の局所最適性の問題を考察する. 変分ベイズ法の目的関数は, EMアルゴリズムの尤度関数と同様, 一般に極大点が多数存在し, 良いパラメータを推定するためには, 初期値を慎重に選ばなければならない. 本章では, 前章で導入した原始初期点を用いた多方向探索アルゴリズムを変分ベイズ法へも適用することで, 良い解を得られることを見る [44].

4.1 ベイズ統計

ベイズ推定は, パラメータの点推定を行う最尤法とは異なり, パラメータを確率的に変動する変数, すなわち確率変数として捉え, 事前分布と尤度の積を考えることによりパラメータの事後分布を導出する. 本節では, ベイズ推定の枠組みを説明した後, ベイズ推定による正規分布のパラメータの事後分布推定について説明を行う.

4.1.1 ベイズ推定

ベイズ統計では例外なくベイズの定理:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)},$$

を出発点とし推定を行なう. ここで, X, Y は確率変数, P は適当な確率分布とする. ベイズの定理を用いると未知パラメータ Θ と観測データ $\mathcal{D}$ に関して以下の式が得

られる:

$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{p(\mathcal{D})}.$$

上記の式は以下を意味する:

$$posterior \propto likelihood \times prior. \tag{4.1}$$

最尤法では尤度方程式を解くことでパラメータを推定するが, ベイズ推定では全ての未知パラメータを確率変数と見なし, それらに何らかの事前分布を与え, データが得られたとき式 (4.1) を用いてそのパラメータの事後分布を求める.

事前分布としてどのようなものを用いるかは様々な選択肢があるが, 一般には **共役事前分布 (conjugate prior)** を用いる. 共役事前分布とは, 尤度関数との積, すなわち事後分布が再び事前分布と同じ関数形となるような分布のことである. このような分布を事前分布として用いれば, 数学的に扱いやすいだけでなく, 新たなデータに対して事後分布を次の事前分布とすることで推定結果の調整が簡単になるという利点もある.

4.1.2 ベイズ推定による正規分布推定

ここでは, 平均ベクトル, 分散共分散行列ともに未知の場合の事後分布を導出する. 前節のように, ベイズ推定では, 式 (4.1) に従って, パラメータの事後分布を求める. いま, n 個の観測データ $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ を用いて, 多変量正規分布の平均ベクトル $\boldsymbol{m}$, 精度行列 $\boldsymbol{\Lambda}$ を推定したいとする. ここで, 本章では分散共分散行列 $\boldsymbol{\Sigma}$ ではなく, 精度行列 $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ を用いることに注意されたい. このとき, 尤度関数は,

$$\begin{aligned} p(\{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}|\boldsymbol{m},\boldsymbol{\Lambda}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Lambda}|^{-1}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{m})^\top\boldsymbol{\Lambda}(\boldsymbol{x}_i-\boldsymbol{m})\right\} \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Lambda}|^{-n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{m})^\top\boldsymbol{\Lambda}(\boldsymbol{x}_i-\boldsymbol{m})\right\} \\ &\propto |\boldsymbol{\Lambda}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{S}_0)\right\}, \end{aligned} \tag{4.2}$$

で与えられる. ここで,

$$\boldsymbol{S}_0 = \sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{x}_i-\boldsymbol{m})^\top,$$

とした. また, ベクトル, $\boldsymbol{a}$, $\boldsymbol{b}$ および行列 $\boldsymbol{X}$ に対して, $\boldsymbol{a}^\top\boldsymbol{X}\boldsymbol{b} = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{b}\boldsymbol{a}^\top)$, という関係式を利用した. 多変量正規分布の平均ベクトルおよび精度行列に対する共役事

前分布はGaussian-Wishart分布 (付録A.6) が用いられる:

$$
\begin{aligned}
p(\boldsymbol{m}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \eta_0, \boldsymbol{W}_0, \nu_0) &= \mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu}_0, (\eta_0\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}_0, \nu_0), \qquad (4.3)\\
\mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu}_0, (\eta_0\boldsymbol{\Lambda})^{-1}) &\propto |\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{\mu}_0)^\top(\eta_0\boldsymbol{\Lambda})(\boldsymbol{m}-\boldsymbol{\mu}_0)\right\},\\
\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}_0, \nu_0) &\propto |\boldsymbol{\Lambda}|^{(\nu_0-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda})\right\}.
\end{aligned}
$$

ただし, $\boldsymbol{\mu}_0$, η_0, $\boldsymbol{W}_0$, ν_0 はハイパーパラメータである. 尤度関数 (式 (4.2)) と事前分布 (式 (4.3)) の積を取り, 整理することで, 平均ベクトル, 精度行列の事後分布は,

$$
p(\boldsymbol{m}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|W, \nu).
$$

と計算される. ここで,

$$
\begin{aligned}
\eta &= \eta_0 + n,\\
\boldsymbol{\mu} &= \frac{\eta_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\eta},\\
\boldsymbol{W}^{-1} &= \boldsymbol{W}_0^{-1} + \boldsymbol{S} + \frac{\eta_0 n}{\eta}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^\top,\\
\nu &= \nu_0 + n,
\end{aligned}
$$

であり,

$$
\begin{aligned}
\bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,\\
\boldsymbol{S} &= \sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top,
\end{aligned}
$$

である.

次に, 単変量正規分布の人工データを用いて, 事前分布およびデータが得られた際の事後分布の様子を可視化する. 実験に用いるデータは $X \sim \mathcal{N}(X|0.8, 0.1)$ から生成された10個のデータを用いる. また, 事前分布のハイパーパラメータは $m_0 = 0$, $\eta_0 = 2$, $W_0 = 0.5$, $\nu_0 = 5$ と設定した (単変量正規分布であるので, 全てスカラーであり, 実際には付録A.3のGaussian-Gamma分布となる). 事前分布を図4.1, 事後分布を図4.2に示す.

ベイズ推定では, データが得られる前は, パラメータ (ここでは, 平均と精度) は事前分布に従う. しかし, 推定に用いるデータが増えるに従って, それらのデータを反映した事後分布が得られる. 図からわかるように, 観測データを得ることで平均は真の値である0.8に近くなっており精度も事前分布よりも大きくなっているこ

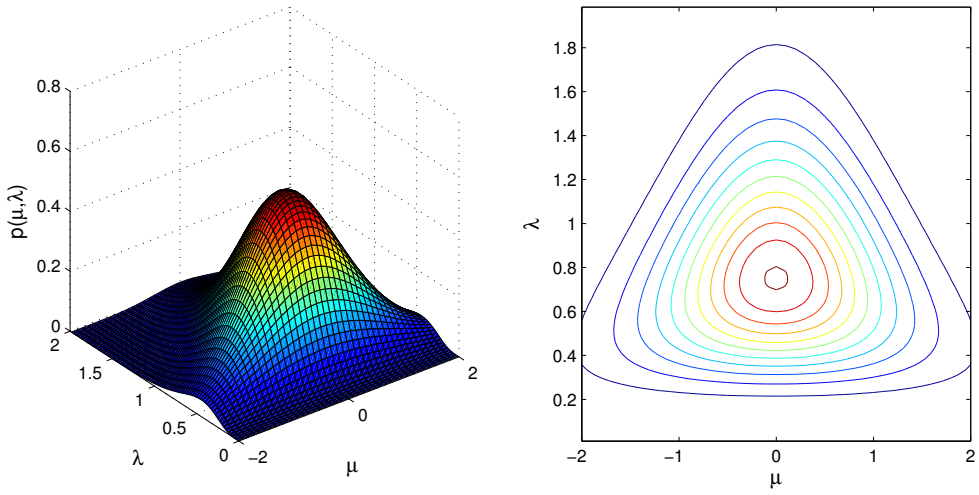


図 4.1: μ, λ の事前分布

とがわかる. さらに分布の分散が事前分布よりも小さくなっており, 最大値付近がより信頼できるものとなっていることが見て取れる.

このように, ベイズ統計の枠組みでは, 事前知識を事前分布として設定し, データが得られた際に, 式 (4.1) を適用することでパラメータの事後分布を推定する. 事後分布は, 利用できるデータが少数の場合は事前分布を重視し, データが多数ある場合は, 事前分布の影響が薄れ, データにフィットした事後分布が得られる仕組みとなっている.

4.2 平均場近似

前節で見たように, ベイズ推定では全ての未知パラメータに対してそれぞれ事前分布を設定し, 尤度との積を取ることによりパラメータの事後分布を得た. しかし, パラメータ数が多い場合や混合正規分布などのように潜在変数を含むような場合は事後分布を解析的に導出することが困難になる. 変分ベイズ法では, 平均場理論に基づいた方法で事後分布の近似分布を求める [8, 9]. この節では変分ベイズ法について説明する [45, 46, 47].

今, $X = \{\boldsymbol{x}_i\}_{i \in \mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ を観測データ, $Z = \{\boldsymbol{Z}_m\}_{m \in \mathbb{N}_M}$, $\boldsymbol{Z}_i \in \mathcal{Z}_m$ を未知パラメータ群とする ($\mathcal{Z}_m$ は m のパラメータ空間). このとき, 任意の確率分布 $q(Z)$(variational distribution と呼ばれる) を用いることで周辺尤度 $\ln p(X)$(自由エネ

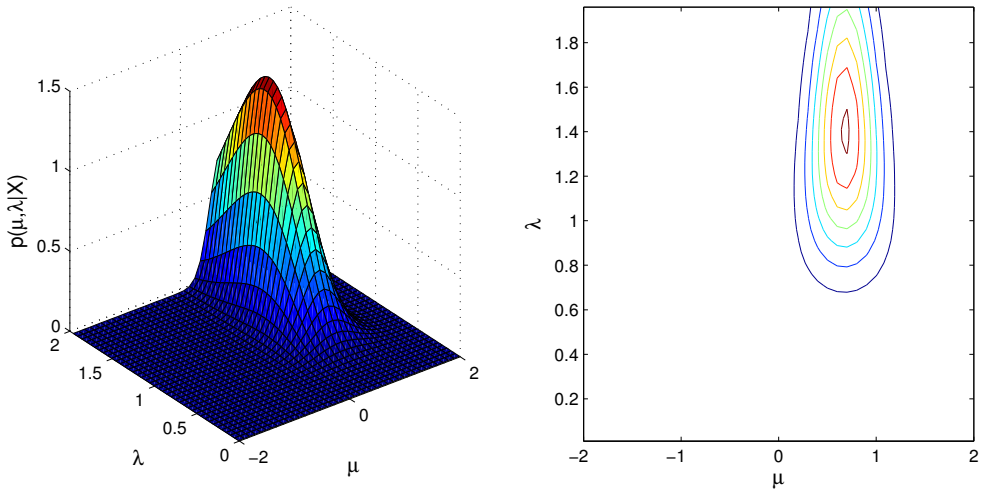


図 4.2: μ, λ の事後分布

ルギーに相当する量) は以下のように分解される.

$$\ln p(X) = \mathcal{L}(q) + \mathrm{KL}(q||p), \tag{4.4}$$

ただし,

$$\begin{aligned} \mathcal{L}(q) &= \int q(Z) \ln \left\{ \frac{p(X,Z)}{q(Z)} \right\} dZ, \\ \mathrm{KL}(q||p) &= -\int q(Z) \ln \left\{ \frac{p(Z|X)}{q(Z)} \right\} \mathrm{d}Z, \end{aligned} \tag{4.5}$$

とおいた. $\mathrm{KL}(q||p)$ はパラメータの事後分布と variational distribution との Kullback-Leibler divergence であり, 常に正となる. したがって, 周辺尤度は以下のように汎関数 $\mathcal{L}(q)$ によって下側の評価ができる:

$$\ln p(X) \geq \mathcal{L}(q).$$

式 (4.4) より, $q(Z) = p(Z|X)$ のときこの下限値は最大となることがわかる. 平均場理論ではこの $\mathcal{L}(q)$ を変分自由エネルギー (variational free energy) と呼び, これを最大化することで事後分布を求める. なお, 今後は $q(\boldsymbol{Z}_m)$ を q_m と略記する.

上で説明したように, 任意の $q(Z)$ を用いれば事後分布を正確に求めることが可能であるが, 実際にはこのような q を用いることは困難である. 従って, q に対して何らかの制約を加えて事後分布を近似的に求めることになる. ここでは, 各々の未

知パラメータ $\{\boldsymbol{Z}_m\}_{m\in\mathbb{N}_M}$ が独立の確率分布に従うと仮定する:

$$q(Z) = \prod_{m=1}^{M} q_m, \tag{4.6}$$

このような仮定のもとで事後分布を求める手法を特に variational mean field method(あるいは naive mean field method) と呼ぶ. この仮定により, 式 (4.5) は以下のように書ける:

$$\mathcal{L}(q) = \int \prod_{m=1}^{M} q_m \left\{ \ln p(X,Z) - \sum_{m=1}^{M} \ln q_m \right\} \mathrm{d}Z.$$

いま, この下限値を q_ℓ に関して最大化したいとすると, $\boldsymbol{Z}_\ell$ に依存する項と依存しない項を以下のようにわけることができる.

$$\begin{aligned}
\mathcal{L}(q) &= \int q_\ell \left\{ \int \ln p(X,Z) \prod_{m\in\mathbb{N}_M, m\neq l} q_m \mathrm{d}\boldsymbol{Z}_m \right\} \mathrm{d}\boldsymbol{Z}_\ell - \int q_\ell \ln q_\ell \mathrm{d}\boldsymbol{Z}_\ell + \mathrm{const} \\
&= \int q_\ell \{ \mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)] + \mathrm{const} \} \mathrm{d}\boldsymbol{Z}_\ell - \int q_\ell \ln q_\ell \mathrm{d}\boldsymbol{Z}_\ell + \mathrm{const} \\
&= \int q_\ell \ln \left\{ \frac{\exp(\mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)]) + \mathrm{const}}{q_m} \right\} \mathrm{d}\boldsymbol{Z}_\ell + \mathrm{const} \\
&= -\mathrm{KL}(\exp(\mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)]) + \mathrm{const} || q_\ell)
\end{aligned}$$

ただし,

$$\int \ln p(X,Z) \prod_{m\in\mathbb{N}_M, m\neq\ell} q_m \mathrm{d}\boldsymbol{Z}_m = \mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)] + \mathrm{const},$$

である. これより, $\mathcal{L}(q)$ を最大化する $q_\ell^*(\boldsymbol{Z}_\ell)$ が以下のように求まる:

$$q_\ell^* = \frac{\exp(\mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)])}{\int \exp(\mathbb{E}_{Z\setminus\boldsymbol{Z}_\ell}[\ln p(X,Z)]) \mathrm{d}\boldsymbol{Z}_\ell}.$$

従って, 全ての未知パラメータ $\{\boldsymbol{Z}_m\}_{m\in\mathbb{N}_M}$ について式 (4.7) を繰り返し求めていくことで周辺尤度の下限値を最大化する.

4.3 変分ベイズ法

いま, $X = \{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}, \boldsymbol{x}_i \in \mathbb{R}^p$ を観測データ集合, $Z = \{\boldsymbol{Z}_i\}_{i\in\mathbb{N}_n}, \boldsymbol{Z}_i \in \mathcal{Z}$ を各観測データに対する潜在変数の集合, $\Theta \in \mathcal{M}$ をモデルパラメータの集合とする. ただし, $\mathcal{Z}$ は潜在変数の空間であり, $\mathcal{M}$ はモデルパラメータ空間 である. また, X, Z, Θ

の同時確率を $p(X, Z, \Theta)$, 変分分布を $q(Z, \Theta)$ とする. このとき, すべての未知量を周辺化した対数周辺尤度 $\ln p(X)$ は,

$$\begin{aligned}
\ln p(X) &= \sum_Z \ln \int p(X, Z, \Theta) \mathrm{d}\Theta \\
&= \sum_Z \ln \int q(Z, \Theta) \frac{p(X, Z, \Theta)}{q(Z, \Theta)} \mathrm{d}\Theta \\
&\geq \sum_Z \int q(Z, \Theta) \ln \frac{p(X, Z, \Theta)}{q(Z, \Theta)} \mathrm{d}\Theta \qquad (4.7) \\
&= \mathcal{L}(q(Z, \Theta)),
\end{aligned}$$

のように変形でき, 下限値 $\mathcal{L}(q(Z, \Theta))$ を得ることができる. ここで, 未知量, $\{Z, \Theta\}$ に対する変分分布 $q(Z, \Theta)$ が前節のように,

$$q(Z, \Theta) = q(Z) q(\Theta),$$

と因子化できると仮定し, 汎関数 $\mathcal{L}(q(Z, \Theta)) = \mathcal{L}(q(Z), q(\Theta))$ を $q(Z), q(\Theta)$ について逐次最大化することで周辺尤度を最大化する. $\mathcal{L}(q(Z), q(\Theta))$ を $q(Z)$ で偏微分し, 0 とおくことで以下を得る:

$$\begin{aligned}
&\frac{\partial}{\partial q(Z)} \mathcal{L}(q(Z), q(\Theta)) = 0, \\
&q(Z) = \frac{1}{C_Z} \exp\Big\{ \mathbb{E}_\Theta \big[\beta \ln p(X, Z|\Theta) \big] \Big\}. \qquad (4.8)
\end{aligned}$$

また, $\mathcal{L}(q(Z), q(\Theta))$ を $q(\Theta)$ で偏微分し, 0 とおくことで以下を得る:

$$\begin{aligned}
&\frac{\partial}{\partial q(\Theta)} \mathcal{L}(q(Z), q(\Theta)) = 0, \\
&q(\Theta) = \frac{1}{C_\Theta} p(\Theta) \exp\Big\{ \mathbb{E}_Z \big[\ln p(X, Z|\Theta) \big] \Big\}. \qquad (4.9)
\end{aligned}$$

ここで, $p(\Theta)$ はモデルパラメータの事前分布, $p(X, Z|\Theta)$ は (完全データの) 対数尤度である. また, C_Θ, C_Z は正規化項である.

変分ベイズ法の流れは以下の通りである:

変分ベイズ法

Initialization: ハイパーパラメータの初期化.

VB-E, M Step 収束するまで以下を繰り返す.

VB-E Step: $q(Z) \leftarrow \frac{1}{C_Z} \exp\Big\{ \mathbb{E}_\Theta [\ln p(X, Z|\Theta)] \Big\}$

VB-M Step: $q(\Theta) \leftarrow \frac{1}{C_\Theta} p(\Theta) \exp\Big\{ \mathbb{E}_Z [\ln p(X, Z|\Theta)] \Big\}$

4.4 確定的アニーリング変分ベイズ法

ここでは, 変分ベイズ法に確定的アニーリングを導入した手法 [13] を紹介する. 変分ベイズ法では, 変分事後分布の初期ハイパーパラメータの設定に依存した解が得られる. これは, 変分ベイズ法の目的関数が一般に多峰であるために生じる局所最適性の問題である. この問題を改善するために提案された手法が, 確定的アニーリング変分ベイズ (deterministic annealing variational Bayes: DAVB) 法である. 確定的アニーリング法による VB 法の局所最適性の問題への対処は, EM アルゴリズムへの確定的アニーリング法の導入と同様の考えに基づく.

DAVB 法では, 式 (4.7) を更に以下のように分解する:

$$\begin{aligned}\sum_Z \int q(Z,\Theta)\ln\frac{p(X,Z,\Theta)}{q(Z,\Theta)}\mathrm{d}\Theta &= \sum_Z \int q(Z,\Theta)\ln p(X,Z,\Theta)\mathrm{d}\Theta \\ &\quad -\sum_Z \int q(Z,\Theta)\ln q(Z,\Theta)\mathrm{d}\Theta. \qquad (4.10)\end{aligned}$$

潜在変数とパラメータの値 (Z,Θ) のエネルギーを $-\ln p(X,Z,\Theta)$ とすると, 式 (4.10) の第 1 項は $q(Z,\Theta)$ のもとでの負の平均エネルギー $-E$, 第 2 項は $q(Z,\Theta)$ の負のエントロピー $-S$ と解釈できる. これを統計力学の関係式 $\mathcal{F}=E-TS$($\mathcal{F}$ は自由エネルギー, T は温度) と比べると, $\mathcal{L}(q)$ は $T=1$ のときの負の自由エネルギー $-\mathcal{F}$ と解釈できる. これらの関係より, DAVB 法では温度パラメータ $\beta=\frac{1}{T}$ を含めた以下の変分自由エネルギー $\mathcal{F}_\beta(q(Z,\Theta))$ を最大化することで周辺尤度を最大化する:

$$\mathcal{F}_\beta(q(Z,\Theta)) = \sum_Z \int q(Z,\Theta)\ln p(X,Z,\Theta)\mathrm{d}\Theta - \frac{1}{\beta}\sum_Z \int q(Z,\Theta)\ln q(Z,\Theta)\mathrm{d}\Theta. \quad (4.11)$$

ここで, $q(Z,\Theta)=q(Z)q(\Theta)$, を式 (4.11) に代入することで, 以下を得る:

$$\begin{aligned}\mathcal{F}_\beta(q(Z),q(\Theta)) &= \sum_Z \int q(Z)q(\Theta)\ln p(X,Z,\Theta)\mathrm{d}\Theta \\ &\quad -\frac{1}{\beta}\sum_Z \int q(Z),q(\Theta)\ln q(Z)q(\Theta)\mathrm{d}\Theta \\ &= \sum_Z \int q(Z)q(\Theta)\Big[\ln p(\Theta)+\ln p(X,Z|\Theta)\Big]\mathrm{d}\Theta \\ &\quad -\frac{1}{\beta}\sum_Z \int q(Z)q(\Theta)\Big[\ln q(Z)+\ln q(\Theta)\Big]\mathrm{d}\Theta.\end{aligned}$$

これを関数 $q(Z)$ に関して偏微分し, 0 とおくことで以下を得る:

$$\begin{aligned}&\frac{\partial}{\partial q(Z)}\mathcal{F}_\beta(q(Z),q(\Theta))=0, \\ &q(Z)=\frac{1}{C_Z}\exp\Big\{\mathbb{E}_\Theta\big[\beta\ln p(X,Z|\Theta)\big]\Big\}. \qquad (4.12)\end{aligned}$$

また, $q(\Theta)$ に関して偏微分し 0 とおいて以下を得る:

$$\frac{\partial}{\partial q(\Theta)}\mathcal{F}_\beta(q(Z), q(\Theta)) = 0,$$
$$q(\Theta) = \frac{1}{C_\Theta} p(\Theta)^\beta \exp\Big\{\mathbb{E}_Z\big[\beta \ln p(X, Z|\Theta)\big]\Big\}. \tag{4.13}$$

ここで, C_Z, C_Θ は正規化項である. DAVB 法では, DAEM アルゴリズムと同様, 逆温度パラメータ β を $\beta \approx 0$ から $\beta = 1$ へ徐々に増加させる各温度において, 式 (4.12), 式 (4.13) を収束するまで繰り返し計算することで, 負の自由エネルギー, 式 (4.11) を最大化する.

以下に DAVB 法の流れを示す.

確定的アニーリング変分ベイズ法

Initialization: $\beta \leftarrow \beta^{\text{init}} (\approx 0)$, ハイパーパラメータの初期化.

1. 収束するまで以下を繰り返す.

VB-E Step: $q(Z) \leftarrow \frac{1}{C_Z} \exp\Big\{\mathbb{E}_\Theta[\beta \ln p(X, Z|\Theta)]\Big\}$

VB-M Step: $q(\Theta) \leftarrow \frac{1}{C_\Theta} p(\Theta)^\beta \exp\Big\{\mathbb{E}_Z[\beta \ln p(X, Z|\Theta)]\Big\}$

2. $\beta \leftarrow \min(1, \beta \times \text{const})$ とする.

3. $\beta < 1$ ならば 1 に戻る, さもなくば終了.

4.5 混合正規分布推定

本節では, 混合正規分布推定問題における変分ベイズ法および確定的アニーリング変分ベイズ法の事後分布のハイパーパラメータ更新式を示す.

いま, n 個の観測データを $X = \{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$ と表し, 混合要素 k, $k = 1, \ldots, K$ の混合比, 平均ベクトルおよび精度行列 (分散共分散行列の逆行列) をそれぞれ π_k, $\boldsymbol{m}_k$, $\boldsymbol{\Lambda}_k$ とする. また, $\boldsymbol{\pi} = \{\pi_k\}_{k\in\mathbb{N}_K}$, $\boldsymbol{m} = \{\boldsymbol{m}_k\}_{k\in\mathbb{N}_K}$, $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}_K}$ とする. このとき, 混合正規分布は以下のように書ける:

$$p(X|\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\boldsymbol{m}_k, \boldsymbol{\Lambda}_k^{-1}).$$

変分ベイズ法の枠組みにおける混合正規分布推定では, データが与えられたとき, モデルパラメータ $\boldsymbol{\pi}$, $\boldsymbol{m}$, $\boldsymbol{\Lambda}$ それぞれの事後分布を求める. しかし, 混合モデル

の場合, データがどの分布に属するものであるかは一般に未知である. 従って, 観測データが所属するクラス情報を潜在変数として扱うことになる. ここでは, 潜在変数を $Z = \{\boldsymbol{z}_i\}_{i\in\mathbb{N}_n}, \boldsymbol{z}_i \in \{0,\ 1\}^K$ と表すことにする. ただし, $\boldsymbol{z}_i$ は K 次元ベクトルで K 個の要素のうちいずれか 1 つが 1 となり, 残りは 0 となる.

4.5.1 パラメータの事前分布

次に, パラメータ $\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}$ に対して事前分布を設定する. 混合比 $\boldsymbol{\pi}$ に対する事前分布は多項分布の共役事前分布である Dirichlet 分布 (付録 A.4) が用いられる:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}. \tag{4.14}$$

ここで, $\boldsymbol{\alpha}_0 = \alpha_0\mathbf{1}_K$($\mathbf{1}_K$ は全ての要素が 1 となる K 次元ベクトル) とした. また, 平均ベクトルおよび精度行列に対する事前分布 $p(\boldsymbol{m}, \boldsymbol{\Lambda})$ は Gaussian-Gamma 分布の多変量分布である Gaussian-Wishart 分布 (付録 A.6) が用いられる:

$$\begin{aligned} p(\boldsymbol{m}, \boldsymbol{\Lambda}) &= p(\boldsymbol{m}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{m}_k|\boldsymbol{\mu}_0, (\eta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0). \end{aligned} \tag{4.15}$$

これらの事前分布における, α_0, η_0, $\boldsymbol{\mu}_0$, $\boldsymbol{W}_0$, ν_0 はハイパーパラメータで, 事前に設定する必要がある.

4.5.2 変分ベイズ法によるパラメータの事後分布推定

まず, 変分ベイズ法による混合正規分布のパラメータの事後分布を変分ベイズ法により導出する. 式 (4.7) 中の $p(X, Z)$ に相当する全ての確率変数の同時分布 $p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})$ は以下のように分解される:

$$p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) = p(X|Z, \boldsymbol{m}, \boldsymbol{\Lambda})p(Z|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{m}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}). \tag{4.16}$$

ここで右辺の第 1, 第 2 因子は以下で与えられる:

$$p(X|Z, \boldsymbol{m}, \boldsymbol{\Lambda}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{m}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}}, \tag{4.17}$$

$$p(Z|\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{z_{ik}}. \tag{4.18}$$

また, 第 3 および第 4, 5 因子はパラメータに関する事前分布で式 (4.14) および式 (4.15) により与えられる. 式 (4.16) は図 4.3 と対応している.

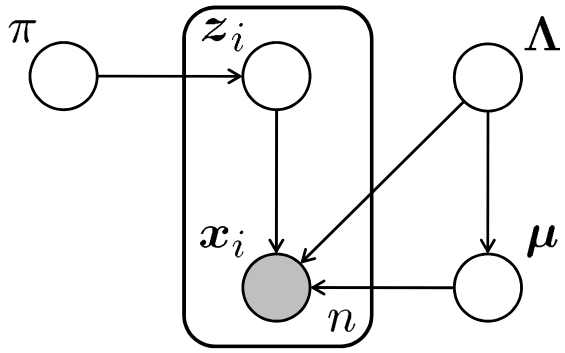


図 4.3: 混合正規分布のグラフィカルモデル

次に, 式 (4.6) に基づき未知パラメータに対する variational distribution を以下のように分解する:

$$q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) = q(Z)q(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}).$$

これで事後分布を計算する準備が整ったので, 式 (4.8) あるいは式 (4.9) に基づいて全ての未知パラメータについてこれを計算していく.

混合正規分布推定では, まず始めに潜在変数に対する事後分布 $q(Z)$ を計算する:

$$q(Z) = \prod_{i=1}^{n}\prod_{k=1}^{K} r_{ik}^{z_{ik}}, \tag{4.19}$$

$$r_{ik} = \frac{\rho_{ik}}{\sum_{j=1}^{K}\rho_{ik}}, \tag{4.20}$$

$$\rho_{ik} = \exp\left\{\mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k(\boldsymbol{x}_i - \boldsymbol{\mu}_k)]\right\}.$$

式 (4.20) は EM アルゴリズムの枠組での潜在変数の事後確率と対応するものと見ることができる. すなわち, データ $\boldsymbol{x}_i$ がクラス C_k に属する確率を表す. 式 (4.20) を用いることで, EM アルゴリズムにおけるパラメータの推定値に類似した以下の量を計算することができる:

$$N_k = \sum_{i=1}^{N} r_{ik}, \tag{4.21}$$

$$\bar{\boldsymbol{x}}_k = \frac{1}{N_k}\sum_{i=1}^{n} r_{ik}\boldsymbol{x}_i, \tag{4.22}$$

$$\boldsymbol{S}_k = \frac{1}{N_k}\sum_{i=1}^{n} r_{ik}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^\top, \tag{4.23}$$

N_k はクラス C_k に属するデータの個数の期待値と見なせる. さらに, $\bar{\boldsymbol{x}}_k$, S_k は r_{ik} のもとでのクラス C_k の平均, 分散の期待値となる.

次に, パラメータ $\Theta = \{\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}\}$ に対する事後分布 $q(\Theta) = q(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})$ を導出する:

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}), \tag{4.24}$$

$$q(\boldsymbol{m}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{m}_k|\boldsymbol{\mu}_k, (\eta_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k), \tag{4.25}$$

ただし, $\boldsymbol{\alpha} = \{\alpha_k\}_{k\in\mathbb{N}_K}$ とし, $k \in \mathbb{N}_K$ に対して,

$$\begin{aligned}
\alpha_k &= \alpha_0 + N_k, \\
\eta_k &= \eta_0 + N_k \\
\boldsymbol{\mu}_k &= \frac{\eta_0\boldsymbol{\mu}_0 + N_k\bar{\boldsymbol{x}}_k}{\eta_k}, \\
\boldsymbol{W}_k^{-1} &= \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{\eta_0 N_k}{\eta_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_0)^\top, \\
\nu_k &= \nu_0 + N_k,
\end{aligned}$$

と計算される. 混合比 $\boldsymbol{\pi} = \{\pi_k\}_{k\in\mathbb{N}_K}$, 平均ベクトル $\boldsymbol{m} = \{\boldsymbol{m}_k\}_{k\in\mathbb{N}_K}$, 精度行列 $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}_K}$ の事後分布は, ハイパーパラメータ $\{\alpha_k, \eta_k, \boldsymbol{\mu}_k, \boldsymbol{W}_k^{-1}, \nu_k\}_{k\in\mathbb{N}_K}$ により特徴づけられるため, 変分ベイズ法による混合正規分布推定では, これらのパラメータを推定することとなる.

事後分布 (式 (4.19), 式 (4.24), 式 (4.25)) はそれぞれも各事前分布 (式 (4.18), 式 (4.14), 式 (4.15)) と同じ分布族に属していることがわかる. また, これらの式を見ると所属するデータが少ないクラス, つまり N_k が小さくなるようなクラスの事後分布の各パラメータは事前分布のパラメータに近くなることがわかる. 逆に, N_k が大きくなるようなクラスでは式 (4.21), 式 (4.22), 式 (4.23) で計算された値に重点がおかれることになる.

変分ベイズ法による混合正規分布推定では以上の計算により $\mathcal{L}(q)$ が変化しなくなるまで各パラメータの事後分布を繰り返し求めていく. ただし, $\mathcal{L}(q)$ は式 (4.5) より以下のように計算される (各項については付録 C を参照されたい):

$$\begin{aligned}
\mathcal{L}(q) &= \sum_Z \int\int\int q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) \ln\left\{\frac{p(X, Z, \boldsymbol{\pi}, \mu, \boldsymbol{\Lambda})}{q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})}\right\} \mathrm{d}\boldsymbol{\pi}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{\Lambda} \\
&= \mathbb{E}[\ln p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] \\
&= \mathbb{E}[\ln p(X|Z, \boldsymbol{m}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(Z|\boldsymbol{\pi})] + \mathbb{E}[p(Z|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{m}, \boldsymbol{\Lambda})] \\
&\quad -\mathbb{E}[\ln q(Z)] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{m}, \boldsymbol{\Lambda})].
\end{aligned} \tag{4.26}$$

変分ベイズ法による混合正規分布推定の流れは以下の通りである:

変分ベイズ法

Initialization: 各事前分布のハイパーパラメータの設定, 事後分布のハイパーパラメータの初期化.

VB-E, M Step 収束するまで以下を繰り返す.

VB-E Step: $q(Z) \leftarrow \exp\Big[\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{m},\boldsymbol{\Lambda}}[\ln p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] + \text{const}\Big]$.

VB-M Step: $q(\boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) \leftarrow \exp\Big[\mathbb{E}_{Z}[\ln p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] + \text{const}\Big]$.

4.5.3 確定的アニーリング変分ベイズ法による混合正規分布推定

ここではDAVB法を混合正規分布推定に適用したときの更新式を導出する.

更新式の導出はVB法の場合とほぼ同様であるので, ここでは更新式のみを示す. なお, DAVB法においても各パラメータの事前分布として式(4.14), 式(4.15)を用いる. また, 各記号の示すものは2.2節と同様である. 潜在変数Zの事後分布$q(Z)$は,

$$
\begin{aligned}
q(Z) &= \prod_{i=1}^{n}\prod_{k=1}^{K} r_{ik}^{z_i k}, \\
r_{ik} &= \frac{\exp[\beta\rho_{ik}]}{\sum_{j=1}^{K}\exp[\beta\rho_{ik}]}, \\
\rho_{ik} &= \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{m}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_i - \boldsymbol{m}_k)^\top \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \boldsymbol{m}_k)],
\end{aligned}
$$

となる. また, パラメータ$\Theta = \{\boldsymbol{\pi},\ \boldsymbol{m},\ \boldsymbol{\Lambda}\}$に対する事後分布, $q(\boldsymbol{\pi})$, $q(\boldsymbol{m}, \boldsymbol{\Lambda})$はそれぞれ,

$$
\begin{aligned}
q(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}|\beta(\boldsymbol{\alpha} - \mathbf{1}) + \mathbf{1}), \\
q(\boldsymbol{m}_k, \boldsymbol{\Lambda}_k) &= \mathcal{N}(\boldsymbol{m}_k|\boldsymbol{\mu}_k, (\beta\eta_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\beta\boldsymbol{W}_k, \beta(\nu_k - d - 1) + d + 1),
\end{aligned}
$$

となる. ここで,

$$
\begin{aligned}
\alpha_k &= \alpha_0 + N_k, \\
\eta_k &= \eta_0 + N_k, \\
\boldsymbol{\mu}_k &= \frac{\eta_0\boldsymbol{\mu}_0 + N_k\bar{\boldsymbol{x}}_k}{\eta_k}, \\
\boldsymbol{W}_k^{-1} &= \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{\eta_0 N_k}{\eta_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_0)^\top, \\
\nu_k &= \nu_0 + N_k,
\end{aligned}
$$

であり,

$$
\begin{aligned}
N_k &= \sum_{i=1}^{n} r_{ik}, \\
\bar{\boldsymbol{x}}_k &= \frac{1}{N_k}\sum_{i=1}^{n} r_{ik}\boldsymbol{x}_i, \\
\boldsymbol{S}_k &= \frac{1}{N_k}\sum_{i=1}^{n} r_{ik}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^\top,
\end{aligned}
$$

である. これらの式から分かるように, EM アルゴリズムの潜在変数の事後分布に相当する r_{ik} に温度パラメータが関与しているため, 他のパラメータにもその影響が伝搬する. r_{ik} を除けば各更新式は VB 法と同様の形となっている. また, 各パラメータの事後分布の引数に温度パラメータを持つため, パラメータが同じであっても温度によって異なる事後分布が得られる.

4.6 多方向探索変分ベイズ法

DAVB 法は, 主に次の 2 つの理由により良い解を得ることができると考えられる. まず 1 つ目は, DAEM アルゴリズムと同様, 原始初期点 (primitive initial point: PIP) を初期値としていることである. このアイディアの背景には, 高温状態 ($\beta \approx 0$) での解空間が, 元の ($\beta = 1$ での) 解空間の大域的構造を近似していると考えられていることがある. 2 つ目は, ある温度における良解が, 温度を少しだけ変化させた状態においても良い場所に位置していると考えられる, ということである. しかしながら, DAVB 法では様々な温度のもとで最適化を行わなけらばならず, アニーリングによる計算コストが高くなる問題がある. そこで, VB 法の局所最適性の問題に対して, DAVB 法の 1 つ目の利点を生かした手法を提案する [44]. 我々の手法では, PIP, すなわち, DAVB 法における $\beta \approx 0$ での解, を初期値とする. 更に, PIP の Hesse 行列の情報を用いた多方向探索アルゴリズムを導入する.

4.6.1 原始初期点

DAEM アルゴリズムの枠組みと同様, DAVB 法の枠組みにおいても, 高温状態 ($\beta = 0$) での解を原始初期点 (primitive initial point: PIP) と定義する. DAVB 法における PIP, $\boldsymbol{\theta}^{\mathrm{PIP}} = \{\alpha^{\mathrm{PIP}}, \eta^{\mathrm{PIP}}, \boldsymbol{\mu}^{\mathrm{PIP}}, \boldsymbol{W}^{\mathrm{PIP}}, \boldsymbol{\nu}^{\mathrm{PIP}}\}$, は前節の更新式において $\beta = 0$ を代

入することで,

$$
\begin{aligned}
\alpha^{\mathrm{PIP}} &= \alpha_0 + N^{\mathrm{PIP}}, \\
\eta^{\mathrm{PIP}} &= \eta_0 + N^{\mathrm{PIP}}, \\
\boldsymbol{\mu}^{\mathrm{PIP}} &= \frac{\eta_0 \boldsymbol{\mu}_0 + N^{\mathrm{PIP}} \bar{\boldsymbol{x}}}{\eta^{\mathrm{PIP}}}, \\
(\boldsymbol{W}^{\mathrm{PIP}})^{-1} &= \boldsymbol{W}_0^{-1} + N^{\mathrm{PIP}} \boldsymbol{S}^{\mathrm{PIP}} + \frac{\eta_0 N^{\mathrm{PIP}}}{\eta^{\mathrm{PIP}}} (\boldsymbol{x} - \boldsymbol{\mu}_0)(\boldsymbol{x} - \boldsymbol{\mu}_0)^\top, \\
\boldsymbol{\nu}^{\mathrm{PIP}} &= \nu_0 + N^{\mathrm{PIP}},
\end{aligned}
$$

と求められる. ただし,

$$
\begin{aligned}
N^{\mathrm{PIP}} &= \frac{n}{K}, \\
\bar{\boldsymbol{x}} &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i, \\
\boldsymbol{S}^{\mathrm{PIP}} &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top,
\end{aligned}
$$

である. ここで, PIP においては全ての混合要素が同一の分布となるため, 添字 k を省略した. PIP を初期値として用いることは, 我々の手法をより有効なものにすることが可能となる. 後で見るように, 負の自由エネルギーの Hesse 行列 (付録 C) は, $\beta = 1$ のとき, PIP において正負両方の固有値を持つ, すなわち目的関数は PIP において鞍点となっている. 負の自由エネルギーの増加方向に探索を行うためには, 慎重に探索方向を決定する必要がある. 我々の手法では, PIP の Hesse 行列を解析することで, PIP から正の固有値に対応する固有ベクトル方向に多方向探索を行う.

4.6.2 PIP の性質

ここでは, 10 個の実データ (表 3.1) を用いて, 混合正規分布推定における PIP の性質を実験的に考察する. 今後, 各データセットは, 表の通り #1, #2,..., #10 と表すこととする. また, 全ての実験において混合数は $K = 5$ と設定する. 各々のデータセットは, 各変数を $[-1, 1]$ の値を取るよう正規化した. まず, PIP が $\beta = 1$ のとき, 実際に鞍点となっていることを確認する. 各データセットにおいて, PIP における目的関数の Hesse 行列を計算し, その固有値を計算する. 結果を表 4.1 に示す. 表中の上付き文字 f は-重固有値を意味する: 例えば, $50^f \times 1$ は Hesse 行列が 1 つの 50 重固有値を持つことを意味する. 表を見ると, 全てのデータセットにおいて Hesse 行列が正負両方の固有値を持つことがわかる. 更に, 正の固有値の多重度を見ると全

表 4.1: 固有値の個数と多重度

Data	Negative	Positive
#1	$single \times 1,\ 4^f \times 1,\ 5^f \times 113, 50^f \times 1$	$5^f \times 11$
#2	$single \times 1,\ 4^f \times 1,\ 5^f \times 184, 65^f \times 1$	$5^f \times 15$
#3	$single \times 1,\ 4^f \times 1,\ 5^f \times 93,\ \ 45^f \times 1$	$5^f \times 10$
#4	$single \times 1,\ 4^f \times 1,\ 5^f \times 45,\ \ 30^f \times 1$	$5^f \times 7$
#5	$single \times 1,\ 4^f \times 1,\ 5^f \times 113, 50^f \times 1$	$5^f \times 11$
#6	$single \times 1,\ 4^f \times 1,\ 5^f \times 14,\ \ 15^f \times 1$	$5^f \times 5$
#7	$single \times 1,\ 4^f \times 1,\ 5^f \times 23,\ \ 20^f \times 1$	$5^f \times 5$
#8	$single \times 1,\ 4^f \times 1,\ 5^f \times 33,\ \ 25^f \times 1$	$5^f \times 6$
#9	$single \times 1,\ 4^f \times 1,\ 5^f \times 32,\ \ 25^f \times 1$	$5^f \times 7$
#10	$single \times 1,\ 4^f \times 1,\ 5^f \times 159, 60^f \times 1$	$5^f \times 13$

て 5 重の固有値となっている. これは, 混合数 K, 本実験では $K = 5$, に関連した値となっている. この結果は, $\beta = 1$ のとき, 自由エネルギー関数が PIP において実際に鞍点となっていることを示唆している. VB 法の目的は負の自由エネルギー関数の増加方向へ探索を行うことであるので, 正の固有値に対応する固有ベクトル方向に探索を行う必要がある.

次に, 各データセットにおいて, PIP からある解までの負の自由エネルギー関数の断面図, すなわち, PIP から解となるあるパラメータまでの直線上の関数値, を考察する. 実験で用いる解は, 4.8 節の計算機実験で得られた最良解を用いた. 図 4.4 に結果を示す. 10 個のデータセット中, 8 個で負の自由エネルギーが単調増加していることが見て取れる. これは, PIP から良解へ, 目的関数を単調に増加する経路が**存在する**ことを意味する (必ずしもこの経路に沿った探索を行うということではない).

これらの経験的な検証実験から, PIP は次のような性質を持つと考えられる:

(i) PIP から良解へ, 目的関数が単調増加する経路が存在する傾向にある,

(ii) $\beta = 1$ において, PIP での目的関数の Hesse 行列は正負両方の固有値を持つ, すなわち, 鞍点となっている,

(iii) 全ての正の固有値は混合数 K に関連した多重度を持つ.

性質 (i) はアニーリングを行わずとも, PIP から良解へ探索を行えることを示唆している. 図 4.4 では, 2 つのデータセットでは単調増加していないが, 他の経路を取

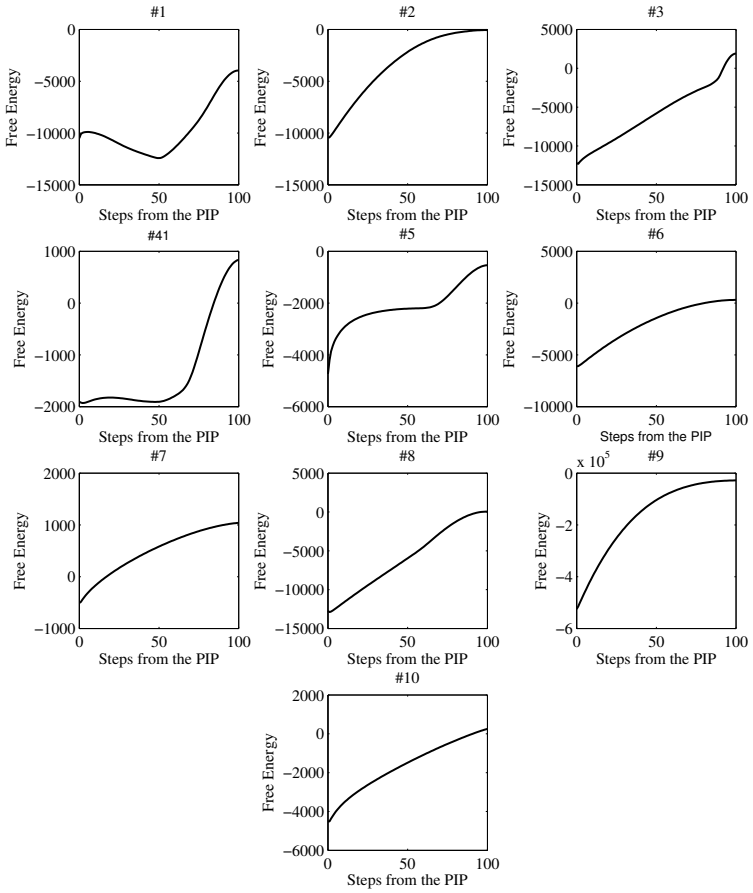


図 4.4: PIP から極値までの直線上のパラメータの自由エネルギー関数値

ることで単調増加経路が存在する可能性もある. 図 4.4 で主張したいことは, PIP から良解への単調増加経路が存在する**傾向にある**ということであり, これによって PIP が良い初期点になりうることを示したいということである. PIP が大域的最適解への単調増加経路を持つかどうか, すなわち, PIP を初期点として用いる手法が大域的最適性を持ちうるかどうかは今後更なる検証が必要である. 性質 (ii) は鞍点では, 目的関数の増加方向を慎重に決定する必要があることを示唆している. 我々の例では, 正の固有値に対応する固有ベクトル方向に探索を行うことで, そのような方向を解析的に求める. また, そのような方向を決定する際には, より大きな固有値を優先的に選ぶのが好ましい. 性質 (iii) は, K 重となっている固有値のうち, 1 つのみを用いればよいことを示唆している. 何故なら, この多重性は混合分布のクラスラベルの割り振りの冗長性を反映したものであり, 全てのラベル割

り振りの組み合わせの解を得る必要がないためである. このことを説明するために, 2 重固有値を持つ 2 混合正規分布の Hesse 行列を考え, 対応する固有ベクトルを $\boldsymbol{u}_1 = [\Theta_1^\top,\ \Theta_2^\top]$, $\boldsymbol{u}_2 = [\Theta_2^\top,\ \Theta_1^\top]$ とする. ここで, $\Theta_1,\ \Theta_2$ はパラメータベクトルであり, それぞれ混合正規分布の 1 つ目と 2 つ目の混合要素のパラメータとする. もし, $\boldsymbol{u}_1, \boldsymbol{u}_2$ の両方向に探索を行った場合, 結果として得られる解は, クラスラベルが入れ替わったもの, $[\Theta_1^{*\top},\ \Theta_2^{*\top}]$, $[\Theta_2^{*\top},\ \Theta_1^{*\top}]$, である. このような冗長性を避け, 計算コストを削減するため, 我々は多重の固有値に対しては, そのうちの 1 つのみを選択することとする. ただし, 多重となっている固有空間において探索を行った場合に, 必ずしもこのような冗長な解が得られるわけではないことに注意されたい.

4.6.3 多方向探索変分ベイズ法

前述した性質を利用し, PIP を初期点とする多方向探索変分ベイズ法を提案する. 我々のアプローチでは, まず初めに PIP, $\boldsymbol{\theta}^{\mathrm{PIP}} = \{\alpha^{\mathrm{PIP}},\ \eta^{\mathrm{PIP}},\ \boldsymbol{\mu}^{\mathrm{PIP}},\ \boldsymbol{W}^{\mathrm{PIP}},\ \boldsymbol{\nu}^{\mathrm{PIP}}\}$ を計算し, PIP における Hesse 行列, $\partial^2\mathcal{L}/\partial\Theta\partial\Theta^\top\big|_{\Theta=\boldsymbol{\theta}^{\mathrm{PIP}}}$, およびその固有値, 固有ベクトルを求める. 次に, 累積寄与率が 80%となるような正の固有値を, 値の大きいものから順に選択する. このような選択方法は, できるだけ大きな固有値を選択するためのヒューリスティックとして導入した. その後, 選択した固有値に対応する固有ベクトルを用いて, 探索方向, $\{\Delta\boldsymbol{\theta}_r\}_{r\in\mathbb{N}_{2R+2^R}}$ を生成する. 探索方向 $\{\Delta\boldsymbol{\theta}_r\}_{r\in\mathbb{N}_{2R+2^R}}$ の生成方法およびその理論的な解釈は EM アルゴリズムにおける多方向探索法と同様である. 最後に, PIP から $\Delta\boldsymbol{\theta}_r$ 方向に微小な摂動を加えたパラメータを VB 法の初期値として生成する: 例えば, $\boldsymbol{\theta}_r^{\mathrm{init}} = \boldsymbol{\theta}^{\mathrm{PIP}} + \Delta\boldsymbol{\theta}_r,\ r \in \mathbb{N}_{2R+2^R}$. それら $2R + 2^R$ 個の初期値から VB 法によるパラメータ推定を行う.

この手法を VB-PIP 法と呼ぶこととする. VB-PIP 法の流れは以下の通りである:

VB-PIP 法

入力. 観測データ $X = \{\boldsymbol{x}_i\}_{i\in\mathbb{N}_n}$, 事前分布のハイパーパラメータ, 累積寄与率.

Step 1. 事前分布のハイパーパラメータの初期化を行う.

Step 2. PIP, $\boldsymbol{\theta}^{\mathrm{PIP}} = \{\alpha^{\mathrm{PIP}},\ \eta^{\mathrm{PIP}},\ \boldsymbol{\mu}^{\mathrm{PIP}},\ \boldsymbol{W}^{\mathrm{PIP}},\ \boldsymbol{\nu}^{\mathrm{PIP}}\}$ を計算する.

Step 3. PIP における負の自由エネルギー関数 (式 (4.26)) の Hesse 行列を計算し, 固有値分解を行う.

Step 4. 累積寄与率 80%となる正の固有値を選択する.

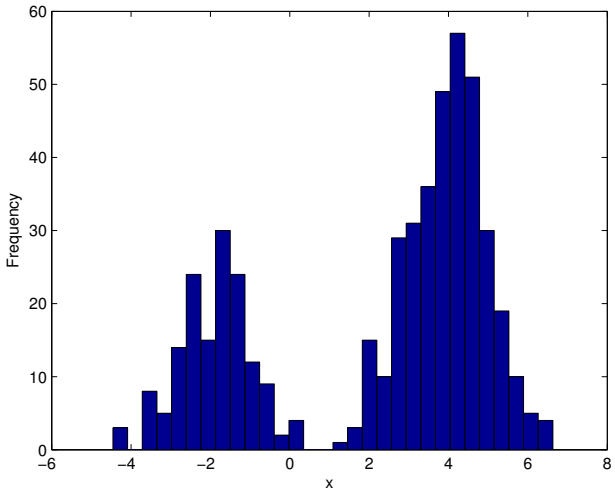


図 4.5: データのヒストグラム

Step 5. Step 4 で選択した固有値に対応する固有ベクトルを用いて, 探索方向 $\{\Delta\boldsymbol{\theta}_r\}_{r\in\mathbb{N}_{2R+2^R}}$ を生成する.

Step 6. 初期値 $\{\boldsymbol{\theta}_r^{\text{init}}\}_{r\in\mathbb{N}_{2R+2^R}} = \{\boldsymbol{\theta}^{\text{PIP}} + \Delta\boldsymbol{\theta}_r\}_{r\in\mathbb{N}_{2R+2^R}}$ を求める.

Step 7. Step 6 で生成した初期値を用いて VB 法によりパラメータを推定する.

出力. 目的関数が最大となるパラメータ $\boldsymbol{\theta}^*$.

4.7 変分ベイズ法の枠組みでの混合正規分布推定のシミュレーション

ここでは, 人工データを用いて, VB 法, DAVB 法, VB-PIP 法のシミュレーションを行う. 実験に用いるデータは, 1 次元 2 混合正規分布によって生成される人工データ 500 個である:

$$p(X|\boldsymbol{\theta}) = 0.3\mathcal{N}(-2,1) + 0.7\mathcal{N}(4,1).$$

混合比 π_k, 平均ベクトル $\boldsymbol{m}_k$ および精度行列 $\boldsymbol{\Lambda}_k$ の事前分布,

$$\begin{aligned} p(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \mathrm{C}(\boldsymbol{\alpha}_0)\prod_{\mathrm{k}=1}^{\mathrm{K}}\pi_{\mathrm{k}}^{\alpha_0-1}, \\ p(\boldsymbol{m},\boldsymbol{\Lambda}) &= \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{m}_k|\mathbf{m}_0,(\eta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0), \end{aligned}$$

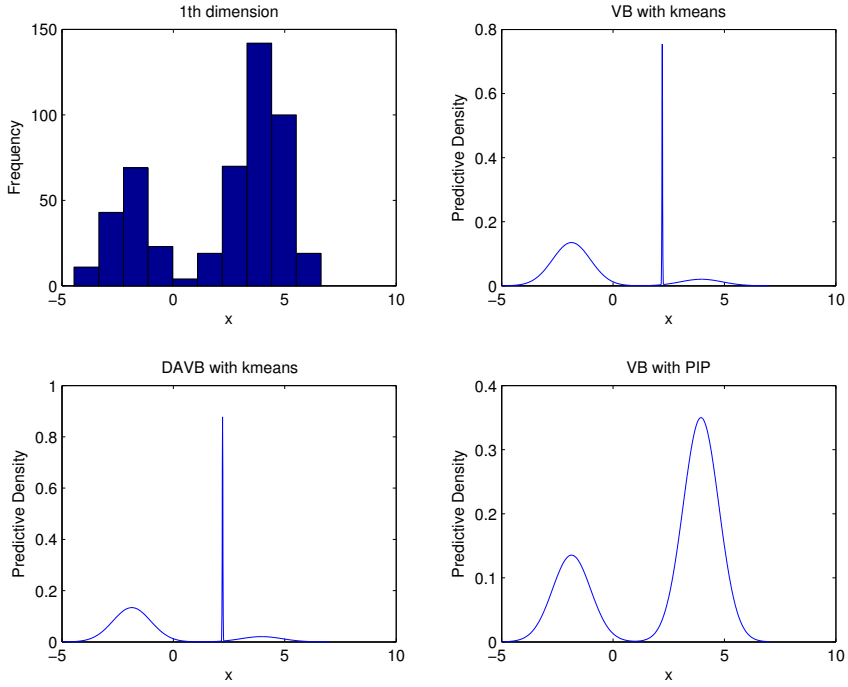


図 4.6: 最良解での予測分布

のハイパーパラメータは,

$$\begin{aligned}\alpha_0 &= 0.01,\\ \eta_0 &= 1,\\ \boldsymbol{\mu}_0 &= \frac{1}{N}\sum_{i=1}^{n}\boldsymbol{x}_n,\\ \mathbf{W}_0 &= 20\boldsymbol{I},\\ \nu_0 &= 50.\end{aligned}$$

と設定した. 次に, 各アルゴリズムで得られた最良解での予測分布を見てみる (図 4.6). データ $\hat{\boldsymbol{x}}$ に対する予測分布は以下の式により与えられる:

$$\begin{aligned}p(\hat{\boldsymbol{x}}|X) &= \frac{1}{\hat{\alpha}}\sum_{k=1}^{K}\alpha_k \mathrm{St}(\hat{\boldsymbol{x}}|\boldsymbol{m}_k, \boldsymbol{L}_k, \nu_k+1-p),\\ \boldsymbol{L}_k &= \frac{(\nu_k+1-p)\eta_k}{1+\eta_k}\boldsymbol{W}_k.\end{aligned}$$

ここで, St は多変量 Student t-分布で以下で与えられる:

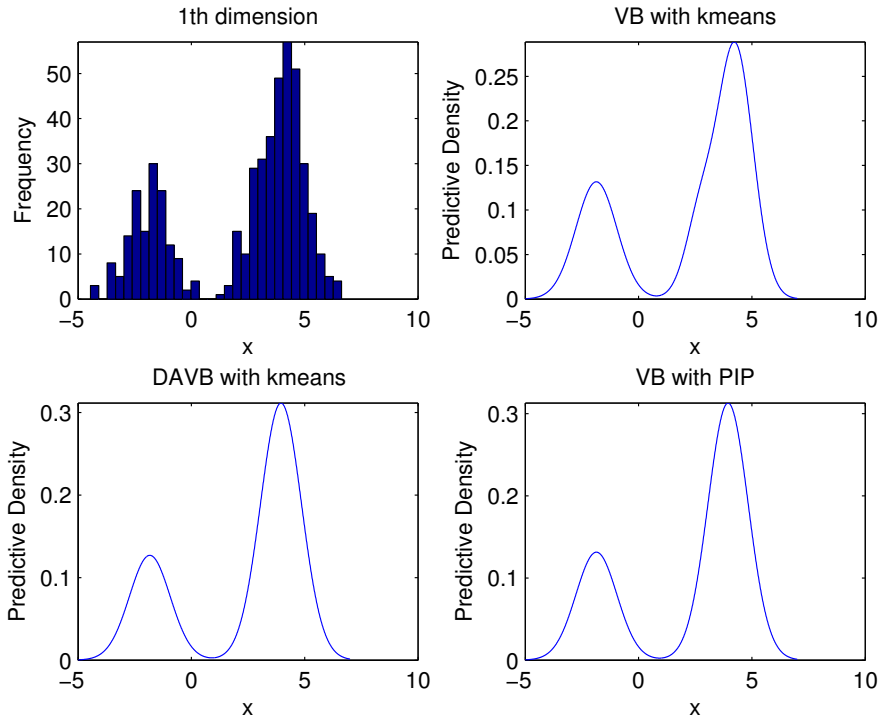


図 4.7: $\boldsymbol{W}_0 = 0.05 \times \boldsymbol{I}$ での結果: 尤度はそれぞれ, VB(kmeans) 1086, DAVB 1051, VB(PIP) 1059.

$$\begin{aligned} \mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &= \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{(\nu\pi)^{\frac{p}{2}}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\frac{\nu+p}{2}}, \\ \Delta^2 &= (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu}). \end{aligned}$$

図を見ると, VB(kmeans) および DAVB は局所最適解へ収束しているように思われる. これらの局所最適性の問題が起こる仕組みは, EM アルゴリズムにおける解空間での探索点の挙動と同様である考えられる. しかし, VB 法の枠組みでは, 事前分布の設定を変更することで, すなわち問題空間に変更を加えることで解決可能な場合もある. この例では, 事前分布付近に存在するデータに, 分布がオーバーフィットしたために, 図のようなピークが現れる. 従って, このような場合は事前分布を変更することで対処可能である. 次に, Wishart 分布のハイパーパラメータである $\boldsymbol{W}_0$ を変更してみる.

今までの設定では $\boldsymbol{W}_0 = 20 \times \boldsymbol{I}$ としていた. 分散共分散行列としては, これの逆行列を考えるのでこのような設定のもとでは分散は極めて小さくなる. これが

ピークの原因の一つと考えられる. まず, $\boldsymbol{W}_0 = 0.05 \times \boldsymbol{I}$ と設定した結果を図 4.7 に示す. 図から分かるように先ほど見られたようなピークはなくなり, 全て同じ予測分布が得られている. これらの例では事前分布を変更することで VB(kmeans) 法や DAVB 法で見られる局所解を避けることができた. しかしながら, 今回の 2 通りの設定で同じ解が得られた VB-PIP 法は, 他の 2 手法よりも局所最適解に陥りにくいことを示しており, VB-PIP 法の有効性を示唆している.

4.8 計算機実験

ここでは, 表 3.1 のデータセットを用いて, VB-PIP 法の性能を評価する. 比較対象は, k-means 法により初期値を生成する方法 (VB(k-means)), DAVB 法である.

本実験では, 3 通りの設定で性能を比較する: 実験 I では初期値の数を統一し, 実験 II では計算時間を, DAVB 法により 1 度推定する計算時間に統一したものである. 事前分布のハイパーパラメータは $\{\alpha_0,\ \eta_0,\ \boldsymbol{\mu}_0,\ \boldsymbol{W}_0,\ \nu_0\} = \{1, 1, \bar{\boldsymbol{x}}, 10 \times I, 50\}$ と設定した. なお $\bar{\boldsymbol{x}}$ はデータ X の標本平均である. DAVB 法における温度スケジューリングは $\beta^{(0)} = 0.1,\ \beta^{(t+1)} = 1.2 \times \beta^{(t)}$ とした.

4.8.1 実験 I

まず, 初期値の数を統一して実験を行う. 初期値数は VB-PIP 法で生成される個数とした. すなわち, もし VB-PIP 法により I 個の初期値が生成された場合, VB 法, DAVB 法の初期値も I 個とする. 結果を表 4.2 に示す. 解の良さは負の自由エネルギーの関数値で評価する. すなわち, 大きい値ほど良い. また, 表中において, “Init.” は初期値生成に要した計算時間, “Est.” はパラメータの推定に要した時間, “Total(=“Init.”+“Est.”)” は総計算時間を表す. さらに, “Initial Points” は初期値の数を表す.

VB-PIP 法は 10 個のデータセット中 5 個で最良の解を, 3 個のデータセットで他手法と同様の解を得ている. また, 計算コストについては, k-means 法により初期値を生成する, 標準的な VB 法と同等である. 何故なら, VB-PIP 法は初期値生成後は通常の VB 法によりパラメータ推定を行うためである. また, DAVB 法と比較すると, VB-PIP 法の計算コストは非常に小さくなっている. これは, DAVB 法ではアニーリングにより, 様々な温度の下で最適化を行わなければならないためである. VB-PIP 法では, 探索方向を生成するために余分な計算時間がかかる. これは主に Hesse 行列の計算がボトルネックになっているためであるが, 付録にあるように混

合正規分布推定では, PIP における目的関数の Hesse 行列がスパースとなっているため, それほど大きなコストが必要なわけではないことに注意されたい.

4.8.2 実験 II

DAVB 法は, アニーリングプロセスにおいてしばしば鞍点, 特に PIP, にトラップされる. 鞍点から抜け出すためには, パラメータにランダムな摂動を加える必要がある. 表 4.2 の結果は, DAVB 法の解の良さはランダムな摂動に依存していることを示唆している. アニーリングプロセスにおいて, 探索点が鞍点にトラップされた場合, ランダムな摂動によって, 他手法と同程度の解のばらつきが見られる: #1, #2, #3, #5, #6, #10. 一方で, 鞍点にトラップされなかった場合は, DAVB 法が得られる解にほとんどばらつきが見られない, すなわち, 何度 DAVB 法により推定を行っても同じような解が得られている. そのような場合は, DAVB 法は 1 度推定を行えば十分であると考えられる. このことを考慮し, ここでは, DAVB 法により 1 度推定を行うために要した計算時間に時間を統一し, 3 手法の性能を比較する. 結果を 4.3 に示す.

この設定においても, ほとんどのデータセットにおいて VB-PIP 法が最も良い性能を示していることがわかる. 一方で, DAVB 法は他手法と比較して解が悪い. 経験的に, 鞍点における DAVB のランダムな挙動は解の質に大きな影響を及ぼすことがわかっている.

これらの結果から, VB-PIP 法は解の質, 計算コストともに他手法と同等かそれ以上であり, 更に, k-means 法などと異なり確定的に初期値を生成できることから, 混合正規分布推定において有効であると考えられる.

4.9 まとめ

本章では, 変分ベイズ法による混合正規分布推定を取り扱った. まず, ベイズ推定による正規分布推定を概略し, 分布のパラメータに対し事前分布を与え, 尤度関数との積を考えることにより事後分布を導出することをみた. 次に変分ベイズ法および確定的アニーリング変分ベイズ法について説明し, 混合正規分布推定における局所最適性について考察を行った. このような問題を改善するために, 前章で導入した, 原始初期点とそこでの目的関数の Hesse 行列を利用した多方向探索アルゴリズムを提案した. 実データを用いた計算機実験により, 本手法が k-means 法による変分ベイズ法および DAVB 法と同等あるいはそれ以上の性能を示すことが確

認された.

多方向探索アルゴリズムにおける今後の課題として, 原始初期点の性質の更なる検証と, その有効性の理論的裏付けなどが挙げられる.

表 4.2: 性能比較 I: 初期値の個数を統一した場合

Data	Method	Best	Average ± S.E.	Computational Time			Initial Points
				Init.	Est.	Total	
#1	VB(kmeans)	-4840	-6912 ± 83.58	0.71	174.94	175.65	142
	DAVB	**-3961**	-6094 ± 59.64	0.72	5238.84	5239.55	142
	VB-PIP	-4049	-5695 ± 69.71	12.42	415.34	427.76	142
#2	VB(kmeans)	143	-184 ± 18.04	0.30	185.94	186.24	142
	DAVB	-65	-119 ± 10.33	0.30	3373.44	3373.74	142
	VB-PIP	**148**	-235 ± 29.42	24.68	226.84	251.52	142
#3	VB(kmeans)	**2342**	1909 ± 53.06	0.21	94.90	95.11	42
	DAVB	1903	1755 ± 6.10	0.21	3578.75	3578.96	42
	VB-PIP	2289	1863 ± 224.88	4.97	100.27	105.24	42
#4	VB(kmeans)	**834**	377 ± 20.87	0.35	313.50	313.85	76
	DAVB	**834**	834 ± 0.00	0.35	9467.25	9467.60	76
	VB-PIP	**834**	730 ± 20.64	2.49	319.70	322.19	76
#5	VB(kmeans)	-575	-1741 ± 89.44	0.12	49.04	49.16	76
	DAVB	-537	-1242 ± 59.96	0.12	748.41	748.53	76
	VB-PIP	**-447**	-730 ± 33.83	6.36	70.46	76.82	76
#6	VB(kmeans)	313	290 ± 5.92	0.02	7.56	7.58	14
	DAVB	301	287 ± 13.97	0.02	135.11	135.13	14
	VB-PIP	**319**	274 ± 19.06	0.06	6.59	6.65	14
#7	VB(kmeans)	1088	1085 ± 2.49	0.05	45.26	45.31	24
	DAVB	1039	1039 ± 0.00	0.05	1354.63	1354.68	24
	VB-PIP	**1092**	1062 ± 8.62	0.39	64.19	64.58	24
#8	VB(kmeans)	42	-219 ± 32.88	0.05	26.22	26.27	24
	DAVB	42	42 ± 0.00	0.06	631.93	631.99	24
	VB-PIP	**54**	-27 ± 31.07	0.42	32.49	32.91	24
#9	VB(kmeans)	**-28032**	-28111 ± 16.36	0.22	304.03	304.25	14
	DAVB	-28139	-28140 ± 0.08	0.24	1164.27	1164.51	14
	VB-PIP	**-28032**	-28099 ± 16.22	1.96	239.24	241.20	14
#10	VB(kmeans)	**288**	52 ± 10.45	0.30	162.15	162.45	272
	DAVB	252	-12 ± 9.81	0.30	1323.17	1323.47	272
	VB-PIP	**288**	-13 ± 10.06	17.66	159.19	176.85	272

表 4.3: 性能比較 II: DAVB1 回の推定にかかる計算時間に統一

Data	Method	Best	Average ± S.E.	Computational Time			Initial Points
				Init.	Est.	Total	
#1	VB(kmeans)	-5304	-7192 ± 170.47	0.17	42.42	42.58	33
	DAVB	-6916	-6916 ± 0.00	0.01	42.78	42.79	1
	VB-PIP	**-4935**	-5821 ± 262.96	12.42	28.30	40.72	8
#2	VB(kmeans)	**107**	-166 ± 48.56	0.04	27.25	27.29	21
	DAVB	-93	-93 ± 0.00	0.00	27.37	27.37	1
	VB-PIP	66	30 ± 36.26	24.68	1.93	26.61	2
#3	VB(kmeans)	**2333**	1901 ± 61.45	0.16	78.07	78.23	33
	DAVB	1744	1744 ± 0.00	0.01	78.90	78.91	1
	VB-PIP	2289	1794 ± 314.96	4.97	70.13	75.10	30
#4	VB(kmeans)	**834**	377 ± 34.85	0.13	125.86	125.99	29
	DAVB	**834**	834 ± 0.00	0.01	128.30	128.31	1
	VB-PIP	**834**	798 ± 16.67	2.49	125.77	128.26	32
#5	VB(kmeans)	-1005	-1861 ± 190.70	0.01	5.99	6.00	9
	DAVB	-537	-537 ± 0.00	0.00	7.68	7.68	1
	VB-PIP	**-447**	-447 ± 0.00	6.36	0.82	7.18	1
#6	VB(kmeans)	313	290 ± 5.92	0.02	7.56	7.58	14
	DAVB	301	301 ± 0.00	0.00	8.97	8.97	1
	VB-PIP	**319**	274 ± 19.06	0.06	6.59	6.65	14
#7	VB(kmeans)	1088	1085 ± 2.49	0.05	45.26	45.31	24
	DAVB	1039	1039 ± 0.00	0.00	55.45	55.45	1
	VB-PIP	**1092**	1057 ± 9.49	0.39	55.01	55.40	21
#8	VB(kmeans)	42	-219 ± 32.88	0.05	26.22	26.27	24
	DAVB	42	42 ± 0.00	0.00	26.43	26.43	1
	VB-PIP	**54**	-46 ± 38.33	0.42	25.69	26.11	19
#9	VB(kmeans)	-28147	-28153 ± 5.21	0.05	76.81	76.86	3
	DAVB	-28139	-28139 ± 0.00	0.03	82.10	82.13	1
	VB-PIP	**-28032**	-28126 ± 31.71	1.96	74.16	76.12	4
#10	VB(kmeans)	**273**	5 ± 42.90	0.01	5.00	5.01	11
	DAVB	-111	-111 ± 0.00	0.00	5.07	5.07	1
	VB-PIP	-91	-91 ± 0.00	17.66	0.46	18.12*	1

第 5 章

マイクロアレイデータ解析

近年, マイクロアレイ技術の発達により, 膨大な量の遺伝子データが得られるようになった. マイクロアレイデータは, データ数 n が比較的小さく, 次元数 (遺伝子数) p が非常に大きいという特徴を持っている. そのようなデータから, 目的に応じた有用な知見を得るために様々な解析手法が研究, 提案されている. 本章では, マイクロアレイの背景となる分子生物学およびマイクロアレイデータについて触れたあと, 2 種類のタスク, 遺伝子群解析とゲノム異常領域同定について説明を行う. また, これらのタスクを行う上で重要となる統計的検定について述べる.

5.1 DNA とセントラルドグマ

ヒトや動物, 植物に関して深く理解するためには, その構成要素を理解することが重要である. これらの生物は細胞と呼ばれる基本単位により構成されており, 各細胞には個体の青写真であるデオキシリボ核酸 (deoxyribose nucleic acid: DNA) が含まれている. DNA 分子は, 染色体と呼ばれる構造の中に存在している. DNA 分子は 4 つの異なるヌクレオチド, アデニン (adenine: A), チミン (thymine: T), グアニン (guanine: G), シトシン (cytosine: C) から成る. これらの 4 つの塩基は, A は T と, G は C と結合しペアを作る. この A, T, G, C の配列が生物の遺伝情報を担っている. A, T, G, C の配列は, 染色体において 2 重螺旋構造を取っており, A-T, G-C のペアを構成するよう互いに相補的な配列となっている [48]. 生命の最も基本的な構成要素と考えられている遺伝子 (gene) は, この DNA の一部分にコード化されている. 各遺伝子は, 生成するタンパク質の量を決定するためのプロモータを含んでおり, 遺伝子のコード領域は, どのタンパク質を合成するかを決める. 染色体もしくは遺伝子全体をゲノム (genome[1]) と呼ぶ. 3 つの塩基 (コドン) からなるコードは, アミ

[1]接尾語-*ome* は “全体” あるいは “群” を意味する単語であり, この場合 *gene-ome*, すなわち遺伝子全体を意味している. 生物学では, 類似の単語としてタンパク質全体を指す *proteome(=proteine-ome)*,

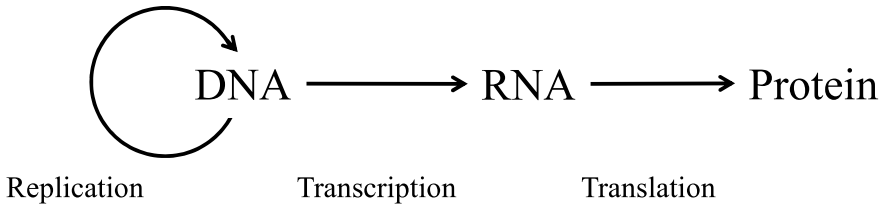


図 5.1: 分子生物学におけるセントラルドグマ

ノ酸の種類およびコード領域の始点, 終点を定義する. DNA からタンパク質が合成される過程で, DNA は一度, リボ核酸 (rebo nucleic acid: RNA) へ転写される. この際, チミン (T) はウラシル (uracil: U) に置き換えられる. DNA から RNA, RNA からタンパク質へと遺伝情報の形が変化していくプロセスをセントラルドグマと呼ぶ (図 5.1).

セントラルドグマにおいては, 主として 3 つ (多くの場合, これに RNA から DNA への逆転写, reverse transcription を含める) の過程からなる:

複製 (Replication): 細胞分裂の際などに DNA のコピーを作るプロセス,

転写 (Transcription): DNA コードをメッセンジャー RNA(messenger RNA: mRNA) へ転写するプロセス,

翻訳 (Translation): DNA 情報がタンパク質へ翻訳されるプロセス (mRNA を介す).

生物は, このように DNA から RNA, タンパク質へと遺伝情報を発現することで, 生きていく上で必要な物質, 機能を維持している. 従って, これらの DNA や RNA, タンパク質の量を解析することで, 生命 (あるいは逆に病気など) に関する重要な知見を得ることができる.

本章で取り扱う 2 種のデータ, 遺伝子発現量データと array CGH データは, 前者が mRNA 量を測定するのに対し, 後者は DNA コピー数を測定する点で異なる.

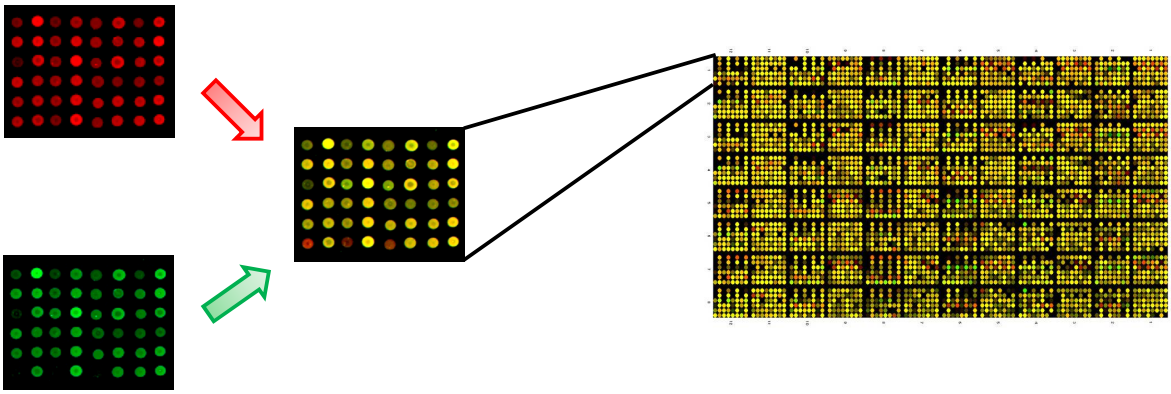


図 5.2: マイクロアレイ技術: 例えば, 健常者の RNA を緑の蛍光色素で着色, 癌患者の RNA を赤の蛍光色素で着色しハイブリダイズする. 得られた画像データ (右) を蛍光スキャナで数値化することでマイクロアレイデータが作成される.

5.2 マイクロアレイ技術とマイクロアレイデータ

本節では, DNA コピー数や mRNA 量を定量化する技術である, マイクロアレイ技術 [17] について説明する. マイクロアレイ技術は, DNA(あるいは RNA) のハイブリダイゼーション (hybridization) を基本原理としている. ハイブリダイゼーションとは, 1 本の DNA が 2 本, あるいは 1 本の DNA と 1 本の RNA が互いに相補的な塩基同士で結合し 2 本の鎖を作ることである. 例えば, mRNA が含まれる試料の中に, 蛍光標識をした DNA の小さな断片 (オリゴヌクレオチド) を添加すれば, オリゴヌクレオチドと相補的な配列を持つ特定の mRNA がハイブリダイズする. すると, ハイブリダイズした mRNA のみが色付けされ蛍光スキャナで検出することができる. 逆に, 調べたいいくつかのオリゴヌクレオチドのプローブ (検出子) を平面上に固定し, そこに試料を流し込めば, 試料に各々のプローブが存在するかどうかおよびその量 (発現量) を検出することができる (図 5.2).

- 遺伝子発現量データ
  遺伝子発現量データは, 前述したような方法で mRNA の濃度 (発現量) を定量化したデータである. mRNA は DNA の遺伝子情報が転写された物質であり, mRNA 量を調べることで, 各 mRNA に対応する遺伝子発現を解析することができる. 例えば, 実験群を緑色で, 対照群を赤色で蛍光標識した場合, 実験群

mRNA 総体を指す *transcriptome(=transcript-ome)* などがある.

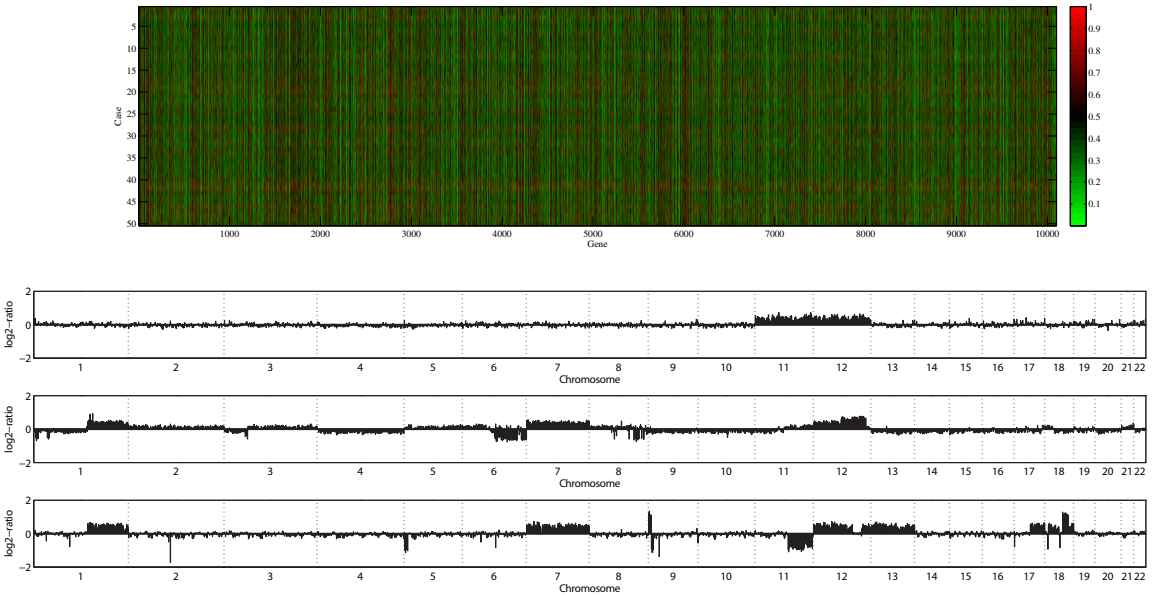


図 5.3: 遺伝子発現量データ 50 症例 (上) と array CGH データ 3 症例 (下): 遺伝子発現量データは空間的な相関はない. 一方, array CGH は隣り合うプローブは相関を持ち, 連続して値が大きく, あるいは小さくなる.

に対して, 対照群の mRNA 発現量が大きければ, そのスポットは赤色を示し, 逆に小さければ緑色を示す. また, 両者の発現量に差がない場合は黄色 (場合によっては黒色) を示す. 遺伝子発現量データ解析により, 特定の機能に必要な遺伝子の特定や, 病気の原因となる遺伝子の特定などを行うことができ, 疾患の予防や治療などに役立つと考えられている.

- array CGH データ
  array CGH とは, 全ゲノムの DNA コピー数を網羅的に測定するための技術である [49]. array CGH マイクロアレイを作成する際, Bacterial artificial chromosome(BAC) クローンや相補的 DNA(complementary DNA: cDNA, オリゴヌクレオチドなどを DNA マイクロアレイへハイブリダイズする方法がある [2] . array CGH データは各プローブにおける実験群の DNA コピー数と対照群のコピー数の $\log_2$-ratio として定量化される. 遺伝子発現量データとは異なり, DNA コピー数異常に起因する疾患の解析などに用いられ, 解析方法も異なる.

図 5.3 にこれらのデータの例を示す. 図上が遺伝子発現量データの例であり, 下が array CGH データ [50, 51, 52] の例である.

[2] 本章で用いる array CGH データは BAC array CGH データである

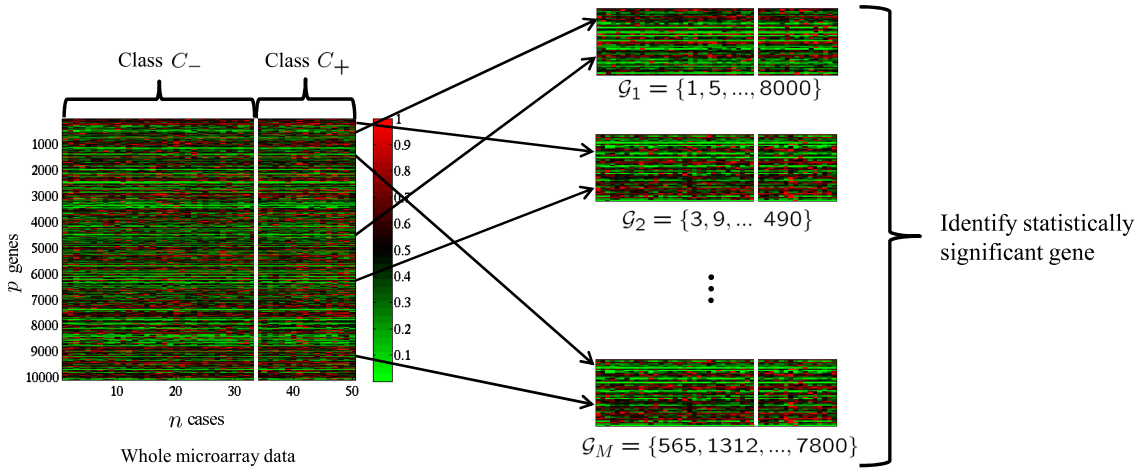


図 5.4: 遺伝子群解析の概要

5.3 遺伝子発現量データと遺伝子群解析

マイクロアレイデータ解析の基本的なタスクは表現型 (例えば, 薬剤効果の有無など) によって発現量の異なる遺伝子を同定することである. 一方で, 多くの遺伝子は他の遺伝子と影響を及ぼしあうため, 関連のある遺伝子をまとめて解析することが有用であると考えられている. このような, 関連のある遺伝子をグループ化したものを遺伝子群 (gene set) と呼び, 発現差のある遺伝子群を同定する問題は遺伝子群解析 (gene set analysis) と呼ばれる. 遺伝子群とは, 生物学的に関連性のある遺伝子をまとめたものであり, 例えば, 特定の疾患に関連する遺伝子の集合, 特定の薬剤に反応する遺伝子の集合などである. また, マイクロアレイデータを遺伝子群に基づいて解析することで, 医学生物学的な解釈が容易となる. 図 5.4 に遺伝子群解析の概念図を示す. 図のヒートマップでは, 横軸方向に n 症例, 縦軸方向に p 遺伝子が並んでおり, 遺伝子発現量の大きいセルが赤色, 小さいものが緑色で表現されている. また, ヒートマップ中の点線の左側が陰性クラス C_-, 右側が陽性クラス C_+ の症例となっている. $\mathcal{G}_\cdot$ は遺伝子群を表しており, いくつかの遺伝子群の中で, 2 クラスの発現パターンの違いを定量化することが遺伝子群解析の目的となる. 遺伝子群解析の研究として有名なものに Gene Set Enrichment Analysis (GSEA) [28] と呼ばれるものがある. GSEA データベースには, 医学生物学的な知見から得られた遺伝子群が多数定義されている.

同一の遺伝子群に含まれる遺伝子の発現量は相互相関を持つことが多いため, 遺伝子群の発現パターンを多変量データと解釈して解析することが有効である.

この解釈のもと, 遺伝子群解析は多重多変量 2 標本検定問題として定式化される. 遺伝子群解析のための検定統計量は遺伝子の相互相関を適切に説明できるものであることが望ましい. GSEA では Enrichment スコアと呼ばれる統計量が用いられる. Enrichment スコアは個々の遺伝子に関する単変量統計量を組み合わせたものであるため, 遺伝子発現量の相互相関を表現することができない. 変数間の相関関係を記述できる多変量 2 標本検定として, Hotelling T^2 検定 [53] と呼ばれるものがある. 2 標本が分散共分散行列の等しい多変量正規分布に従うときには, Hotelling T^2 検定は最強力検定となることが知られている. しかしながら, 遺伝子群の発現パターンの分布は未知であるため, ノンパラメトリックな多変量 2 標本検定を導入する必要がある. ノンパラメトリックなものとしては, runs test [54], nearest-neighbor test [55, 56] などが提案されている. また, 複数の遺伝子群から統計的に有意なものを同定するタスクでは, 検定の多重性を考慮する必要がある. このように, 遺伝子群解析は多重多変量 2 標本検定として定式化される.

5.4 array CGH データとゲノム異常領域同定

Array Comparative Genomic Hybridization (array CGH) はゲノムスケールでの DNA コピー数の測定に有効な技術である. 癌細胞における array CGH 解析では, 多数の BAC, cDNA, オリゴヌクレオチドなどのゲノムクローンのマイクロアレイに, 癌細胞から得た DNA と正常細胞 (あるいは参照細胞) から得た DNA をコハイブリダイズすることによって array CGH データを作成する. array CGH は多数のプローブの各々について, ゲノム上のプローブに対応する領域での癌細胞の DNA コピー数と正常 (参照) 細胞の DNA コピー数の $\log_2$-ratio を与える. $\log_2$-ratio が正の大きな値を取る場合, 癌細胞はプローブに対応する領域において DNA コピー数の増幅を表し, 負で絶対値の大きな値を取る場合は欠損を表す. 図 5.3 下はリンパ腫に疾患した 3 症例から得た array CGH データの例である. 図において, 横軸は染色体番号, $1, \ldots, 22$, を表し, 縦軸は各々の領域のプローブの $\log_2$-ratio を表す. 図の array CGH データでは, ゲノム全体を 2,035 の BAC プローブで表している.

array CGH データ解析の, 最初の基本的なタスクは, 1 つのアレイ (1 人の患者) に見られるゲノム異常領域を探すことである. このタスクに対しては, 多くの計算アルゴリズムが提案されている [33, 36, 37, 57]. 次のタスクは, あるグループ, 例えば同じ癌に疾患した数人の患者など, において共通して見られる異常領域を同定する問題である. 最初のタスクと比べ, このタスクに対する研究はそれほど多くは

なされていない [33]. しかしながら, 生物学的な観点からはそのような共通異常領域を検出することが, より重要である. 現状では, このような共通異常領域の同定問題に対しては, 形式的な方法が確立されていない. そのようなアプローチの1つは, 各々のプローブにおいて, グループの $\log_2$-ratio を平均化し, 最初のタスクの (1つのアレイに対する) アルゴリズムを, その平均化された $\log_2$-ratio に対して適用するというものである. 次いで, 3つ目のタスクは2グループ, あるいはそれ以上のグループにおいて, あるグループでは共通して異常が見られるが, 他のグループでは異常が見られない, もしくは異なる異常パターンを示すような領域を同定する問題である. このような特徴の異なる異常領域を同定する問題は, 生物学的な研究を進める上で重要であると考えられるが, 我々の知る限り, このようなタスクに対処する方法は今のところない. 現在では, 第2, 第3のタスクに対する, 体系化された解析方法が必要とされている.

array CGH解析における計算アルゴリズムを確立する際には, プローブ間の空間的な相関を考慮することが重要である. 遺伝子発現量データとは異なり, CGHマイクロアレイのプローブはDNAのある断片を表しており, 各染色体において, 隣り合うプローブは系列的につながりを持っている. 例えば, 図5.3では, まとまった増幅, 欠損が見られる. このような理由から, 1つ目のタスクに対する手法は系列データのセグメント化, ブレークポイントを検出するようなアルゴリズムとなっている. 一方で, 2, 3のタスクでは, array CGHデータを症例数を次元数とみなした多変量シーケンシャルデータとして取り扱わなければならず, 空間的な相関を考慮することが難しくなる.

グループ間で特徴の異なる異常領域を同定する3つ目のタスクでは, 特徴の違いを定量化し, その統計的検定を行うことが望ましい. 発現量データ解析では, このような統計的信頼性と多重検定に関する研究が進められている [58, 59]. もし, 空間的な相関関係を無視すれば, 発現量データにおいて研究されている方法論をarray CGHデータ解析へ適用することができる. しかしながら, 相関関係を考慮し, 連続する複数プローブからなる領域に関心がある場合, そのような方法ではいくつかの問題がある.

まず初めに, 複数プローブからなる領域に対して統計的検定を行う際には, 単変量検定ではなく, 多変量検定を採用すべきである. 2つ目の問題は, 染色体22本の各々について, 全ての可能な部分領域を調べる必要があり, その数は非常に大きくなることである. 例えば, 図5.3 下は, 2,035プローブからなるBAC arrayであり, 全ての可能な (異常領域の) 候補領域は, 141,452領域となる. これは, 多重検定補正

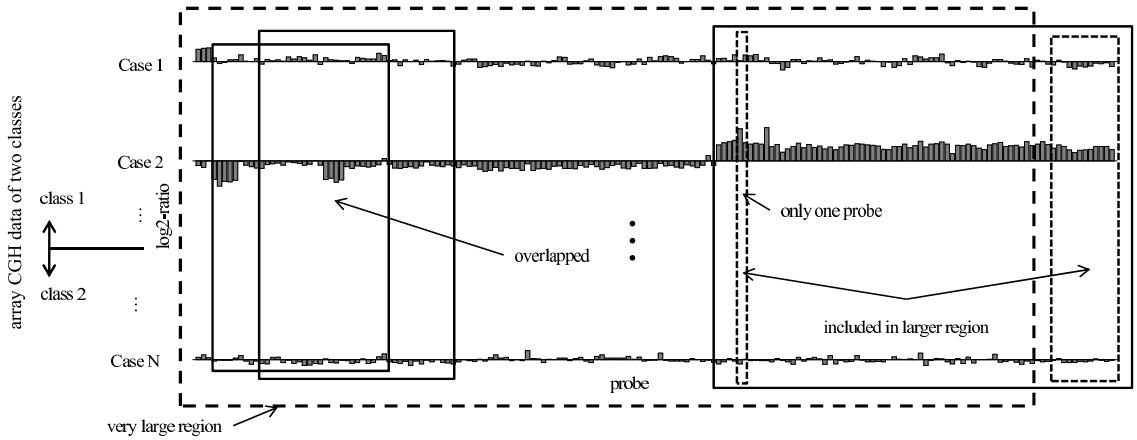


図 5.5: array CGH の異常領域同定問題における候補領域の例

において, 非常に多くの高度に相関を持った検定統計量を取り扱う必要があることを意味する. 最後に, 3 つ目のタスクにおいては, 次元数に依存しない多変量検定統計量を用いる必要がある. 例えば, 1 プローブのみからなる領域と, 1 染色体全体からなるような領域を同じ基準で比較しなければならないためである. array CGH データのグループ間で特徴の異なるゲノム領域を同定するタスクも, 遺伝子群解析と同様, 多重多変量 2 標本検定として定式化される. 図 5.5 は 1 番染色体における array CGH の $\log_2$-ratio 系列の例である. ここでは, 1 番染色体には 173BAC プローブが含まれている ($\ell_1 = 173$). 図では, 異常領域の候補となる部分領域の例が示されている. ゲノム異常領域同定問題では, 例えば, これら 173 個のプローブの全ての部分領域, 図の例では $\frac{1}{2} \times 173 \times 174 = 15,051$ 通り, を調べる必要がある.

5.5 多変量 2 標本検定と多重検定

前述した通り, 遺伝子群解析, array CGH データ解析は, ともに多重多変量 2 標本検定として定式化される. 本節では, 多変量 2 標本検定および多重検定について説明する.

5.5.1 多変量 2 標本検定と統計的機械学習

$D = \{(\boldsymbol{x}_i, y_i)\}_{\mathbb{N}_n}$ を n 個の観測データ, $\boldsymbol{x}_i \in \mathbb{R}^p$ を p 次元の入力ベクトル, $y_i \in \{-1, +1\}$ を出力ラベルとする. また, 2 クラスをそれぞれ C_-, C_+ とし, 各々のクラスに属するデータ数をそれぞれ n_-, n_+ とする: $n = n_- + n_+$. 更に, クラス C_- に属す

るデータは多変量分布 P_- から, クラス C_+ に属するデータは多変量分布 P_+ から生成されたものとする. 多変量2標本検定問題は, データ D を用いて, 2つの多変量分布 P_-, P_+ が同一の分布であるか, あるいは異なる分布であるかを評価する.

多変量2標本検定は主に統計学において研究が行われてきた. もし, 多変量分布 P_-, P_+ が分散共分散行列が等しい多変量正規分布であれば, Hotelling T^2 検定 [53] が検出力の最も高い検定となることが知られている. しかし, 分布 P_-, P_+ に対する事前知識がない場合には, ノンパラメトリックな多変量検定を用いる必要がある. 多くのノンパラメトリック多変量検定の検定統計量は, データ点のペアの距離に基づいて定義される. 例えば, 最近傍検定 (nearest-neighbor test) [55, 56] は, 最近傍法の分類精度を検定統計量として用いている.

多変量2標本検定の重要なあるクラスは, 統計的機械学習における2クラス分類器に基づいて構築できる. $f : \mathcal{X} \in \mathbb{R}^p \to \{-1, +1\}$ を観測データ D を用いて作成された2クラス分類器とする. 分類器 f の任意の分類性能の指標が検定統計量となりうる. 例えば, 訓練誤差, 汎化誤差, あるいはサポートベクトルマシン (SVM) のヒンジ損失などの損失値などを用いることができる. 良い分類性能は分布 P_-, P_+ が異なることを示唆しており, また分類性能の悪さは2つの分布に差異がないことを示唆する. ここでは, 検定統計量を $s(f, D)$ と表す. 分類器が強力であればあるほど, すなわち, 帰無分布と検定統計量の分布 (対立仮説のもとでの分布) が大きく異なるほど, Type II error を小さくすることができる.

統計的検定においては, P_-, P_+ の違いの統計的信頼性を定量化する必要がある. 典型的な例としては, 統計的信頼性は p-値により量られる. 例えば, 分類誤差を検定統計量として用い, 訓練誤差が 20%であったとする. このとき, p-値は, P_-, P_+ が同一のとき, 訓練誤差が 20%未満である確率と定義される.

5.5.2 ラベル並べ替えによる帰無分布推定

前述した通り, 検定統計量の帰無分布は p-値など, 統計的信頼性を評価する際に必要になる. 帰無分布を推定するための, 簡単かつ単純な方法の1つはラベル並べ替えを用いることである [58]. $\boldsymbol{y}$ を, i 番目の要素が y_i であるような n 次元ベクトルとし, π を n 個の要素の並べ替え演算子とする. ラベル並べ替えは $\boldsymbol{y}' = \pi(\boldsymbol{y})$ と表され, 対応するデータセットは, $D' \equiv \{(\boldsymbol{x}_i, y'_i)\}_{i \in \mathbb{N}_n}$ と表される. B 回の並べ替えを考えることとし, 各々の並べ替えを上付き添字 $b \in \mathbb{N}_B$ とインデックス表記する. 並べ替え演算は $\pi^{(b)}$ のようにインデックス化し, 並べ替えられたラベルは $\boldsymbol{y}^{(b)} = \pi^{(b)}(\boldsymbol{y})$ と表す. また, 対応するデータセットは $D^{(b)} = \{(\boldsymbol{x}_i, y_i^{(b)})\}_{i \in \mathbb{N}_n}$ と表し, $D^{(b)}$ を用いて学

習された分類器を $f^{(b)}$ と表す. これらを用いると, B 回のラベル並べ替えによって得られる p-値は,

$$p = \frac{\sum_{b=1}^{B} \mathcal{I}(s(f^{(b)}, D^{(b)}) \leq s(f, D))}{B}, \tag{5.1}$$

となる. ただし, $\mathcal{I}(\cdot)$ はインデックス関数である. ここで, 検定統計量が小さいものがより有意であるとした.

5.5.3 多重検定

複数の統計的検定を同時に行うと, 本来は棄却されるべきではない仮説が棄却される可能性が増加する (例えば P_- と P_+ が同一のものであるにも関わらず, 異なると判断されるなど). 例えば, 有意水準 5%で, 100 個の検定を行った場合, そのうち 5 個は仮説が誤って棄却されることを意味する. 検定対象が複数ある場合には全体としての誤検出の割合が有意水準を越えてしまう問題が生ずる. これを多重検定の問題と呼ぶ. このような場合, 多重検定補正を行わなければならない.

検定統計量が独立である多重検定の場合, 様々な補正法を適用できる. 一方, 遺伝子群解析, array CGH データの異常領域同定を含む多くのマイクロアレイデータ解析などのような, 検定統計量が複雑な相関を持つ場合や検定対象が非常に多いような場合には, それらを考慮した多重検定補正が必要になる. マイクロアレイデータ解析における多重検定については文献 [60] が詳しい. 前節で述べたラベル並べ替え検定は, このような状況においても有効な検定法である. というのは, ラベル並べ替え演算により, 複数の検定統計量の (同時) 確率分布をノンパラメトリックに推定することが可能であるからである.

しかしながら, p-値 (しばしば, nominal p-value と呼ばれる) は多重検定においては適切な指標ではない. 多重検定において用いられるいくつかの統計的有意性の指標の中で, family-wise error rate(FWER) [61] と false discovery rate(FDR) [62] が広く用いられている. 検定統計量が複雑な相関を持つような状況において, ラベル並べ替え検定は FWER, FDR を計算する際に重要な役割を果たす.

いま, M 個の検定対象があり, 各々の対象について B 回のラベル並べ替え演算により帰無分布を推定するとする. このとき, m 個目の検定統計量は $s(f_m,\ D_m)$ とし, m 個目の検定対象における並べ替え演算のうち b 番目の統計量を $s(f_m^{(b)},\ D_m^{(b)})$ と表す. FWER は "M 回の検定のうち, 1 度でも Type I error が起こる確率" と定義され,

$$FWER_m = \frac{\sum_{b=1}^{B} \mathcal{I}(\min_{m'} s(f_{m'}^{(b)}, D_{m'}^{(b)}) \leq s(f_m, D_m))}{B},$$

と計算される. 一方, FDRは“棄却された仮説の中の帰無仮説の割合=(# of false positive)/(# of positive)”と定義される. 統計量の小さな順にk個の検定統計量を選択した場合のFDRは次のように計算される. M個の統計量を有意な順(ここでは小さいほど有意としている)に, インデックスを$j = 1, \ldots, M$とソートしたとき,

$$\begin{aligned} FDR_m &= \min\left\{1.0, \frac{\frac{1}{B}\sum_{b=1}^{B}\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}^{(b)}, D_{m'}^{(b)}) \leq s(f_m, D_m))}{\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}, D_{m'}) \leq s(f_m, D_m))}\right\} \\ &= \min\left\{1.0, \frac{\frac{1}{B}\sum_{b=1}^{B}\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}^{(b)}, D_{m'}^{(b)}) \leq s(f_m, D_m))}{k}\right\} \end{aligned} \tag{5.2}$$

と計算される. 式(5.2)の分母は, “M回の検定においてType I errorを起こす回数を平均”を表している. この部分を, “M回の検定においてType I errorを起こす回数の**メジアン**”とすることも可能である. その場合, FDRは,

$$\begin{aligned} FDR_m &= \min\left\{1.0, \frac{\mathrm{med}_{b=1,\ldots,B}\left\{\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}^{(b)}, D_{m'}^{(b)}) \leq s(f_m, D_m))\right\}}{\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}, D_{m'}) \leq s(f_m, D_m))}\right\} \\ &= \min\left\{1.0, \frac{\mathrm{med}_{b=1,\ldots,B}\left\{\sum_{m'=1}^{M}\mathcal{I}(s(f_{m'}^{(b)}, D_{m'}^{(b)}) \leq s(f_m, D_m))\right\}}{k}\right\} \end{aligned} \tag{5.3}$$

と計算される. また, ある統計量とそれよりも小さな(有意な)統計量を選択したときのFDRの最小値をq-値と呼び,

$$q\text{-value} = \min_{k|k\geq \mathrm{order}(m), k\in\mathbb{N}_M} FDR^{(k)},$$

で与えられる. ここで, $\{FDR^{(k)}\}_{k\in\mathbb{N}_M}$は検定統計量$\{s(f_m, D_m)\}_{m\in\mathbb{N}_M}$によってあらかじめ昇順にソートされた$\{FDR_m\}_{m\in\mathbb{N}_M}$のリストとし, $\mathrm{order}(m)$はm番目の検定の順序とする. これらの測度は, p-値とは異なり, M回全ての検定における各々の統計量の有用性を評価しているため, 多重性を考慮したものとなっている. また, 統計量の相関を考慮できる, すなわち, 統計量が同じような値を取る場合に, それらを用いた検定はともに棄却あるいは採択される傾向を示す. なお, 本稿では, FDRについては平均を用いた前者の定義を用い, q-値を求める際は, 統計量でソートする代わりにp-値によりソートを行った.

第 6 章

サポートベクトルマシンを用いた遺伝子群解析

本章では, 生物学的に関連の深い複数の遺伝子からなるグループ (遺伝子群) のうち, 2 標本の発現パターンの違いを定量化し, その統計的信頼性を評価する問題を考察する. このような問題は遺伝子群解析 (gene set analysis) と呼ばれ, 多重多変量 2 標本検定として定式化される. 遺伝子群解析を行うことで, 解析に用いた遺伝子発現量データがどのような遺伝子群と関連が深いかを知ることができ, 解析の結果から有益な知見を得ることができると期待される. ここでは, この問題のためにサポートベクトルマシン (SVM) に基づく多重多変量 2 標本検定を導入する. このアプローチにおいては, 統計量の帰無分布をラベル並べ替え演算により推定しなければならないため, SVM の学習 (最適化) を多数回行わなくてはならない. 本稿では, SVM のラベル並べ替え解を効率的に計算するため, 最小全域木 (MST) とパス追跡を用いるアプローチを提案する [63].

6.1 SVM を用いた多重多変量 2 標本検定

本節では, 遺伝子群解析および SVM の leave-one-out cross validation (LOOCV) 誤差を検定統計量とした多変量 2 標本検定について説明をする.

6.1.1 遺伝子群解析の定式化

遺伝子群解析では遺伝子群と呼ばれる単位でマイクロアレイデータを解析する. 典型的な遺伝子群解析の問題は以下のように定式化される. 2 つのクラス (例えば, 健常者と癌患者など) をそれぞれ C_-, C_+ とし, 各々のデータ数を n_-, n_+, 全データ数を $n(= n_- + n_+)$ とする. また, 全遺伝子数を p とする. 症例 $i \in \mathbb{N}_n$, 遺伝子 $j \in \mathbb{N}_p$ の発現量を x_{ji} と表し, 症例 $i \in \mathbb{N}_n$ のラベルを $y_i \in \{-1, +1\}$ とする. クラス C_-, C_+ に

所属するデータは行列を用いて,

$$X_- = \begin{bmatrix} x_{11} & \dots & x_{1n_1} \\ \vdots & \ddots & \vdots \\ x_{p1} & \dots & x_{pn_1} \end{bmatrix}, X_+ = \begin{bmatrix} x_{11} & \dots & x_{1n_2} \\ \vdots & \ddots & \vdots \\ x_{p1} & \dots & x_{pn_2} \end{bmatrix},$$

と表され, 全データは,

$$\begin{aligned} X &= [X_- \; X_+] \\ &= \begin{bmatrix} x_{11} & \cdots & x_{1n_1} & x_{1,n_1+1} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn_1} & x_{p,n_1+1} & \cdots & x_{pn} \end{bmatrix}, \end{aligned}$$

と表現できる. 遺伝子群は遺伝子の集まりであるので, 行番号の集合として定義される. 遺伝子群の数を M とし, 遺伝子群を $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ と表すと, ある遺伝子群 $\mathcal{G}_m$ の発現量は $\{x_{ji}\}_{j\in\mathcal{G}_m, i\in\mathbb{N}_n}$ と表される. 本研究の目的は, いくつかの遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ の各々に対して, C_- と C_+ の発現パターンの違いを定量化し, その統計的信頼性を評価することである. この問題は, 多重多変量2標本検定として定式化される. 遺伝子群解析では遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ に対して, 各々の遺伝子群に含まれる遺伝子, すなわち, 行の部分集合, を抽出してそれぞれを多変量2標本データとみなした統計的検定を行う.

6.1.2 多変量2標本検定

本節では, 多変量2標本検定について説明する. 遺伝子群解析のタスクは遺伝子群に含まれるすべての遺伝子の発現量の (多変量) 分布が2標本で異なっているかを統計的に検証することであり, 多変量2標本検定として定式化される.

多変量2標本検定では, 2標本 X_-, X_+(前節参照) を用いて, それらを生成した多変量分布に違いがあるかどうかを評価する. 2標本が分散共分散行列の等しい多変量正規分布に従うとわかっている場合, Hotelling T^2 検定が最も検出力の高い検定であり, マイクロアレイデータ解析においても応用されている [59]. 一方, 2標本の従う分布が未知の場合は, ノンパラメトリック検定を採用する必要がある. ノンパラメトリック検定の例として, runs test [54], nearest-neighbor test [55, 56] などがある. 本研究ではSVMの分類誤差を検定統計量とした多変量2標本検定により遺伝子群解析を行う.

6.1.3 遺伝子群解析における多重検定

本節では遺伝子群解析におけるラベル並べ替え検定について説明する. 5.5.2 節で述べたように, 検定対象が複数ある場合は多重検定補正を行う必要がある. 遺伝子群解析では, 複数の遺伝子群に対して同時に検定を行う必要がある. 従って, ラベル並べ替えによる多重検定補正を行う. また, SVM 分類誤差検定統計量はその帰無分布が未知であるため, ラベル並べ替えにより帰無分布の推定を行う. すなわち, 並べ替えられたラベルのそれぞれに対する SVM を学習し, 帰無分布の推定を行う.

6.1.4 サポートベクトルマシン

サポートベクトルマシン (Support Vector Machine: SVM) [31] とは, 2 クラス分類問題に対して高い識別性能を持つパターン認識の手法の一つである. 学習データを $\{(\boldsymbol{x}_i, y_i)\}_{i\in\mathbb{N}_n}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, とすると, SVM の学習は以下のような 2 次計画問題として定式化される:

$$\min_{b,\boldsymbol{w}} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)). \tag{6.1}$$

ここで, $C \in [0, \infty)$ は正則化パラメータである. ラグランジュ未定乗数として $\{\alpha_i\}_{i\in\mathbb{N}_n}$ を導入すると, 式 (6.1) の双対問題は

$$\begin{aligned} \max_{\{\alpha_i\}_{i\in\mathbb{N}_n}} \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n} \alpha_i, \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i \in \mathbb{N}_n \end{aligned} \tag{6.2}$$

と表される. ただし, $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ はカーネル関数を表わしている. 主形式および双対形式における分類境界は, それぞれ,

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = 0,$$

および

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b = 0.$$

と与えられる. また, 最適性条件 (KKT 条件) は

$$\begin{aligned} \alpha_i = 0 \quad &\Leftrightarrow \quad y_i f(\boldsymbol{x}_i) \ge 1, \\ 0 \le \alpha_i \le C \quad &\Leftrightarrow \quad y_i f(\boldsymbol{x}_i) = 1, \\ \alpha_i = C \quad &\Leftrightarrow \quad y_i f(\boldsymbol{x}_i) \le 1, \end{aligned}$$

と整理される.

6.1.5 SVM分類誤差統計量を用いた多重多変量2標本検定

本節で説明したSVM分類誤差を検定統計量としたラベル並べ替え検定を用いた多重多変量2標本検定のアルゴリズムは以下の通りである.

SVMとラベル並べ替えによる多変量2標本検定

入力 マイクロアレイデータ $\{x_{ji}\}_{j\in\mathbb{N}_p,i\in\mathbb{N}_n}$, ラベル $\{y_i^*\}_{i\in\mathbb{N}_n}$, 遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$, ラベル並べ替え回数 B, シャッフルされたラベル群 $\{\{y_i^{(b)}\}_{i\in\mathbb{N}_n}\}_{b\in\mathbb{N}_B}$, 有意水準 θ, $m\leftarrow 1$.

Step 1 データ $\{x_{ji}\}_{j\in\mathcal{G}_m,i\in\mathbb{N}_n}$ およびラベル $\{y_i^*\}_{i\in\mathbb{N}_n}$ を用いてSVMによる分類器を作成し, 検証用データあるいはLOOCVなどにより分類誤差 s_m^* を計算, $b\leftarrow 1$.

Step 2 $\{x_{ji}\}_{j\in\mathcal{G}_m,i\in\mathbb{N}_n}$ とシャッフルされたラベル $\{y_i^{(b)}\}_{i\in\mathcal{N}}$ を用いて分類器を作成し, LOOCVなどによりSVM分類誤差 $s_m^{(b)}$ を計算. $b<B$ ならば $b\leftarrow b+1$ としStep 2を繰り返す. $b=B$ で $m<M$ であれば $m\leftarrow m+1$ としてStep 1へ. $b=B$ かつ $m=M$ であればStep 3へ.

Step 3 遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ の各々の $\mathcal{G}_m$ に対する,

$$p\text{-value}_m = \frac{\sum_{b=1}^{B}\mathcal{I}(s_m^{(b)}\le s_m^*)}{B} \tag{6.3a}$$

$$FWER_m = \frac{\sum_{b=1}^{B}\mathcal{I}\{\min_{m'} s_{m'}^{(b)}\le s_m^*\}}{B}, \tag{6.3b}$$

$$FDR_m = \frac{\sum_{b=1}^{B}\sum_{m'=1}^{M}\mathcal{I}\{s_{m'}^{(b)}\le s_m^*\}}{B\sum_{m'=1}^{M}\mathcal{I}\{s_{m'}^*\le s_m^*\}}, \tag{6.3c}$$

$$q\text{-value}_m = \min_{k\in\{k|k\ge\text{order}(m),k\in\mathbb{N}_M\}} FDR^{(k)}, \tag{6.3d}$$

を計算する. ここで, $\{FDR^{(k)}\}_{k\in\mathbb{N}_M}$ は $\{p\text{-value}\}_{m\in\mathbb{N}_M}$ の大小によってあらかじめ昇順にソートされた $\{FDR_m\}_{m\in\mathbb{N}_M}$ のリストとする. また, order(m) は遺伝子群 m の順序とする.

出力 基準, 式(6.3a)〜式(6.3d)などが閾値 θ 以下の遺伝子群.

6.2 SVMラベル並べ替え解計算の効率化

前述のように, 多重検定補正に必要なラベル並べ替え回数は通常, 1000〜10000と多い. 従って, これらのラベル群それぞれに対してSVMの学習を行うと計算コストが膨大になる. 本稿では以下の2つのアプローチによりSVMの学習の効率化を図る:

(i) パス追跡による並べ替えラベル群の最適解の追跡,

(ii) 最小全域木 (MST) による効率的なパス追跡スケジューリング.

(i) のパス追跡とは, 学習データ点の追加, 削除や正則化係数の変化などが起こった場合, 最初から学習をしなおすのではなく, 最適解の感度分析に基づいて変更後の最適解を効率的に計算する方法である. 本研究では, 並べ替えられたラベル変化を連続的に追跡する必要があり, 離散的なラベル $\{y_i\}_{i\in\mathbb{N}_n} \in \{-1,+1\}^n$ を実数値も許すように $\{y_i\}_{i\in\mathbb{N}_n} \in [-1,+1]^n$ と緩和した SVM を考える必要がある. 6.2.1 節にて (i) を説明する.

パス追跡では変更前後で最適解が大きく異なる場合, 追跡が非効率となるため, 並べ替えられたラベル群のパス追跡の順序を上手くスケジューリングする必要性が生ずる. この問題に対して, (ii) の方法により対処する. ラベル群をグラフのノード, ラベル間の距離をエッジの重みと考え, ラベル群の MST を構築し, ラベル間の総距離が全体として最小となるようスケジューリングを行う. 6.2.2 節にて (ii) を説明する.

6.2.1 SVM パス追跡によるラベル並べ替え解追跡

本節では, ラベル変化に対する SVM の最適解のパス追跡を説明する.

あるクラスのパラメータにより特徴づけられた凸計画問題では, パラメータが微小変化した際の最適解の変化のパスを厳密に求めることができる. この手法はパラメトリック計画法 (parametric programming), あるいは, パス追跡と呼ばれる. 機械学習の文脈におけるパス追跡アルゴリズムは, パスの区分線形性を利用するものが多い. もし, 最適解がパラメータに関して区分線形で表されるなら, 区分線形パス追跡アルゴリズムによって最適解を効率的に計算することができる. 本研究では, 並べ替えられたラベルの SVM の学習にパス追跡を適用する. 以降, ラベル $\{y_i\}_{i\in\mathbb{N}_n}$ をベクトル $\boldsymbol{y} = [y_1,\ldots,y_n]^\top$ と表記する.

並べ替えられたラベル $\boldsymbol{y}^{(b_1)} \in [-1,+1]^n$ に対する SVM の最適解が得られているとき, 別のラベル $\boldsymbol{y}^{(b_2)} \in [-1,+1]^n$ に対する SVM の最適解を計算したいとする. 通常の SVM の (双対) 最適化問題, 式 (6.2) において, ラベル $\{y_i\}_{i\in\mathbb{N}_n}$ が実数値 $[-1,+1]$ を取り得るように緩和すると, $\boldsymbol{y}$ を変化させたときの最適解のパスは区分線形とな

らない. そこで, 式 (6.2) を以下のような凸計画問題として緩和する:

$$\begin{aligned} \max_{\{\alpha_i\}_{i\in\mathbb{N}_n}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \mathrm{sgn}(y_i)\mathrm{sgn}(y_j)K(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i, \\ \text{s.t.} \quad & \sum_{i=1}^{n}\alpha_i \mathrm{sgn}(y_i) = 0, \\ & 0 \le \alpha_i \le y_i C, \quad \text{for } i \in \{i|y_i > 0\}, \\ & 0 \le \alpha_i \le -y_i C, \quad \text{for } i \in \{i|y_i < 0\}. \end{aligned} \tag{6.4}$$

ここで,

$$\mathrm{sgn}(z) = \begin{cases} +1 & \text{if } z > 0, \\ 0 & \text{if } z = 0, \\ -1 & \text{if } z < 0. \end{cases}$$

である. 式 (6.4) は, $\boldsymbol{y} \in \{-1,+1\}^n$ のとき, 元の SVM の最適化問題, 式 (6.2) と一致する. さらに, 式 (6.4) の最適解はラベル $\boldsymbol{y} \in [-1,+1]^n$ の区分線形関数として表される. 式 (6.4) の最適性条件は以下のようにまとめられる:

$$\begin{aligned} &\mathrm{sgn}(y_i)g(\boldsymbol{x}_i) \ge 1, \quad \text{if} \quad \alpha_i = 0, \\ &\mathrm{sgn}(y_i)g(\boldsymbol{x}_i) = 1, \quad \text{if} \quad 0 < \alpha_i < C|y_i|, \\ &\mathrm{sgn}(y_i)g(\boldsymbol{x}_i) \le 1, \quad \text{if} \quad \alpha_i = C|y_i|, \end{aligned} \tag{6.5}$$

$$\sum_{i=1}^{n}\mathrm{sgn}(y_i)\alpha_i = 0. \tag{6.6}$$

ただし,

$$g(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i \mathrm{sgn}(y_i)K(\boldsymbol{x},\boldsymbol{x}_i) + b$$

である.

ここで, 以下のような 3 つの集合を定義する:

$$\begin{aligned} \mathcal{O} &= \{i|\alpha_i = 0\}, \\ \mathcal{M} &= \{i|0 < \alpha_i < C|y_i|\}, \\ \mathcal{I} &= \{i|\alpha_i = C|y_i|\}. \end{aligned}$$

また, 各集合の要素数を $|\mathcal{O}|, |\mathcal{M}|, |\mathcal{I}|$ と表記する. 今後, ベクトル $\boldsymbol{v} \in \mathbb{R}^n$ に対して $\boldsymbol{v}_{\mathcal{I}}$ のように表記する場合, 集合 $\mathcal{I}$ に含まれるインデックスの要素を取り出した部分ベクトルを表すものとする. 同様に, 行列 $M \in \mathbb{R}^{n\times n}$ に対して $M_{\mathcal{M},\mathcal{O}}$ とした場合,

行列 M の集合 $\mathcal{M}$ に含まれるインデックスの行および集合 $\mathcal{O}$ に含まれるインデックスの列を取り出した部分行列を表すものとする. また, $M_{\mathcal{M},\mathcal{M}}$ のような部分行列は $M_{\mathcal{M}}$ と略記する.

今, (i,j) 要素が $Q_{ij} = \mathrm{sgn}(y_i)\mathrm{sgn}(y_j)K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ で与えられる行列を $Q \in \mathbb{R}^{n\times n}$ とする. また, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\top$ とする. このとき, KKT 条件, 式 (6.5) は次のように書き表せる:

$$Q_{\mathcal{M}}\boldsymbol{\alpha}_{\mathcal{M}} + Q_{\mathcal{M},\mathcal{I}}C|\boldsymbol{y}_{\mathcal{I}}| + \mathrm{sgn}(\boldsymbol{y}_{\mathcal{M}})b = \mathbf{1}_{|\mathcal{M}|}. \tag{6.7}$$

ただし, $\mathrm{sgn}(\boldsymbol{y})$ は $\boldsymbol{y}$ の各要素の符号を表すベクトルであり, $\mathbf{1}_{|\mathcal{M}|}$ は全ての要素が 1 となる $|\mathcal{M}|$ 次元ベクトルである. また, $|\boldsymbol{y}|$ は $\boldsymbol{y}$ の各要素の絶対値をとったベクトルを表す. 同様に, KKT 条件, 式 (6.6) は次のように表される:

$$\mathrm{sgn}(\boldsymbol{y}_{\mathcal{M}})^\top\boldsymbol{\alpha}_{\mathcal{M}} + \mathrm{sgn}(\boldsymbol{y}_{\mathcal{I}})^\top C|\boldsymbol{y}_{\mathcal{I}}| = 0. \tag{6.8}$$

次に, 以下のような行列を定義する:

$$A = \begin{bmatrix} 0 & \mathrm{sgn}(\boldsymbol{y}_{\mathcal{M}})^\top \\ \mathrm{sgn}(\boldsymbol{y}_{\mathcal{M}}) & Q_{\mathcal{M}} \end{bmatrix}.$$

この行列 A および式 (6.7), 式 (6.8) を用いて $|\mathcal{M}|+1$ 元連立方程式を構成し, それを解くと,

$$\begin{bmatrix} b \\ \boldsymbol{\alpha}_{\mathcal{M}} \end{bmatrix} = -A^{-1}\begin{bmatrix} \mathrm{sgn}(\boldsymbol{y}_{\mathcal{I}})^\top \\ Q_{\mathcal{M},\mathcal{I}} \end{bmatrix} C|\boldsymbol{y}_{\mathcal{I}}| + A^{-1}\begin{bmatrix} 0 \\ \mathbf{1}_{|\mathcal{M}|} \end{bmatrix}, \tag{6.9}$$

が得られる. ここで, 行列 A が逆行列を持つと仮定した. b および $\boldsymbol{\alpha}_{\mathcal{M}}$ が $\boldsymbol{y}_{\mathcal{I}}$ に関するアフィン関数であるため, ラベル $\boldsymbol{y}$ が変化したときのパラメータ $b, \boldsymbol{\alpha}_{\mathcal{M}}$ を求めることができる. 集合 $\mathcal{O},\mathcal{I}$ の定義から, 残りのパラメータは以下のように表される:

$$\boldsymbol{\alpha}_{\mathcal{O}} = \mathbf{0}_{|\mathcal{O}|} \text{ and } \boldsymbol{\alpha}_{\mathcal{I}} = C|\boldsymbol{y}_{\mathcal{I}}|. \tag{6.10}$$

もし, 集合 $\mathcal{O},\mathcal{M},\mathcal{I}$ およびラベル $\boldsymbol{y}$ の各要素の符号に変化がなければ, 式 (6.4) の解は式 (6.9) および式 (6.10) により, $\boldsymbol{y}$ のアフィン関数として計算可能である. 一方, いずれかの集合あるいはラベルのある要素の符号に変化がある場合, 式 (6.9) および式 (6.10) を更新しなければならない. ラベル $\boldsymbol{y}^{(b_1)}$ から $\boldsymbol{y}^{(b_2)}$ へのパスを考えるため, $\eta \in [0,1]$ を導入する. ある $\eta \in [0,1]$ により, ラベルが $\Delta\boldsymbol{y} = \eta\big(\boldsymbol{y}^{(b_2)} - \boldsymbol{y}^{(b_1)}\big)$ だけ更新され, $\boldsymbol{y} = \boldsymbol{y}^{(b_1)} + \Delta\boldsymbol{y}$ と表されているとする. また, 現在の最適解を $\boldsymbol{\alpha}, b$ とする. 以

下の条件が保たれるとき, 式 (6.9) および式 (6.10) が満たされる:

$$\mathrm{sgn}(y_i)g(\boldsymbol{x}_i) + \mathrm{sgn}(y_i)\Delta g(\boldsymbol{x}_i) \geq 1, \qquad i \in \mathcal{O}, \tag{6.11a}$$

$$\alpha_i + \Delta\alpha_i > 0, \qquad i \in \mathcal{M}, \tag{6.11b}$$

$$\alpha_i + \Delta\alpha_i - C(|y_i + \Delta y_i|) < 0, \qquad i \in \mathcal{M}, \tag{6.11c}$$

$$\mathrm{sgn}(y_i)g(\boldsymbol{x}_i) + \mathrm{sgn}(y_i)g(\boldsymbol{x}_i) \leq 1, \qquad i \in \mathcal{I}, \tag{6.11d}$$

$$\mathrm{sgn}(y_i)(y_i + \Delta y_i) > 0, \qquad i \in \mathbb{N}_n. \tag{6.11e}$$

ここで, 演算子 Δ は各変数の変化量を表す. もし, η を増加させたとき, 式 (6.11a)〜式 (6.11e) のいずれかの条件が破られる (パス追跡の文脈ではイベントと呼ばれる) ようであれば, 式 (6.9) と式 (6.10) を更新しなければならない. 上記の条件を監視することにより, このようなイベントが起こる η を厳密に求めることが可能である. ここで, スカラー ϕ を,

$$\phi = -A^{-1} \begin{bmatrix} \mathrm{sgn}(\boldsymbol{y}_\mathcal{I})^\top \\ Q_{\mathcal{M},\mathcal{I}} \end{bmatrix} C(\boldsymbol{y}_\mathcal{I}^{(b_2)} - \boldsymbol{y}_\mathcal{I}^{(b_1)})$$

と定義すると, 式 (6.9) より, $\boldsymbol{\alpha}_\mathcal{M}$ および b の変化量は,

$$\begin{bmatrix} \Delta b \\ \Delta\boldsymbol{\alpha}_\mathcal{M} \end{bmatrix} = \eta\phi \tag{6.12}$$

と表される. さらに, $\mathrm{sgn}(y_i)\Delta g(\boldsymbol{x}_i)$ は,

$$\begin{aligned} &\mathrm{sgn}(y_i)\Delta g(\boldsymbol{x}_i) = \eta\psi_i, \\ &\psi_i = \begin{bmatrix}\mathrm{sgn}(y_i) & Q_{i,\mathcal{M}}\end{bmatrix}\phi + Q_{i,\mathcal{I}}C(\boldsymbol{y}_\mathcal{I}^{(b_2)} - \boldsymbol{y}_\mathcal{I}^{(b_1)}), \end{aligned} \tag{6.13}$$

と表される. $y_i + \Delta y_i$ の符号は,

$$\eta = -\frac{y_i}{y_i^{(b_2)} - y_i^{(b_1)}}, \tag{6.14}$$

と変化する. インデックス集合 $\mathcal{M}$ 内の要素を $\{m_1, \ldots, m_\mathcal{M}\}$ と表すとする. 式 (6.12), 式 (6.13) および式 (6.14) より, 次のイベントが起こる η は,

$$\eta = \min_{\substack{i\in\mathbb{N}_{|\mathcal{M}|},\\ j\in\mathcal{O}\cup\mathcal{I},\\ k\in\mathbb{N}_n}}{}^{+}\left\{ -\frac{\alpha_{m_i}}{\phi_{i+1}}, \frac{C|y_{m_i}| - \alpha_{m_i}}{\phi_{i+1} - C(y_i^{(b_2)} - y_i^{(b_1)})}\frac{1 - \mathrm{sgn}(y_i)g(\boldsymbol{x}_i)}{\psi_i}, -\frac{y_k}{y_k^{(b_2)} - y_k^{(b_1)}} \right\} \tag{6.15}$$

と計算される. ここで, $\min_i^+\{z_i\}$ は $\min_i\{z_i | z_i \geq 0\}$ の簡略表記である. 式 (6.15) によって得られる η によりパラメータを更新後, イベントの種類に応じて集合の更新

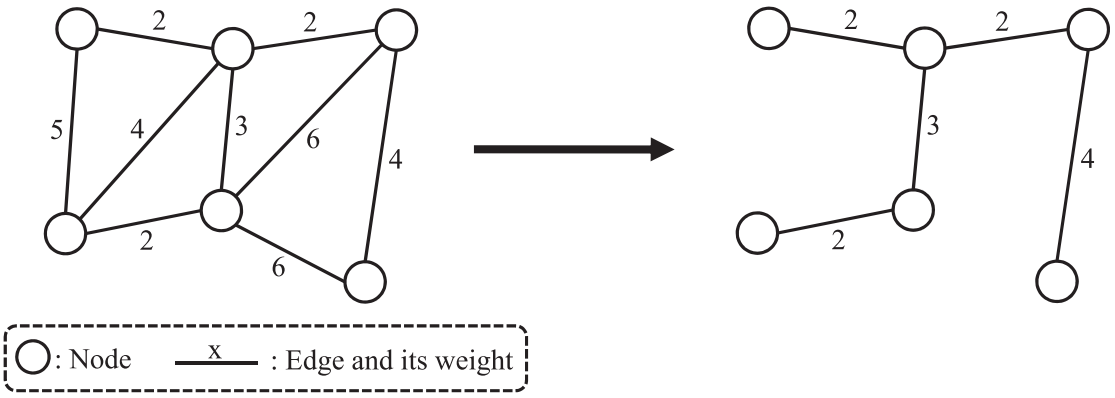


図 6.1: MST の例

を行う. 以上によって得られる η を順次計算していき, $\boldsymbol{y} = \boldsymbol{y}^{(b_2)}$ となったとき, すなわち, η が 1 となった時点でパスが終了し, $\boldsymbol{y}^{(b_2)}$ での SVM の最適解が得られる.

以上のパス追跡アルゴリズムは, ラベル $\boldsymbol{y}$ の符号に関するイベントを除き, $C_i = C|y_i|$ と定義される個別の正則化係数を持つような重み付き SVM [64] のパス追跡 [65] と解釈することができる.

6.2.2 最小全域木

最小全域木 (mininum spanning tree: MST) とは, 与えられた重み付き無向グラフにおいて, エッジの総コストが最小となるループのない木のことを指す [66]. そのような木を作成する問題を Minimum Spanning Tree Problem (MSTP) と呼び, 動的計画法を用いることにより効率的に解くことができる. MSTP の標準的な解法として Prim のアルゴリズム [67] がある. 詳細は割愛するが, 本研究ではこの Prim のアルゴリズムによりラベル群の MST を構築する. 図 6.1 に MST の例を示す.

6.2.3 MST の拡張

本研究では, パス追跡の効率化のために MST に対して以下のような改良を加える.

与えられたデータの (正しい) ラベルを $\boldsymbol{y}^* = [y_1^*, \ldots, y_n^*]^\top, \quad y_i \in \{-1, 1\}$, とする. また, ラベル並べ替え回数を B, シャッフルされたラベル集合を $\mathcal{Y}_B = \{\boldsymbol{y}^{(b)}\}_{b \in \mathbb{N}_B}$, $\mathcal{Y} = \{\boldsymbol{y}^*\} \cup \mathcal{Y}_B$ とする.

エッジの重みとして以下で与えられるラベル間の距離を用いる:

$$D(\boldsymbol{y}, \boldsymbol{y}') = \min(||\boldsymbol{y} - \boldsymbol{y}'||, ||\boldsymbol{y} - \tilde{\boldsymbol{y}}'||), \quad \boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}. \tag{6.16}$$

ここで, $\tilde{\boldsymbol{y}}$ は, $\tilde{\boldsymbol{y}} = -\boldsymbol{y}$ と定義され, 全てのラベルの符号を反転させたラベルを表す. このような距離を考える理由は次の通りである. 現在の目的は, 最適解の近さ, すなわち, 分類境界の変化の少なさをラベル変化量を見ることで測るのであった. 従って, 単純に 2 つのラベルの変化量 $D(\boldsymbol{y}, \boldsymbol{y}') = ||\boldsymbol{y} - \boldsymbol{y}'||$ を用いれば良いように思われるが, ある分類境界を導くラベルが完全に反転した状態での分類境界は, 反転前のものと一致している. 従って, あるラベルを基準として, 他方のラベルが元の状態および反転した状態で距離を計算し, より小さい方をその 2 つのラベル間の距離とすれば, より最適解の近いラベルを選択し, パス追跡の効率を高めることができると考えられる. パス追跡を行なう際にも, ラベルの反転の有無の情報を用いて最適化を行なう.

以上のように, 各ラベル間の距離を計算し MST を構成する. 次に, MST の総コストをより小さくするため, 仮想ラベルの概念を導入する. 仮想ラベルとは, 実際のラベル群には含まれていないが, 仮想ラベルを経由することで MST のコストをより小さくできるような仮想のラベルを指す. これは Steiner Tree [68] と呼ばれるものの概念を用いたものであるが, 今回は次のようなヒューリスティックな方法により仮想ラベルを作成する. なお, ラベル群 $\mathcal{Y}$ からなる MST のコスト $C(\mathcal{Y})$ は, ラベル間の距離の総和と定義する. すなわち, ラベル $\boldsymbol{y}_i$, $\boldsymbol{y}_j$ が連結しているとき, (i, j) と表記し, 連結関係の集合を $\mathcal{E} = \{(i, j) | i, j \in \mathbb{N}_{|\mathcal{Y}|}\}$ としたとき,

$$C(\mathcal{Y}) = \sum_{(i,j)\in\mathcal{E}} D(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

と定義する.

まず, 先に作成した MST の中からあるラベルをランダムに選択する. 次に, そのラベルから辿ることのできる $K-1$ 個のラベルをランダムに辿りながら選択する. 得られた K 個のラベルを $\mathcal{Y}_K = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$, また, $\mathcal{Y}_K$ に対して張られた MST のコストを $C(\mathcal{Y}_K)$ とする. 次に $\mathcal{Y}_K$ 内のラベルの重心を計算する. この際, 重心から $\mathcal{Y}_K$ に属する各ラベルへの距離の総和が最小となる重心を求めたい. そこで, K 個のラベルに対する全ての反転のパターンを考慮した 2^K 個の重心を計算する. 便宜上, 次のようなベクトルを定義する. 先に述べたように, あるラベル $\boldsymbol{y}$ を反転させたラベル $\tilde{\boldsymbol{y}}$ は $\tilde{\boldsymbol{y}} = -\boldsymbol{y}$ により与えられる. そこで, $\boldsymbol{r}_c = [r_1, \ldots, r_K]^\top$, $r_k \in \{-1, +1\}$ なるベクトルを考え, あるラベル $\boldsymbol{y}_k \in \mathcal{Y}_K$ に対して, 反転させる場合は $r_k = -1$,

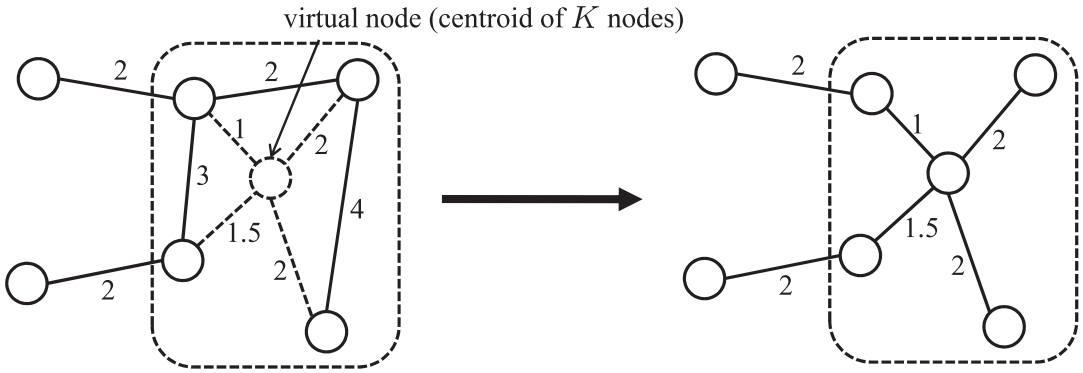


図 6.2: 仮想ラベル作成の例 ($K = 4$): 枠内のコストは 9 であるが, 仮想ラベルを経由し 4 つのラベルを接続した場合, コストが 6.5 となる. このような場合は仮想ラベルを追加する.

反転させない場合は $r_k = +1$ とする. $\{-1, +1\}$ からなる K 次元ベクトルの全てのパターンを列挙し, $\{\boldsymbol{r}_c\}_{c\in\mathbb{N}_{2^K}}$, $\boldsymbol{r}_c \in \{-1, +1\}^K$, と表す. 例えば, $K = 2$ のとき, $\{\boldsymbol{r}_c\}_{c\in 2^2} = \{[-1, -1]^\top, [-1, +1]^\top, [+1, -1]^\top, [+1, +1]^\top\}$ である. これらのラベル反転情報を用いて, 2^K 通りの重心,

$$\boldsymbol{g_c} = \frac{1}{K}\sum_{k=1}^{K} r_k \boldsymbol{y}_k, \quad c \in \mathbb{N}_{2^K},\ r_k \in \boldsymbol{r}_c,\ \boldsymbol{y}_k \in \mathcal{Y}_K, \tag{6.17}$$

が計算される. 式 (6.17) 中で, $r_k = -1$ に対しては $r_k\boldsymbol{y}_k = -\boldsymbol{y}_k = \tilde{\boldsymbol{y}_k}$ と反転されたラベルで計算されることになる. 従って, 全ての $\boldsymbol{r}_c$ に対して式 (6.17) を計算することで, 全ての反転パターンの重心が得られる. これらの重心 $\mathcal{G} = \{\boldsymbol{g}_c\}_{c\in\mathbb{N}_{2^K}}$ のうち, 重心から $\mathcal{Y}_K$ 内の各ラベルへエッジを作成した場合の最小コスト,

$$\min_{\boldsymbol{g}_c\in\mathcal{G}} C(\mathcal{Y}_k \cup \{\boldsymbol{g}_c\}) = \min_{\boldsymbol{g}_c\in\mathcal{G}} \sum_{k=1}^{K} D(\boldsymbol{y}_k, \boldsymbol{g}_c),$$

が元のコスト $C(\mathcal{Y}_K)$ よりも小さければ, その最小コストとなる重心を仮想ラベルとして $\mathcal{Y}$ へ追加する:

$$\mathcal{Y} \leftarrow \mathcal{Y} \cup \arg\min_{\boldsymbol{g}_c\in\mathcal{G}} D_{\boldsymbol{g}_c} \quad \text{if} \quad \min_{\boldsymbol{g}_c\in\mathcal{G}} D_{\boldsymbol{g}_c} < C(\mathcal{Y}_k). \tag{6.18}$$

ただし,

$$D_{\boldsymbol{g}_c} \equiv \sum_{k=1}^{K} D(\boldsymbol{y}_k, \boldsymbol{g}_c)$$

とした. 仮想ラベルが追加される場合, $\mathcal{Y}_K$ で張られている MST を, 仮想ラベルを中心とした star 型木で置き換える. これにより MST の一部を見たときに, その部分のコストが仮想ラベル追加前よりも小さくなる. 以上の操作を, 仮想ラベルが T 個追加されるまで繰り返す. その後, 仮想ラベルを加えた $\mathcal{Y}$ に関して再度 MST を構築する. 仮想ラベルを Greedy に加えることで, 仮想ラベルを加える前の状態よりも総コストが減少し, パス追跡を効率的に行なうことが可能となる. 以下に仮想ラベル追加の流れを示す:

仮想ラベルを用いた MST 構築

入力 オリジナルラベル $\boldsymbol{y}^*$, 並べ替えラベル群 $\mathcal{Y}_B = \{\boldsymbol{y}^{(b)}\}_{b\in\mathbb{N}_B}$, ランダムに選ぶラベル数 K, 繰り返し回数 T.

Step 1 $\mathcal{Y} = \{\boldsymbol{y}^*\} \cup \mathcal{Y}_B$ に対して MST を構築する. $t \leftarrow 1$.

Step 2 仮想ラベルを追加する.

Step 2-1 $\mathcal{Y}$ からラベルをランダムに選択する.

Step 2-2 Step 2-1 で選択したラベルから辿ることのできる $K-1$ 個のラベルを選択し, これら K 個のラベルを $\mathcal{Y}_K$, $\mathcal{Y}_K$ に対して張られた MST のコストを $C(\mathcal{Y}_K)$ とする.

Step 2-3 K 個のラベル反転の有無を表すベクトル群 $\{\boldsymbol{r}_c\}_{c\in\mathbb{N}_{2^K}}$, $\boldsymbol{r}_c \in \{-1,+1\}^n$ を用いて, 2^K 個の重心,

$$\boldsymbol{g_c} = \frac{1}{K}\sum_{k=1}^{K} r_k \boldsymbol{y}_k, \quad c \in \mathbb{N}_{2^K},\ r_k \in \boldsymbol{r}_c,\ \boldsymbol{y}_k \in \mathcal{Y}_K,$$

を計算する.

Step 2-4 以下の式により, ラベル群 $\mathcal{Y}$ を更新する:

$$\mathcal{Y} \leftarrow \mathcal{Y} \cup \arg\min_{\boldsymbol{g}_c\in\mathcal{G}} D_{\boldsymbol{g}_c} \quad \text{if} \quad \min_{\boldsymbol{g}_c\in\mathcal{G}} D_{\boldsymbol{g}_c} < C(\mathcal{Y}_k).$$

ただし,

$$D_{\boldsymbol{g}_c} \equiv \sum_{k=1}^{K} D(\boldsymbol{y}_k, \boldsymbol{g}_c)$$

である.

Step 2-5 Step 2-4 で仮想ラベルが追加される場合は, $\mathcal{Y}_K$ に対して張られた MST を, 重心を中心とする star 型木で置き換える.

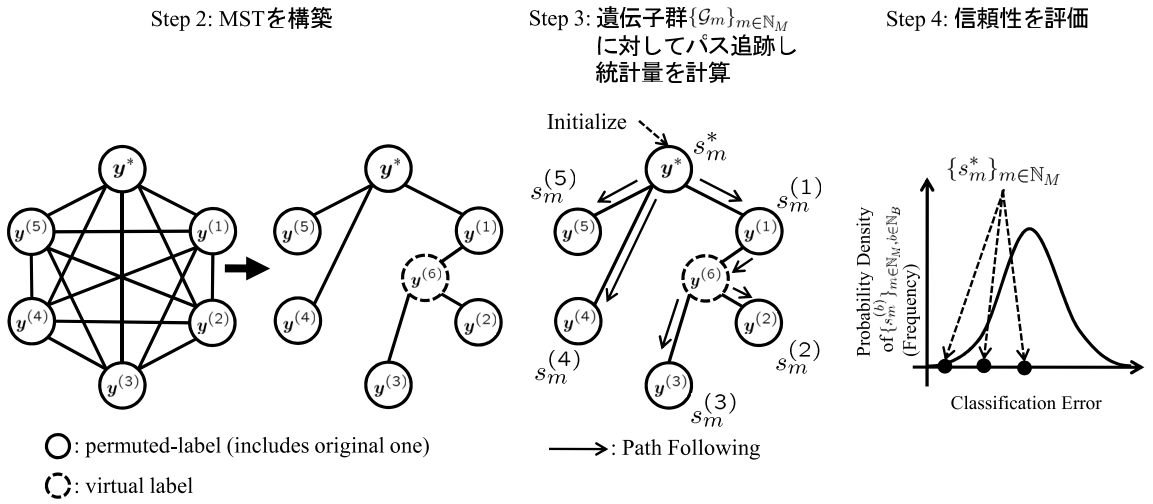


図 6.3: MST に基づく SVM パス追跡を用いた多重多変量 2 標本検定の概念図: Step 2 では生成されたオリジナルラベルおよび並べ替えラベル群に対して, 仮想ラベルを加えた MST を構築する. Step 3 では遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ に対して統計量 $\{s_m^*\}_{m\in\mathbb{N}_M}$ および帰無統計量 $\{s_m^{(b)}\}_{m\in\mathbb{N}_M, b\in\mathbb{N}_B}$ を計算する. Step 4 では推定された帰無分布を用いて, 各々の遺伝子群の統計的信頼性を評価する.

Step 3 $t = T$ であれば Step 4 へ. $t < T$ かつ Step 2 で仮想ノードが追加された場合は $t \leftarrow t+1$ として Step 2 へ. $t \neq T$ で仮想ラベルが追加されなければ t を更新せずに Step 2 へ.

Step 4 $\mathcal{Y}$ に対して MST を再構築する.

出力 仮想ラベルを含むラベル群 $\mathcal{Y}$ に対して構築された MST.

図 6.2 に仮想ラベル作成の例を示す.

ただし, 仮想ラベルは $\{-1, +1\}$ とは限らず, 一般に実数ラベル持つ. 3.1 節のパス追跡では 実数ラベルを取り扱うことができるため, このような拡張が可能である.

6.2.4 MST とパス追跡を用いた多重多変量 2 標本検定

ここでは, MST およびパス追跡により効率化した多重多変量 2 標本検定についてまとめる. MST に基づく SVM パス追跡を用いた多重多変量 2 標本検定による遺伝子群解析の流れは以下の通りである:

MST およびパス追跡を用いた遺伝子群解析

入力 マイクロアレイデータ $\{x_{ji}\}_{j\in\mathbb{N}_p,i\in\mathbb{N}_n}$, (オリジナル) ラベル $\boldsymbol{y}^* \in \{-1,+1\}^n$, 遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$, ラベル並べ替え回数 B, MST の重心計算に用いるラベル数 K, 追加する仮想ラベル数 T, 有意水準 θ. $m \leftarrow 1$.

Step 1 シャッフルされたラベル群 $\mathcal{Y}_B = \{\boldsymbol{y}^{(b)}\}_{b\in\mathbb{N}_B}$ を生成する.

Step 2 MST を構築する:

Step 2-1 $\mathcal{Y} = \{\boldsymbol{y}^*\} \cup \mathcal{Y}_B$ および距離関数式 (6.16) を用いて MST を構築する.

Step 2-2 T 個の仮想ラベルが追加されるまで式 (6.18) を繰り返し計算する.

Step 2-3 再度ラベル群 $\mathcal{Y}$ を用いて MST を再構築する.

Step 3 遺伝子群 $\mathcal{G}_m$ の分類誤差統計量を計算する:

Step 3-1 MST を順に辿りながら, ラベル $\boldsymbol{y} \in \mathcal{Y}$ とデータ $\{x_{ji}\}_{j\in\mathcal{G}_m,i\in\mathbb{N}_n}$ を用いて以下のステップを実行する. ただし, MST 作成時に $\boldsymbol{y}$ が反転されている場合は, $\boldsymbol{y} \leftarrow -\boldsymbol{y}$ を用いる:

Step 3-1(a) 親ラベルからのパス追跡により $\boldsymbol{y}$ での最適解を計算, 記憶する.

Step 3-1(b) もし $\boldsymbol{y}$ が仮想ラベルでなければ, LOOCV により分類誤差統計量を計算したものを, $\boldsymbol{y} = \boldsymbol{y}^*$ ならば s_m^*, $\boldsymbol{y} \in \mathcal{Y}_B$ ならば $s_m^{(\cdot)}$ とする.

Step 3-2 $m = M$ ならば Step 4 へ. そうでなければ $m \leftarrow m+1$ として Step 3 を繰り返す.

Step 4 式 (6.3a)〜 式 (6.3d) などにより遺伝子群 $\{\mathcal{G}_m\}_{m\in\mathbb{N}_M}$ の統計的信頼性を評価する.

出力 基準, 式 (6.3a)〜式 (6.3d) などが閾値 θ に関して有意な遺伝子群.

ここで, Step 3-1(b) においてもパス追跡を用いることにより, 効率的に LOOCV Error を求められることに注意されたい.

図 6.3 に上記アルゴリズムの概略を示す.

6.3 遺伝子群解析への応用

ここでは, 実際のマイクロアレイデータと遺伝子群を用いた計算機実験を行う. 実験の主旨は以下の 2 点である:

1. MST とパス追跡を用いた効率化の検証,

2. 既存の遺伝子群解析手法との比較.

実験に用いるデータはGSEAデータベース[1] [28, 69] から取得可能なC2.Diabetes($n = 34, M = 331$), C2.p53($n = 50, M = 291$) を用いる. なお, 遺伝子群は, 各々の遺伝子群に含まれる遺伝子数$|\mathcal{G}_m|$が15個以上500個以下となるもののみを用いることとした.

6.3.1 MSTとパス追跡を用いた効率化の検証

まず, 計算時間の比較を行う. 比較対象には, SVMの代表的な学習法であるSMOアルゴリズムを用いる. SMOアルゴリズムはLIBSVM [70] のプログラムに若干の改良を加えたものを用いた. また, SMOの初期値を全てのラベルで$\boldsymbol{\alpha} = \mathbf{0}$として設定するもの (SMO), LOOCVにおいて全データを用いた学習結果を初期値として用いるもの (SMO_hot) の二通りで実験を行った. SMOアルゴリズムの終了条件は10^{-6}とし, 正則化係数は$C \in \{10, 100, 1000\}$とした. また, カーネルにはガウシアンカーネル$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$を用い, ハイパーパラメータ$\gamma$は$\gamma = 10/$(データの次元数)とした. ラベル並べ替え回数は$B \in \{1000, 10000\}$とし, MSTのパラメータは$\{K, T\} = \{3, 100\}$と設定した. 結果を表6.1および6.2に示す. なお, パス追跡ではMSTを構築するために必要な時間が余分にかかることに注意されたい.

並べ替えられた全てのラベルにおいてその都度SVMを学習するSMOと比較して, MSTスケジューリングを用いてSVM最適解のパス追跡を行うことで計算時間が大幅に削減されている. MSTスケジューリングを行う場合, MSTを構築するプロセスが必要となるため, その分計算時間が増加する. しかし, SVMの学習と比較して, MSTの構築に必要な計算時間はそれほど大きくない. ただし, MSTを構築する際のラベル間距離の計算オーダは, 症例数n, ラベル並べ替え回数Bに対して$\mathcal{O}(nB^2)$, MST構築の計算オーダは$\mathcal{O}(B^2)$と, nやBが大きい場合には何らかの対処が必要になると思われる.

これらの結果から, MSTスケジューリングに基づくパス追跡による最適解の追跡は, SVM分類誤差を検定統計量とした多変量検定を効率的に行う上で効果的であることがわかる.

[1] http://www.broadinstitute.org/gsea/index.jsp

表 6.1: 計算時間の比較 (sec.): C2.Diabetes

	B	MST	$C = 10$	$C = 100$	$C = 1000$
MST_Path	1000	0.69589	794.32	771.77	771.77
SMO		-	2,115.6	2,206.6	2,210.5
SMO_hot		-	1,647.1	1,702.7	1,703.8
MST_Path	10000	61.225	7,427.8	7,119.3	7,113.2
SMO		-	31,430	22,017	22,120
SMO_hot		-	16,305	16,904	17,056

表 6.2: 計算時間の比較 (sec.): C2.p53

	B	MST	$C = 10$	$C = 100$	$C = 1000$
MST_Path	1000	0.89386	1,942.2	1,976.8	1,957.7
SMO		-	5,343.7	11,334	12,142
SMO_hot		-	3,538.2	6,986.5	7,709.1
MST_Path	10000	72.482	19,129	19,303	19,110
SMO		-	53,067	112,999	116,744
SMO_hot		-	35,615	69,883	73,776

6.3.2 遺伝子群解析結果の考察

次に, いくつかの手法により遺伝子群解析を行った結果を考察する. 比較に用いる手法は, GSEA[2] [28], t-値の平均を用いた手法 (Ave. T) [71], 最近傍検定 (NN) [55, 56], Hotelling T^2 検定 (Hot. T) [53], SVM 分類誤差統計量を用いた検定 (SVM) の 5 つである. GSEA の統計量は t-値を用いたものとした. t-値の平均を用いた手法の統計量は,

$$s_m = \frac{1}{\sqrt{|\mathcal{G}_m|}} \left(\frac{1}{|\mathcal{G}_m|} \sum_{j \in \mathcal{G}_m} t_j \right),$$

を用いる. ここで, $\{t_j\}_{j \in \mathcal{G}_\cdot}$ は遺伝子 j における t-値であり, 係数 $\frac{1}{\sqrt{|\mathcal{G}_\cdot|}}$ は遺伝子群のサイズによる統計量の補正である. Hotelling T^2 検定の統計量はデータの次元数に依存するため, 各々のラベルを用いて理論的に計算される p-値を用いた. 最近傍検定, SVM 分類誤差による検定では LOOCV による分類誤差を統計量として用いた.

[2]GSEA の結果は GSEA の Web サイトからダウンロード可能なソフトウェアを利用した. Web サイトおよび [28] の情報だけでは, 解析結果の FWER, FDR, q-値の計算方法が明確でないため, 我々が用いたこれらの値と基準が異なる可能性があることに注意されたい.

また, SVM のハイパーパラメータ C は前節の $C = 10$ のものを用いることとした. 全ての手法においてラベル並べ替え回数は $B = 10000$ とした.

これらの手法は単変量アプローチか多変量アプローチかにより以下のように分類される:

単変量アプローチ: GSEA, Ave. T

多変量アプローチ: NN, Hot. T^2, SVM

単変量アプローチの 2 手法は, 片側検定が可能である. 従って, クラス 1, クラス 2 の各々で, 各遺伝子群の FDR などの検証を行うことが可能である. 一方で, 多変量アプローチの 3 手法は統計量として, NN, SVM は判別率, Hot. T^2 は p-値を用いているため, 片側検定を行うことができない. 従って, 各遺伝子群が 2 クラスの判別に有用かどうかという判断のみに留まる. すなわち, 今回用いる 5 つの検定手法は,
単変量アプローチ : GSEA, Ave. T : 片側検定可
多変量アプローチ : NN, Hot. T^2, SVM : 片側検定不可
とまとめられる.

t-値の平均を用いた手法と GSEA は, 統計量の計算方法が似ているため, これらの手法は比較的似た結果が得られると予想される. なお, 遺伝子群解析では, 遺伝子群の全ての遺伝子を多変量的に解析することが好ましいが, 今回の比較はどの手法が良いかを検証するためではなく, 解析結果にどのような特徴がみられるかを考察するためであることに注意されたい.

表 6.3 に C2.p53 に対して各検定法により q-value < 0.25 に関して有意と検出された遺伝子群において, 他の検定法においても検出された遺伝子群の個数を示す. なお, C2.Diabetes では GSEA 以外の検定法では q-value < 0.25 に関して遺伝子群が検出されなかったため, 省略する. 表中で, 各セルの分子は左側の検定法において検出された遺伝子群のうち, 上側の検定法においても検出されていた個数を表す. また, 分母は左側の検定法で検出された遺伝子群の個数を表している. 例えば, 検定法 X で検出された遺伝子群の集合を $\mathcal{S}(X)$, そのサイズを $|\mathcal{S}(X)|$ としたとき, 表中の 1 行 2 列目は,

$$\frac{|\mathcal{S}(\mathrm{G}) \cap \mathcal{S}(\mathrm{A})|}{|\mathcal{S}(\mathrm{G})|},$$

を表す. 一番右の列には各手法で検出された遺伝子群に対する, 他の手法で検出された全ての遺伝子群との一致率を表す. 例えば, 1 行 6 列目は,

$$\frac{|\big(\mathcal{S}(\mathrm{A}) \cup \mathcal{S}(\mathrm{H}) \cup \mathcal{S}(\mathrm{N}) \cup \mathcal{S}(\mathrm{S})\big) \cap \mathcal{S}(\mathrm{G})|}{|\mathcal{S}(\mathrm{G})|},$$

表 6.3: 各検定法において検出された遺伝子群の一致率: C2.p53

	GSEA	Ave. T	Hot. T^2	NN	SVM	全手法との一致率
GSEA	-	5/5	3/5	4/5	4/5	5/5
Ave. T	5/66	-	8/66	8/66	20/66	27/66
Hot. T^2	3/36	8/36	-	8/36	19/36	22/ 36
NN	4/19	8/19	8/19	-	13/19	16/19
SVM	4/75	20/75	19/75	13/75	-	37/75

などとなる. ここで, GSEA, Ave. T, Hot. T^2, NN, SVM をそれぞれ, G, A, H, N, S と略記した.

GSEA で検出された 5 つの遺伝子群は全て, 同じ単変量アプローチである Ave. T でも検出されている. しかし, 多変量アプローチである Hot. T^2, NN, SVM では検出されなかった遺伝子群が存在する. 各検定法で検出された遺伝子群の多くは SVM による検定においても検出されている (4/5, 20/66, 19/36, 13/19 など). すなわち, SVM は単変量, 多変量の両アプローチの遺伝子群を検出していると考えられる. 特に, 多変量アプローチである Hot. T^2, NN における一致率 (19/36, 13/19) は単変量アプローチである Ave. T(20/66) よりも高いことがわかる. 逆に, SVM では多くの遺伝子が検出されている (75 遺伝子群) が, それらのうち, 約半数 (37 遺伝子群) が他の手法でも検出されており, SVM による遺伝子群解析の有用性を示唆している. SVM で検出された遺伝子群が医学的に意味のあるものであるかは, 医学生物学的な知見での検証が必要であるが, 本研究の主目的は, SVM による遺伝子群解析およびその効率化である. 従って, ここでは他手法との比較に関する考察に留める.

次に各手法で q-値が最小となった遺伝子群の p-値, FWER, q-値を表 6.4, 6.5 に示す. q-値が同位のものがあった場合は, FWER が小さいものを掲載した.

いずれのデータセットにおいても, GSEA と Ave. T で最も q-値が最小となる遺伝子群が一致していることがわかる. しかしながら, その値は大きく異なっている. 両手法とも t-値を基準とした統計量を用いているため, 似た結果が予想されたが, 注釈で述べたように FWER などの計算方法が異なる可能性があるため, それらの基準を統一して比較する必要がある. また, C2.Diabetes では, GSEA 以外の検定法では q-値の値が大きくなっており, q-value < 0.05 に関して有意な遺伝子群が検出されないという結果になった. この結果についても, FWER, FDR などの定義を統一して比較しなければならない. 一方, C2.p53 では, Hot. T^2 を除く 4 つの検定法にお

表 6.4: q-値最小の遺伝子群: C2.Diabetes

Test	Gene Set ID	p-value	FWER	q-value
GSEA	298	0.0008	0.0543	0.04441
Ave. T	298	0.0074	0.3122	0.7751
Hot. T^2	315	0.0065	0.744	0.7072
NN	76	0.0161	0.9177	0.9086
SVM	165	0.0017	0.2089	0.3541

表 6.5: q-値最小の遺伝子群: C2.p53

Test	Gene Set ID	p-value	FWER	q-value
GSEA	163	0	0.0203	0.0227
Ave. T	164	0	0.0078	0.0105
Hot. T^2	164	0	0.905	0.1203
NN	163	0.0005	0.0355	0.0405
SVM	75	0.0001	0.0029	0.0031

いて gene set ID 163 が検出されており, その q-値も比較的小さい (SVM では 0.03376 であった). また, C2.p53 においては, SVM の q-値が他の検定法と比較して小さくなっており, 検出力が高いことを示している.

6.4 まとめ

本章では, SVM 分類誤差を検定統計量とした多重多変量 2 標本検定による遺伝子群解析を行った. 関連のある遺伝子をグループ化した遺伝子群の解析においては, 遺伝子群に含まれる遺伝子の発現パターンを多次元的に捉える必要があるため, 多変量検定を行う必要があることを述べ, SVM の分類後差を用いることで多変量検定を行えることを見た. また, 複数の遺伝子群の同時検定による多重検定の問題に対してはラベル並べ替え検定により対処した. しかしながら, 全てのラベルに対してナイーブに SVM 学習を行うと, 計算コストが膨大になる. そこで, MST によりスケジューリングされたパス追跡により SVM を学習する枠組みを提案した. 実際の遺伝子群データを用いた計算機実験により, 提案手法により効率的に検定を行うことができることを示した. また, いくつかの検定法を用いた遺伝子群解析の結果について考察を行った.

今後の課題として, より洗練された仮想ラベルの追加方法の考案, FWER, FDR などの基準を統一した解析結果の比較, SVM を用いた遺伝子群解析結果の医学生物学的知見からの考察などが挙げられる.

第 7 章

最近傍法を用いた array CGH データ解析

近年, ゲノムのコピー数異常を網羅的に検出するための技術として array CGH と呼ばれる技術が開発され注目されている. array CGH により検出されたコピー数データを用いることで, 疾患の原因となっている異常を領域的に同定することが可能になる. しかし, 異常領域を同定するには次元の異なるデータの統計量を比較する必要があり, 更にそれらを全ての可能なゲノム領域について行なわなければならない. 従って, 検定の多重性やそれらの相関, 更に適切な統計量を用いるなどの対処が不可欠となる. 本章では, 最近傍多変量検定を用いることで, これらの問題に対処する [75]. また, 最近傍多変量検定により計算コストも低く抑えることができることを示す.

7.1　array CGH データ解析

本節では, まず, 2 標本の array CGH のゲノム異常領域同定問題を定式化する. 問題設定を簡単にするために, 定式化においては 2 標本問題に限定するが, ここでの話題は, 多標本の問題へ直接的に拡張することができることに注意されたい. 定式化を行った後に, ゲノム異常領域同定問題を困難にしている 3 つの話題について述べる.

7.1.1　array CGH のゲノム異常領域同定問題の定式化

いま, n 症例の array CGH データがあり, それらは異なる 2 つのグループ (例えば, 健常者と癌患者, ある疾患の 2 種のサブタイプなど), C_1, C_2 から得られたものであるとする. n_1, n_2 をそれぞれ C_1, C_2 の症例数とする: $n_1 + n_2 = n$. さらに, c 番染色体のプローブ数を ℓ_c, $c = 1, \ldots, 22$ とし, 各々のプローブは染色体内で $1, \ldots, \ell_c$ とインデックス化されているものとする. 症例 i の c 番染色体の j 番目のプローブの $\log_2$-ratio を x_{icj}, $i \in \mathbb{N}_n$, $c \in \mathbb{N}_{22}$, $j \in \mathbb{N}_{\ell_c}$ と表記する. また, 症例 i の所属するグ

ループを y_i で表し, C_1 に対しては $y_i = 1$, C_2 に対しては $y_i = 2$ と表すこととする.

議論を簡単化するため, しばらくは症例 i の c 番染色体に関する異常領域同定のタスクを考えることとする: 症例 i の c 番染色体に含まれるプローブの $\log_2$-ratio は $x_{ic1}, x_{ic2}, \ldots, x_{ic\ell_c}$ で与えられる. 第5章で述べたように, 各プローブはゲノムの物理的なある領域から得られたものであるため, array CGH データは空間的な相関を持つ. 異常領域同定のタスクでは, $x_{ic1}, x_{ic2}, \ldots, x_{ic\ell_c}$ の全ての可能な部分領域を調べなければならない. これらの部分領域の総数は $\sum_{h=1}^{\ell_c} h = \frac{1}{2}\ell_c(\ell_c+1)$ で与えられる. 何故なら, 領域幅 ℓ_c の領域は1通り, $\ell_c - 1$ の領域は2通り, ..., 領域幅1の領域は ℓ_c 通り存在するためである. 今後は, 全ての染色体の全ての部分領域の総数を $M \equiv \sum_{c=1}^{22}\sum_{h=1}^{\ell_c} h = \frac{1}{2}\sum_{c=1}^{22}\ell_c(\ell_c+1)$ と表し, 各々の領域は $m = 1, \ldots, M$ とインデックス化されているものとする. また, $\mathcal{R}_m$ を, 領域 m を構成する遺伝子番号 c とプローブ番号 j のインデックスのペア (c, j) の集合として用い, そのサイズを $|\mathcal{R}_m|$ で表す.

7.1.2 ゲノム異常領域同定問題における多変量2標本検定

5.4節で触れたように, array CGH のゲノム異常領域同定問題は多重多変量2標本検定として定式化される. 本節では, 本問題における多変量2標本検定について詳しく述べる.

検出された異常領域の統計的信頼性 (例えば, p-値 など) を評価するためには, 統計的検定を行う必要がある. $\{\mathcal{R}_m\}_{m\in\mathbb{N}_M}$ の各々の領域に関して, その領域での2標本 C_1, C_2 の違いを定量化したい. この問題は, 2標本, $\{x_{icj}\}_{i\in\{1,\ldots,n_1|y_i=1\},(c,j)\in\mathcal{R}_m}$, $\{x_{icj}\}_{i\in\{1,\ldots,n_2|y_i=2\},(c,j)\in\mathcal{R}_m}$, を用いた $|\mathcal{R}_m|$-次元の多変量2標本検定として定式化される. ここでの多変量2標本検定の帰無仮説は, "n 症例全ての $|\mathcal{R}_m|$ 次元ベクトル $\{x_{icj}\}_{i\in\mathbb{N}_n,(c,j)\in\mathcal{R}_m}$ は共通の $|\mathcal{R}_m|$-次元多変量分布から独立に生成されたものである, すなわち, 独立同分布 (independently and identically distributed: i.i.d.) である", であり, 対立仮説は "2標本 $\{x_{icj}\}_{i\in\{1,\ldots,n_1|y_i=1\},(c,j)\in\mathcal{R}_m}$, $\{x_{icj}\}_{i\in\{1,\ldots,n_2|y_i=2\},(c,j)\in\mathcal{R}_m}$ はそれぞれ異なる2つの $|\mathcal{R}_m|$-次元多変量分布から独立に生成されたものである", となる[1] .

統計学の文脈では, パラメトリックな多変量検定とノンパラメトリックな多変量検定が研究されている. もし, 得られたデータが分散共分散行列の等しい多変量正規分布から生成されたものであれば, Hotelling T^2 検定 [53] が最も検出力の高い

[1]通常の統計的検定では, 例えば帰無仮説として, 2分布の平均が等しい ($\mu_1 = \mu_2$), 対立仮説として, 2分布の平均が異なる ($\mu_1 \neq \mu_2$) などのように数学的に表されるが, ここではノンパラメトリック検定を行うことを想定し, 若干抽象的な仮説を立てることとする.

検定であることが知られている. しかしながら, array CGH データの $\log_2$-ratio シグナルが多変量正規分布に従うとは限らないため, 本問題において Hotelling T^2 検定を用いることは適切でない. 一方で, ノンパラメトリック検定の多くは 2 つのデータ間の距離を用いて検定統計量を定義する. 例えば, 文献 [54] では, 2 つのデータ間の距離に基づいて最小全域木 (minimum spanning tree: MST) を構築することで, あるクラスのノンパラメトリックな単変量検定を多変量検定に拡張した. また, 最近傍検定 [55, 56] では, より単純な方法によりノンパラメトリックな多変量検定を行っている. これらのタイプの多変量検定を**多変量アプローチ**と呼ぶことにする.

しかしながら, 実際の応用において, これらの多変量アプローチが用いられることは少なく, より単純なアプローチが用いられる (そのようなアプローチを, 多変量アプローチと比較して**単変量アプローチ**と呼ぶことにする). $|\mathcal{R}_m|$-次元の 2 標本の違いを定量化するためには, 例えば $|\mathcal{R}_m|$ 個の単変量の検定統計量の平均を用いることができる. 通常の t-検定をこのような目的で用いる場合, $T_{\mathcal{R}_m} = \frac{1}{|\mathcal{R}_m|}\sum_{(c,j)\in\mathcal{R}_m} t_{cj}$ を多変量統計量とすることが最も単純な方法である. ここで, t_{cj} は単変量の 2 標本, $\{x_{icj}\}_{i\in\{1,\ldots,n_1|y_i=1\}}$, $\{x_{icj}\}_{i\in\{1,\ldots,n_2|y_i=2\}}$, の t-値 である. この他にも多くの単変量アプローチが可能である. 例えば, 文献 [74] では, t-値 の平均の代わりに, t-値 の最大値を用いる方法が提案されている: 領域 $\mathcal{R}_m$ に対する統計量を, $T_{\mathcal{R}_m} = \max_{(c,j)\in\mathcal{R}_m} t_{cj}$ と定義する.

単変量アプローチの利点は, 解釈のしやすさと計算の容易さである. もし, 本章で扱うタスクにこのような方法を用いるのであれば, まず初めに各プローブの t-値 を計算し, 次に各々の領域 $\mathcal{R}_m$, $m \in \mathbb{N}_M$ に含まれるプローブの t-値 の平均を計算し, その領域の多変量統計量として用いればよい. 一方で, もし変数間, すなわちプローブ間に相関がある場合, このような単変量アプローチは多変量アプローチと比較して検出力が劣る. というのも, このような方法は各々のプローブで独立に計算された統計量を平均化, あるいはその最大値を取っているため, プローブ間の相関を考慮していないためである. しかしながら, 通常多変量アプローチの計算量は, 単変量アプローチの計算量と比較して大きくなる. 加えて, 本章で扱うタスクでは, 多くの候補領域に対して並べ替え検定を行わなければならない. 多変量アプローチのこのような計算量的な問題が, 異常領域同定問題を考える際に大きな制限となっている.

7.1.3 ゲノム異常領域同定における多重検定

2標本で特徴の異なる異常領域同定問題では, 多数の統計的検定を同時に行うため, 多重検定の問題に対処する必要がある. 5.5.3節で触れた通り, 多重検定補正を行う際には検定統計量の相関を考慮しなければならない.

本章で取り扱う問題では, 多重性が大きく ($M \approx 10^5$ 個の検定), 多くの領域が重なり合っているため, 検定統計量が相関を持っている. 従って, 本章においてもラベル並べ替えにより多重検定補正および検定統計量の帰無分布推定を行う.

7.1.4 異なる領域幅の検定統計量の比較

ゲノム異常領域同定問題では, 様々な幅のゲノム領域の統計的検定を行わなければならない. 極端な例では, 1つのプローブからなるゲノム領域の検定統計量と染色体全体からなる領域の検定統計量とを比較する必要がある. 従って, ここで用いる検定統計量は異なる領域幅で比較可能なものでなければならない. 言い換えれば, 次元数 $|\mathcal{R}_m|$ に依存しないような検定統計量を用いる必要がある. 多くの多変量統計量はその次元数に依存したものである. 例えば, t-値の平均を用いた統計量 $T_{\mathcal{R}_m} = \frac{1}{|\mathcal{R}_m|}\sum_{(c,j)\in\mathcal{R}_m} t_{cj}$ の帰無分布は次元数 $|\mathcal{R}_m|$ により異なる: 帰無分布の分散は領域幅が大きいほど小さくなる. 異なる次元数のデータから得られたスケールの異なる統計量の正規化や標準化は, 各変数がi.i.d. であるか, 変数間の相関が既知の場合にのみ可能であることに注意されたい.

7.2 最近傍多変量検定

本節では, 特徴の異なるゲノム領域を同定する問題に対して最近傍多変量検定を導入する. 前述の通り, この問題を考える際には次のような要件を満たす検定統計量を用いる必要がある:

1. プローブ間の相関を考慮できる検定法であること, すなわち, 連続する複数のプローブ多次元的なパターンを反映した検定統計量を用いること,
2. 領域幅 $|\mathcal{R}_m|$ に依存しない統計量であること,
3. ラベル並べ替え演算をある程度低い計算コストで実行可能であること.

本節では, 最近傍多変量検定を用いることでこれらの要請を満たすことができることを見る.

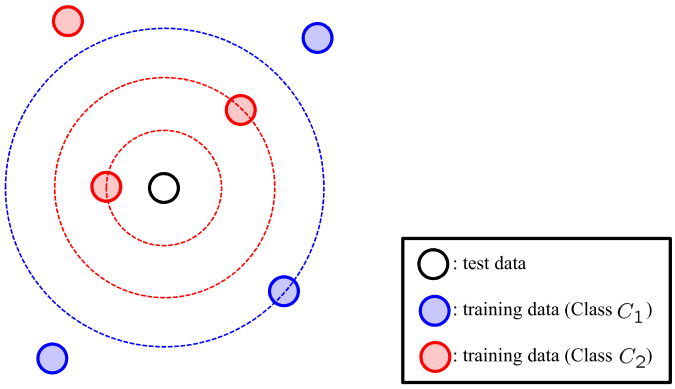


図 7.1: k-NN の例: ($k = 3$): この場合, テストデータに最も近い 3 つのデータ点のうち, 2 点がクラス C_2 に属するため, テストデータはクラス C_2 に分類される.

7.2.1 k-最近傍法

本節では k-最近傍法 (k nearest neighbor 法: k-NN) [73] について説明する. k-NN 法では, 特徴空間においてテストデータに最も近い k 個のデータ点のうち, 多数派のクラスにテストデータを分類するという単純な分類アルゴリズムである. 図 7.1 に k-NN 法の例を示す.

k-NN 法は単純でありながら良い性能を示すため, パターン認識を始め, 多くのアプリケーションで利用されている. ここでは, 検定統計量として k-NN 法の分類誤差を用いることで, 多変量検定を実現する.

7.2.2 最近傍多変量検定

各領域 $\{\mathcal{R}_m\}_{m\in\mathbb{N}_M}$ において, $\{x_{icj}\}_{i\in\mathbb{N}_n,(c,j)\in\mathcal{R}_m}$ を用いて k-NN の 1 つ抜き交差検証誤差 (leave-one-out cross-validation error: LOOCV error) を計算する. もし, LOOCV error が小さければ, それは領域 $\mathcal{R}_m$ において C_1 と C_2 に違いが見られることを示唆しており, 一方, LOOCV error が大きければそれほど違いがないということを示唆する. 異常領域をランク付けしたいだけであれば, M 個の領域を LOOCV error に基づいてを昇順にソートすればよい. しかしながら, 我々は各領域の検定統計量の統計的信頼性も評価したい. 故に, LOOCV error の帰無分布を計算する必要がある. 前節で述べたように, 帰無分布はラベル並べ替えにより推定される. その際, k-NN の LOOCV error を多数回計算しなければならず, もしナイーブにこれを行うと計算コストが膨大になる. しかし, ラベル並べ替え k-NN の LOOCV error を求める際

に, 簡単なトリックを用いることで, 効率的な計算が可能である:

最近傍多変量検定

入力: $\log_2$-ratio $\{x_{icj}\}_{i\in\mathbb{N}_n, c\in\mathbb{N}_{22}, j\in\mathbb{N}_{\ell_c}}$, ラベル $\{y_i\}_{i\in\mathbb{N}_n}$, ラベル並べ替え回数 B, 並べ替えられたラベル群 $\{\{y_i^{(b)}\}_{i\in\mathbb{N}_n}\}_{n\in\mathbb{N}_B}$, 近傍点数 k, 距離関数 d, 有意性の基準 θ, 最大異常領域数 γ.

Step 1-1: 各染色体 $c = 1, \ldots, 22$ に対して, 全ての部分領域を $\mathcal{R}_m$, $m = 1 \ldots, M$ と順序づける.

Step 1-2: 領域 $\{\mathcal{R}_m\}_{m\in\mathbb{N}_M}$ に対して, ラベル $\{y_i\}_{i\in\mathbb{N}_n}$ を用いて k-NN LOOCV error $\{s_m^*\}_{m\in\mathbb{N}_M}$ を計算する.

Step 2: $m \leftarrow 1$ として以下を繰り返す:

Step 2-1: $b \leftarrow 1$ とする.

Step 2-2: ラベル $\{y_i^{(b)}\}_{i\in\mathbb{N}_n}$ を用いて, 領域 $\mathcal{R}_m$ の k-NN LOOCV error $s_m^{(b)}$ を計算する.

Step 2-3: もし $b < B$ ならば $b \leftarrow b + 1$ として Step 2-2 へ. $b = B$ かつ $m < M$ であれば, $m \leftarrow m + 1$ として Step 2-1 へ. $b = B$ かつ $m = M$ であれば Step 3 へ.

Step 3: 各領域 $\{\mathcal{R}_m\}_{m\in\mathbb{N}_M}$ に対して,

$$
\begin{aligned}
p\text{-value}_m &= \frac{\sum_{b=1}^{B} \mathcal{I}(s_m^{(b)} \le s_m^*)}{B}, \\
FWER_m &= \frac{\sum_{b=1}^{B} \mathcal{I}\{\min_{m'} s_{m'}^{(b)} \le s_m^*\}}{B}, \\
FDR_m &= \frac{\sum_{b=1}^{B}\sum_{m'=1}^{M} \mathcal{I}\{s_{m'}^{(b)} \le s_m^*\}}{B\sum_{m'=1}^{M} \mathcal{I}\{s_{m'}^* \le s_m^*\}}, \\
q\text{-value}_m &= \min_{k\in\{k|k\ge \mathrm{order}(m), k\in\mathbb{N}_{\mathbb{M}}\}} FDR^{(k)},
\end{aligned}
$$

を計算する. ここで, $\{FDR^{(k)}\}_{k\in\mathbb{N}_{\mathbb{M}}}$ は $\{p\text{-value}\}_{m\in\mathbb{N}_{\mathbb{M}}}$ の大小によってあらかじめ昇順にソートされた $\{FDR_m\}_{m\in\mathbb{N}_{\mathbb{M}}}$ のリストとする. また, $\mathrm{order}(m)$ は遺伝子群 m の順序とする.

Step 4: Step 3 で計算された基準が閾値 θ よりも小さい遺伝子領域を選択する. もし, 領域が重複したものがある場合, 基準の小さいものを選択する. また, もし γ より多くの領域がある場合, 上位 γ 個を選ぶ.

出力: 有意水準 θ に関して有意な, 異常領域群.

k-NN による分類において, 近傍数 k を事前に決定する必要があり, この値は分類性能に大きく影響する. 一方で, k-NN を統計的検定に用いる場合, いくつかの k, 例えば $k = 1, 3, 5$ の, k-NN LOOCV error の平均などを用いることができる.

7.2.3 最近傍多変量検定の利点

最近傍法は単純な分類アルゴリズムであるが, 現実の応用においてしばしば良い性能を示す. 例えば, マイクロアレイ発現量データを用いた癌の分類問題において, 決定木, SVM などの他の分類アルゴリズムの中で, 最近傍法が最も良い性能を示したと報告されている [72]. このことは, k-NN LOOCV error は, 各候補領域の分類性能を量る上で有効な指標であり, 更に, 最近傍多変量検定は異常領域同定において高い検出力を持ち得ることを示唆している. 従って, 本アプローチは前節で述べた (1) の要請を満たしていると考えられる.

加えて, (2) の要請も満たす. 何故なら, k-NN LOOCV error 自体は領域幅 $|\mathcal{R}_m|$ に依存しないためである. 一方で, LOOCV error の不利な点として, 離散的な値しか取り得ない点がある: データ数が n のとき, LOOCV error の取り得る値は高々 n 通りしかない. このことは, 領域の分類性能の微妙な違いを表現するには粒度が粗すぎ, LOOCV error の等しい領域が多数現れるということが起こり得ることを意味している. このような場合, 例えば, より領域幅が小さいものを選択する等のヒューリスティックを用いる必要がある.

最後に, 最近傍多変量検定が要請 (3) を満たすことを強調しておく. 一般に, 多変量アプローチは単変量アプローチと比較して, 計算コストが高くなる. 従って, 本章の問題のように, 候補領域が多く, 多数のラベル並べ替え演算が必要なケースでは適切でない. しかしながら, 我々の方法では, 単純なトリックを用いることにより, 計算コストを大幅に削減可能である. 最近傍法を用いた異常領域同定問題では, M 個の領域全てに対して n 症例の k-最近傍点を求める必要があり, この計算コストが非常に高い. これを並べ替えられた B 通りのラベルに対して繰り返さなければならず, 毎回 k-最近傍点を, 距離の再計算により求めると, 計算コストが膨大になる: 症例数 n, 近傍数 k, 並べ替え回数 B, 領域数 M に対して, オーダ $\mathcal{O}(kn^3BM)$.

この問題を解決するために, 我々は Step 1 を Step 1-1, Step 1-2 と分けた. Step 1-1 においてあらかじめ各々の症例に対する k-近傍表 $\{T_m\}_{m\in\mathbb{N}_M}$ を求め, 記憶しておけば, Step 1-2 では LOOCV error s^*_m をオーダ $\mathcal{O}(kn)$ で計算可能である. Step 2 の並べ替えラベル群 $\{\{y_i^{(b)}\}_{i\in\mathbb{N}_n}\}_{b\in\mathbb{N}_B}$ に基づく統計量 $\{s_m^{(b)}\}_{m\in\mathbb{N}_M,b\in\mathbb{N}_B}$ に対して, Step 1-1, 1-2 と同様の手順を踏む必要はない, すなわち, $\{T_m\}_{m\in\mathbb{N}_M}$ を**再度計算する必要はな**

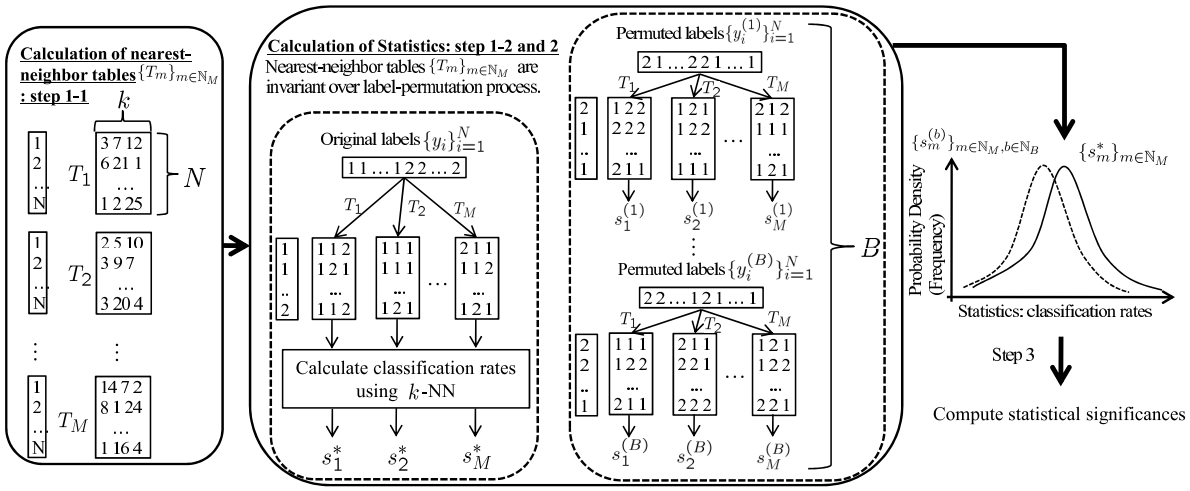


図 7.2: 最近傍多変量検定の概念図: Step 1-1 では, k-近傍表を作成する. この表はラベル並べ替えに対して不変であるため, Step 2 ではこれを再計算する必要がない. Step 3 では推定された帰無分布に基づいて統計的信頼性を評価する.

い. 言い換えれば, Step 1-1 において計算された k-近傍表 $\{T_m\}_{m\in\mathbb{N}_M}$ は, **ラベル並べ替えに対して不変である**. この事実を用いれば, 各並べ替えプロセスにおいて, LOOCV error を計算する際に Step 1-2 のみを行えば良く, ここでの計算オーダは $\mathcal{O}(kn)$ となる. 従って, その都度 $\{T_m\}_{m\in\mathbb{N}_M}$ を計算する場合のオーダ $\mathcal{O}(kn^3BM)$ に対して, $\mathcal{O}(n^2)+\mathcal{O}(knBM)$ と計算量が削減される. その代わり, 我々のアルゴリズムではサイズ Mnk の整数型の記憶領域を必要とする. しかしながら, k-近傍表の記憶や計算に対してビット演算を用いることで, 計算量, 記憶領域の両面での改善が可能である.

図 7.2 に最近傍多変量検定の概念図を示す.

7.3 array CGH データ解析への応用

本節では, 先に紹介した array CGH データに対して最近傍多変量検定によるゲノム異常領域同定を行う. 次に, 検出された異常領域を用いた未知データの癌のサブタイプ分類を行う.

7.3.1 データセットと前処理

本節の実験では, リンパ腫患者から得られた 75 症例の BAC array CGH に対して, 最近傍多変量検定を適用する. これらのサンプルは愛知県がんセンターにおいて収

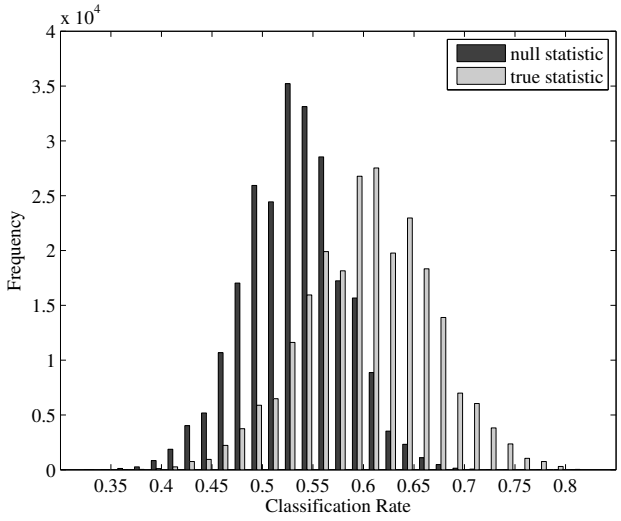
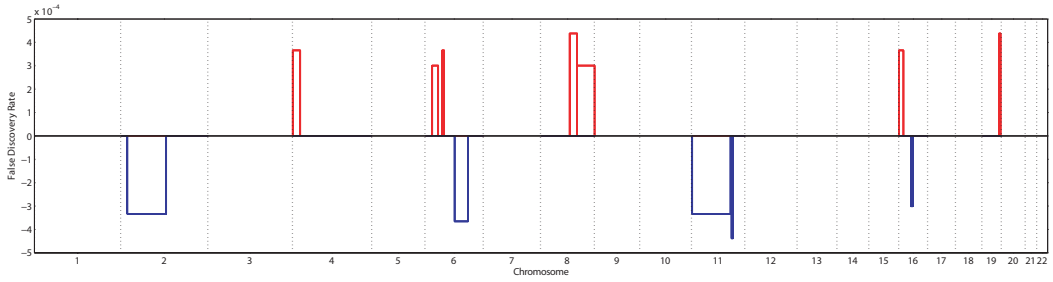


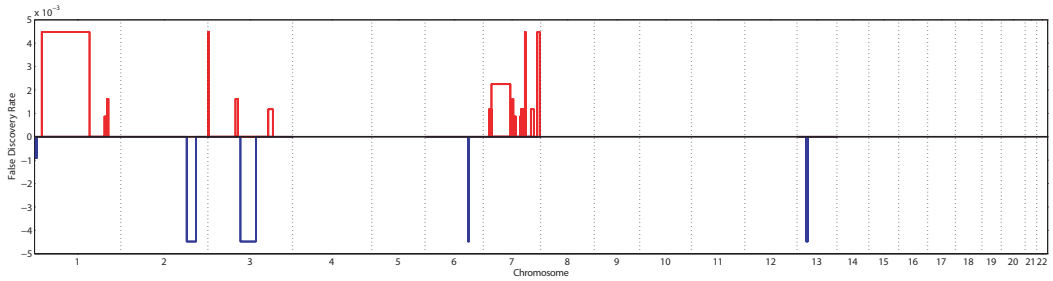
図 7.3: 検定統計量および帰無統計量のヒストグラム: 帰無分布は 0.5 付近が最も出現頻度が高いが検定統計量はそれよりもやや上方に偏っていることがわかる.

集, 調査されたものである [50, 51, 52]. 75 症例のうち, 46 症例は (病理学者によって) びまん性大細胞リンパ腫 (diffuse large B-cell lymphoma: DLBCL) と診断された症例であり, 29 症例はマントル細胞リンパ腫 (mantle cell lymphoma: MCL) と診断された症例である. また, DLBCL の 46 症例は, 活性化 B 細胞型 (activated B-cell: ABC)18 症例, 胚細胞 B 細胞型 (germinal center B-cell: GCB)28 症例に分けられる. ここでは, 2 つのタスクについて実験を行う. 1 つ目のタスクは DLBCL46 症例と MCL29 症例の間で特徴の異なるゲノム異常領域を同定するもの, 2 つ目のタスクは ABC18 症例と GCB28 症例の間で特徴の異なるゲノム異常領域を同定するものである. 各アレイは 2,035 BAC プローブからなる.

各アレイの $\log_2$-ratio はメジアンが 0 となるように標準化した. この標準化処理により, 多くの増幅を含むアレイは減少方向に, 多くの欠損を含むアレイは増加方向に, 若干の偏りが生じる. gain と loss のそれぞれに対する異常領域を個別に同定するために, 元の $\log_2$-ratio シーケンスを gain $\log_2$-ratio シーケンスと loss $\log_2$-ratio シーケンスに分割して解析を行った. gain $\log_2$-ratio シーケンスにおいては, 値が 0.1 より小さいものを $[0.0,\ 0.1]$ のランダムな値で置き換えた. また, loss $\log_2$-ratio シーケンスでは, 値が -0.1 より大きいものを $[-0.1,\ 0.0]$ のランダムな値で置き換えた.



(a)DLBCL vs. MCL



(b)ABC vs. GCB

図 7.4: 検出された異常領域: 横軸は染色体を, 縦軸は FDR の値を表している. また, 赤線は gain 異常領域であり, 青線は loss 異常領域を表している. loss 異常領域の FDR は, -FDR と負の値としてることに注意されたい.

7.3.2 2標本間で特徴の異なる領域の同定

本節の実験では有意性の基準に FDR を用い, その閾値を, DLBCL vs. MCL タスクでは 0.0005 に, ABC vs. GCB タスクでは 0.005 と設定した. ラベル並べ替え回数は $B = 1,000$ とした.

ゲノム異常領域同定では, まず Step 1-1 において k-近傍表 $\{T_m\}_{m\in\mathbb{N}_M}$ を作成する. 次に Step 1-1 において (正しい) ラベルでの検定統計量 $\{s^*_m\}_{m\in\mathbb{N}_M}$, Step 2 において並べ替えられたラベルでの検定統計量 $\{s^{(b)}_m\}_{m\in\mathbb{N}_M,b\in\mathbb{N}_B}$ を計算する. 図 7.3 に検定統計量 $\{s^*_m\}_{m\in\mathbb{N}_M}$, 帰無統計量 $\{s^{(b)}_m\}_{m\in\mathbb{N}_M,b\in\mathbb{N}_B}$ のヒストグラムを示す. なお, **ここでは検定統計量は分類誤差ではなく, 分類精度とした**. すなわち, 大きいほど分類性能が高い (有意である). また, 帰無統計量は $M \times B$ 個存在するため, 図の見やすさのためランダムに選択した M 個を掲載した.

次に, 推定された帰無分布を用いて, M 個の領域それぞれの p-値, FWER, FDR, q-値 等を計算し, それらの基準が閾値 θ よりも小さいものを異常領域として出力す

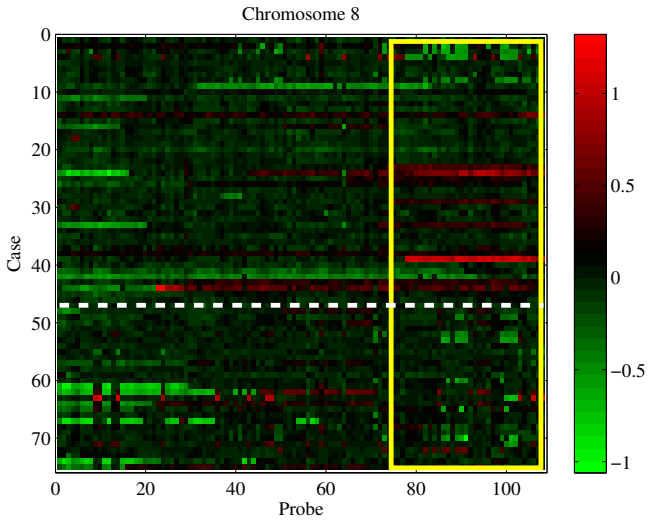


図 7.5: 検出された異常領域の例 (8 番染色体): DLBCL vs. MCL タスクにおいて, 最も FDR が小さかった異常領域を黄色の長方形で示す. 横軸はプローブを表し, 縦軸は症例を表す. また, 白色の点線より上側が DLBCL 症例, 下側が MCL 症例である. $\log_2$-ratio の値が大きいプローブを赤色で, 小さいプローブを緑色で表現してある.

る. 図 7.4(a) に DLBCL vs. MCL タスクにおいて検出された異常領域を, (b) に ABC vs. GCB タスクにおいて検出された異常領域を示す.

これらの結果から, 我々の手法は数個のプローブからなる小さな領域, 染色体の大部分からなるような大きな領域ともに検出していることがわかる. 図 7.5 に検出された異常領域の $\log_2$-ratio のヒートマップを示す. 異常領域 (図中の黄色実線の長方形で囲まれた領域) において, MCL 症例と比較して DLBCL 症例はより gain の傾向 (より赤に近い色) を示していることが見て取れる. このような結果は, 我々のアプローチが異種の癌の違いを表すような重要なゲノム異常領域を検出する能力を有していることを示唆している.

7.3.3 異常領域を用いた癌の分類

本節では, 検出された異常領域を用いた癌の分類タスクの実験を行う. 分類性能の評価には LOOCV を用いた. まず, あるアレイを検証用に抜いておき, 残ったアレイで異常領域を同定する. 次に, 検証用に抜いたアレイが分類される癌タイプ (DLBCL と MCL あるいは ABC と GCB) の事後確率を**投票**を用いて推定する. 検出された各々の異常領域は 1 票持っており, 各領域に含まれるプローブの $\log_2$-ratio を用いた距離に基づいて k-NN 分類を行うことで分類すべき癌タイプを決定し 1 票を

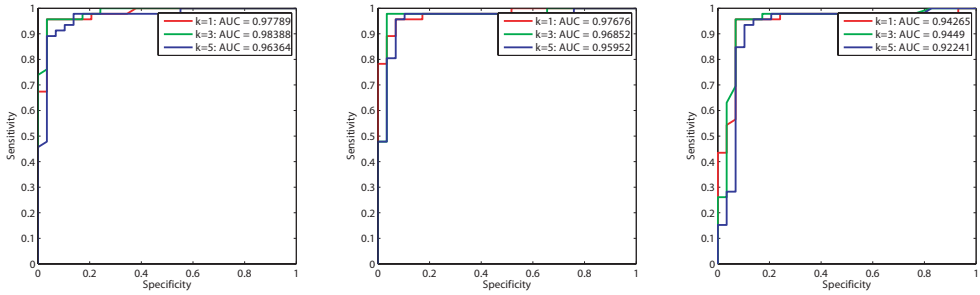


図 7.6: DLBCL vs. MCL タスクにおける ROC 曲線 ($k = 1, 3, 5$): 左から順に最近傍多変量検定, ADM, CLAC

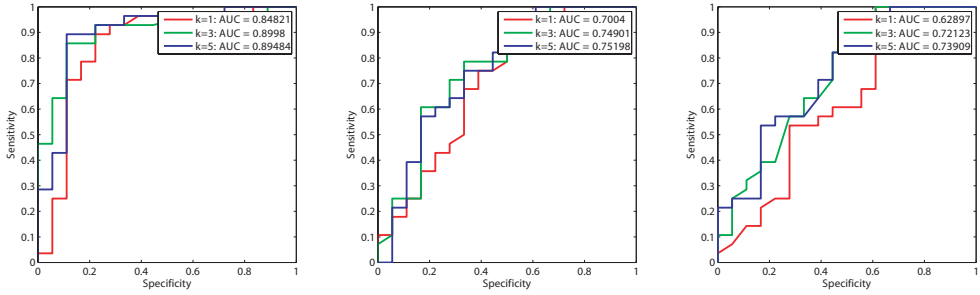


図 7.7: ABC vs. GCB タスクにおける ROC 曲線 ($k = 1, 3, 5$): 左から順に最近傍多変量検定, ADM, CLAC

入れる. それを検出された全ての異常領域について行い, 票の多い癌タイプに検証用アレイを分類する. 例えば, もし 10 個の異常領域が検出され, 3 領域で DLBCL, 7 領域で MCL と分類されたとすると, そのアレイが DLBCL である確率は 0.3, MCL である確率は 0.7 と計算される.

我々は, 最近傍多変量検定により検出された異常領域の分類性能を, ADM [57] および CLAC [33] と比較する. ADM は array CGH データ解析に用いられる標準的な手法であるが, この手法は 1 つのアレイの異常領域を同定するために提案された手法である. 従って, ここでは, 初めに各々のプローブにおいて 2 標本の t-値を計算し, それらの t-値を ADM への入力として与えることで本章の問題に適用した. また, CLAC も ADM と同様に 1 つのアレイの異常領域を同定する目的で提案されたものである. そこで, ADM と同様, CLAC への入力も各々のプローブの t-値とした. なお, ADM, CLAC は単変量アプローチであることに注意されたい. ここでの癌タイプ分類タスクでは, FDR の閾値 θ は DLBCL vs. MCL タスクでは 0.00125, ABC vs. GCB

タスクでは 0.0125 と設定した. ADM は FDR を計算する枠組みが与えられていないため, ADM の有意水準は Bonferroni 補正を用いた p-値とした. すなわち, 本問題においては, 候補領域の総数が 118,328 であるため, ADM の有意水準は, DLBCL vs. MCL タスクでは 0.00125 / 118328, ABC vs. GCB タスクでは 0.0125 / 118328 とした.

分類精度は ROC 曲線および AUC により評価する. ROC(receiver operating characteristic) 曲線は 2 クラス分類器において, 閾値を変化させたときの sensitivity vs. specificity を表したものである. DLBCL vs. MCL タスクでは DLBCL を陽性, MCL を陰性とし, ABC vs. GCB タスクでは ABC を陽性, GCB を陰性とした. 図 7.6, 7.7 において, 横軸は偽陽性 (false positive), 縦軸は真陽性 (true positive) を表している. AUC(area under curve) は ROC 曲線の下側の面積を表しており, 様々なコスト (偽陽性と偽陰性の相対コスト) での分類精度を評価するために用いられる. ROC 平面において, 点 (0, 1) が完全な分類器 (false positive, false negative がない) を与える. 従って, AUC の最大値は 1 となる. 結果を図 7.6, 7.7 に示す. なお, 前述した分類過程における k-NN の近傍数 k は $k = 1, 3, 5$ の 3 通りで行った.

DLBCL vs. MCL タスクでは, 最近傍多変量検定, ADM, CLAC 全てがかなり良い性能を示している. 一方で, ABC vs. GCB タスクでは, DLBCL vs. MCL タスクと比べ全ての手法で分類性能が下がっている. これは ABC, GCB というサブタイプの差異が医学生物学的に見ても曖昧であることに関連する. しかしながら, どちらのタスクにおいても我々の手法が, 他の手法と比較して良い性能を示していることがわかる. 特に, ABC vs. GCB タスクでの分類性能の違いは顕著である. これらの結果から, 最近傍多変量によるアプローチは, 2 種, あるいはより多くのサブタイプ間で特徴の異なる領域を同定できる可能性を持っていることがわかる.

7.4 まとめ

本章では, 最近傍多変量検定を用いた array CGH データのゲノム異常領域同定を行った. ゲノムコピー数を定量化したデータである array CGH データ解析においては, 1 疾患 1 症例に特異的な異常領域を同定するタスク, 1 疾患複数症例に共通の異常領域を同定するタスク, 2(あるいは複数) 疾患で特徴の異なる異常領域を同定するタスクがある. 本章では, 重要であるがあまり研究がなされていな 3 つ目のタスクに取り組んだ. このタスクにおいては, 多変量検定の必要性, 多重検定への対処, 次元の異なるデータから得られた統計量の比較, という 3 つの問題があるが, これらの問題に対して, 最近傍法の分類誤差を検定統計量とし, ラベル並べ替え検定

を行うことにより対処した. array CGH データ解析では, 全ての可能な部分領域に対して統計量を計算する必要があり, 計算コストが心配されるが, ラベル並べ替えに対して最近傍表が不変であることを利用することで, 効率的な統計量の計算が可能となる. 実際の array CGH データを用いた計算機実験により, 提案手法が疾患分類に有用な異常領域を同定できることが確認された.

今後の課題として, さらに多くの実データを用いた実験による, 提案手法の有用性の評価, ビット演算などによるアルゴリズムの更なる効率化などが挙げられる.

第 8 章

おわりに

本研究では, 統計的機械学習の 1 分野である密度推定問題, 特に EM アルゴリズム, 変分ベイズ法による混合正規分布推定の局所最適性の問題に取り組んだ. また, 統計的機械学習の応用例としてマイクロアレイデータ解析を行った. 特に, 遺伝子群解析, ゲノム異常領域同定問題に対し, サポートベクトルマシンおよび最近傍分類器の分類誤差を用いたノンパラメトリック検定により各々の問題を取り扱った.

第 1 章では, 本研究の背景及び目的について述べた. また, 本稿の構成について説明をした.

第 2 章では, 第 3 章, 第 4 章で必要となる正規分布及び混合正規分布の定式化を行った. 正規分布は, 確率密度分布の中でも基本的かつ重要な分布であり, 様々な現実データの分布に適用されている. また, 正規分布の重み付け和として表される混合正規分布は, 複雑な形状の分布を表現することができ, パターン認識などの多くの現実問題に適用されている. この章では, これらの分布を可視化することで, 分布の形状の具体例を示した.

第 3 章では, EM アルゴリズムを用いた混合正規分布推定の局所最適性について考察し, 効果的な初期値生成法を提案した. まず, 最尤法による正規分布推定, EM アルゴリズムによる混合正規分布推定について説明を行い, 定式化した. EM アルゴリズムは混合正規分布推定問題における代表的な解法であるが, その目的関数が多峰であるが故に, 推定されるパラメータが初期値に依存するという局所最適性の問題を有する. そこで, この問題に対する対処法の 1 つである DAEM アルゴリズムを紹介し, その挙動及び特徴について述べた. DAEM アルゴリズムは, 統計力学のアナロジーを用いて, 目的関数を単純な形状から元の形状へ変化させつつ推定を行うことで, 局所最適性の問題に対処している. しかし, 探索点が鞍点へトラップされた場合の対処が問題となる. そこで, DAEM アルゴリズムの枠組みより得られる原始初期点に着目し, 原始初期点における (元の) 目的関数の Hesse 行列を調べ

ることで, 解析的かつ効率的な初期値生成法提案した.

第4章では, 第3章で定式化した多方向探索アルゴリズムを変分ベイズ法を用いた混合正規分布推定へも適用した. ベイズ推定及び変分ベイズ法は, 最尤法, EMアルゴリズムとは異なるアプローチの混合正規分布の推定法であり, 近年EMアルゴリズムと並んで多くのアプリケーション分野で用いられている. 変分ベイズ法を用いた混合正規分布推定においても, EMアルゴリズムと同様, 局所最適性の問題を有しており, この問題への対処法の1つとしてDAEMアルゴリズムのアイディアを用いたDAVB法が提案されている. そこで, 本研究ではEMアルゴリズムにおいて提案した多方向探索アルゴリズムの変分ベイズ法への適用を試みた. 計算機実験により, 本手法は変分ベイズ法においても良い性能を示すことが確認された.

第5章では, 第6章, 第7章における統計的機械学習を用いたマイクロアレイデータ解析の基礎となるマイクロアレイデータ, 多変量2標本検定, 多重検定とラベル並べ替え検定について説明を行った. 本研究で取り扱う遺伝子群解析とゲノム異常領域同定問題は, ともに多重多変量2標本検定として定式化される. ここでは一般的な2クラス分類器の分類誤差を検定統計量として用いた多変量2標本検定の定式化を行った. また, 複数の検定を同時に行う際に生じる多重検定の問題について言及し, その対処法であるラベル並べ替え検定の説明を行った. さらに, 検定の多重性を考慮した有意性の基準であるFDR, FWER, q-値を紹介した.

第6章では, サポートベクトルマシンの分類誤差を検定統計量として用いた遺伝子群解析について考察した. 遺伝子群解析においては, 個々の遺伝子の発現パターンではなく, 遺伝子群に含まれる複数の遺伝子の発現パターンを多次元データとして解析することが望ましい. 本研究では, サポートベクトルマシンの分類誤差を検定統計量として用いることで, 多変量2標本検定を行った. しかしながら, 分類誤差の帰無分布推定及び多重検定への対処としてラベル並べ替え検定を用いることで, SVM学習の計算コストが高くなる問題が生じる. この問題に対しては, 最小全域木(MST)によりスケジューリングされたパス追跡を用いることで対処した. 並べ替えラベル群のパス追跡を実現するため, 実数ラベルを許容するよう最適化問題を緩和し, パス追跡を再定式化した. パス追跡を行う場合, 最適解が大きく異なると非効率となる. そこで, 最適解の距離をラベル間の距離で測り, 並べ替えラベル群のMSTを構築することで, パス追跡を行うラベル順序を最適化した. この際, 仮想ラベルを導入することで, 並べ替えラベル群の総距離をより小さくなるようMSTを拡張した. 計算機実験により, 並べ替えラベルに対してナイーブにSVMを学習するよりも計算コストが小さくなることが確認された. また, いくつかの遺伝

子群解析手法を比較することで, SVM による遺伝子群解析結果の特徴を検証した.

第 7 章では, 最近傍多変量 2 標本検定による array CGH データ解析を行った. ここでは, 異なる 2 標本, 例えば, 健常者と癌患者, ある疾患の 2 種のサブタイプ, ある薬剤に対する反応の有無など, でゲノムコピー数の特徴の異なる領域を同定する問題を考えた. この問題においては, 領域内のプローブを多次元データとして取り扱うこと, 領域幅の異なる領域の比較, 多数の候補領域を同時に検定することによる多重検定の問題に対処する必要があった. 本研究では, 最近傍法の分類誤差を検定統計量として用いることで第 1, 第 2 の問題に対処した. 第 3 の問題についてはラベル並べ替え検定により対処した. ラベル並べ替え検定では, 多数の並べ替えラベルに対する統計量を計算しなければならず, 通常計算コストが高くなるが, 最近傍多変量検定においては, ラベル変化に対して最近傍表が不変であることを利用することで, 計算コストを削減することができた. また, 実際の array CGH データを用いた計算機実験により, 様々な領域幅の異常領域の検出が可能であること, また, 検出された異常領域を用いたサブタイプ分類において, よい分類性能を示すことが確認された.

最後に本研究の今後の課題について触れる. まず, 本研究では, EM アルゴリズム, 変分ベイズ法による混合正規分布推定を取り扱った. ここでは, 原始初期点を用いた多方向探索法を提案したが, 原始初期点が大域的最適解への経路を持つかどうか, すなわち, 多方向探索法が大域的最適性を持ちうるかどうかはまだ未検証である. この問題に対する理論的な検証を行うことが重要な課題であると考えられる. しかしながら, 多方向探索アルゴリズムは, 他のモデルへも適用可能である. 他の混合モデルや回帰モデル, あるいはニューラルネットワークなどの別の枠組みへの応用も考えられる. また, 現実のアプリケーションへの応用も課題の一つである. EM アルゴリズム, 変分ベイズ法は音声認識などの信号処理, 画像処理, ネットワークのクラスタリング, バイオインフォマティクスなど様々な場面で用いられる. 本研究で提案した, 多方向探索アプローチをこれらのアプリケーション分野に応用し, その有用性や問題点を検証することも重要な課題と考えられる. 一方, マイクロアレイデータ解析においては, より多くの現実データへの適用および考察が考えられる. 本研究では, サポートベクトルマシンや最近傍法によるマイクロアレイデータ解析タスクの枠組みを導入し, その有効性を少数のデータセットにおいて検証するに留まった. 今後は, さらに多くの現実データを用いて解析を行い, 解析結果を医学, 生物学的見地から考察を行うことが必要であると考えられる. また, これらの手法を様々な場面で利用してもらえるよう, ソフトウェアとして公開するこ

とも今後の課題である.

謝辞

本研究を進めるにあたり, 多くの方々のご支援とご協力をいただきました. ここに感謝の意を表すとともに, お礼申し上げます.

名古屋工業大学大学院工学研究科創成シミュレーション工学専攻教授 杉山 勝先生には, 本論文の主査として多大なご尽力をいただき, 多数の有益なご意見, ご指摘を賜りました. 心からお礼申し上げます.

名古屋工業大学大学院工学研究科情報工学専攻教授 犬塚 信博先生, 同 和田山正先生, 同大学大学院工学研究科創成シミュレーション工学専攻教授 徳田 恵一先生には,　本論文の副査としてご尽力をいただき, 貴重なご意見を賜りました. 厚くお礼申し上げます.

名古屋工業大学大学院工学研究科創成シミュレーション工学専攻准教授 竹内一郎先生には, 博士後期課程に進学して以降 3 年間に渡り, 研究のご指導のみならず, 共同研究を行う機会や, 勉強会などを通して, 関連分野の幅広い知識をご教授頂きました. また, 本論文の執筆に関する懇切丁寧な指導をして頂きました. 更に, 卒業後の進路についても親身になってご相談に乗っていただきました. 心から感謝いたします.

中部大学工学部情報工学科教授 中野良平先生には, 本研究の機会を与えていただき, 研究の進め方から論文の書き方に至るまで, 懇切丁寧にご指導いただきました.

東京工業高等専門学校情報工学科講師 北越大輔先生には, 公私に渡って多大なご支援をいただきました. 心から感謝いたします.

名古屋工業大学 中野・竹内研究室所属の皆様には, 研究についてのみならず, 研

究室での日々の生活においても多大なご協力をいただきました. 特に, 博士後期課程の先輩である棚橋裕輔氏には, 同じ博士後期課程の学生として, 様々なご相談に乗っていただきました. 厚くお礼申し上げます.

最後に, 今日まで暖かく見守り激励してくださった友人, 知人の皆様, そして長きにわたる学生生活を支えてくださった家族に深く感謝いたします.

2010 年 12 月

石川 勇太

参考文献

[1] E. Parzen. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, Vol. 33, No. 3, pp.1065–1076, 1962.

[2] K. Fukunaga, and L. Hostetler. Optimization of k nearest neighbor density estimates. *IEEE Transaction on Information Theory*, Vol. 19, Issue 3, pp.320–326, 1973.

[3] G. J. McLachlan, and D. Peel. Finite Mixture Models. Wiley, 2000.

[4] J. Aldrich. R.A. Fisher and the Making of Maximum Likelihood 1912–1922. *Statistical Science*, Vol. 12, No. 3 pp.162–176, 1997.

[5] J. O. Berger. Statistical Decision Theory and Bayesian Analysis. *Springer Series in Statistics*, Springer-Verlag, 1985.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B, 39, 1-38, 1997.

[7] G. J. McLachlan, and T. Krishnan. The EM Algorithm and Extensions. Wiley, 1997.

[8] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[9] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research*, 1996.

[10] W. K. Hastings. Monte Carlo Methods Using Markov Chains and Their Applications. *Biometrika*, Vol.57, No.1, pp.97–109, 1970.

[11] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, Vol.11, Issue 2, pp.271-282, 1998.

[12] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Split and Merge EM Algorithm for Improving Gaussian Mixture Density Estimation. *Journal of VLSI Signal Processing*, 26, pp.133–140, 2000.

[13] K. Katahira, K. Watanabe and M. Okada. Deterministic Annealing Variant of Variational Bayes Method. *Journal of Physics: Conference Series*, 95, 012015, 2008.

[14] Z. Zhang, B. T. Dai, and A. K. H. Tung. Estimating Local Optimums in EM Algorithm over Gaussian Mixture Model. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[15] I. Sato, K. Kurihara, S. Tanaka, H. Nakagawa, and S. Miyashita. Quantum Annealing for Variational Bayes Inference. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[16] 永田賢二, 片平健太郎, 岡ノ谷一夫, 岡田真人. 変分ベイズ法における確定的アニーリングとハイパーパラメータの部分最適化について. 情報論的学習理論テクニカルレポート, pp.144–151, 2009.

[17] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:5235:467–470, 1995.

[18] P. Larrañaga, B, Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, Vol. 7, No. 1, pp.86–112, 2005.

[19] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick. Gene selection: a Bayesian variable selection approach. *Bioinformatics*, Vol. 19, No. 1, pp.90–97, 2003.

[20] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F.X. Mayer, and H. W. Mewes. Gene selection from microarray data for cancer classification–a machine learning approach. *Computational Biology and Chemistry*, Vol. 29, Issue 1, pp.37–46, 2005.

[21] H. H. Zhang, J. Ahn, X. Lin, and C. Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, Vol. 22, No. 1, pp.88–95, 2005.

[22] H. Xiong, and X. Chen. Kernel-based distance metric learning for microarray data classification. *BMC Bioinformatics*, 7:299, 2006.

[23] K. Blekas, N. P. Galatsanos, A. Likas, and I. E. Lagaris. Mixture model analysis of DNA microarray images. *IEEE Transactions on Medical Imaging*, Vol. 24, pp.901–909, 2005.

[24] M. Ouyang, W. J. Welsh, and P. Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, Vol. 20, No. 6, pp.917–923, 2004.

[25] Y. Qu, and S. Xu. Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, Vol. 20, No. 12, pp.1905–1913, 2004.

[26] Y. Lee, and C. K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, Vol. 19, No. 9, pp.1132–1139, 2003.

[27] F. Chu, and L. Wang. Applications of Support Vector Machines to Cancer Classification with Microarray Data. *International Journal of Neural Systems*, Vol. 15, No. 6, pp.475–487, 2005.

[28] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, Vol. 102, No. 43, pp. 15545–15550, 2005.

[29] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, Vol. 21, No. 9, pp. 1943–1949, 2005.

[30] I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:242, 2007.

[31] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, 1999.

[32] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, Vol. 258, No. 5083, pp.818–821, 1992.

[33] D. Lipson, Y. Aumann, A. Ben-Dor, N. Linial, and Z. Yakhini. Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis. *Journal of Computational Biology*, Vol. 13, No. 2, pp. 215–228, 2006.

[34] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, and A. N. A. N. Jain. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, Vol.90, Issue 1, pp.132–153, 2004.

[35] J. C. Marioni, N. P. Thorne, and S. Tavaré. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, Vol. 22, No. 9, pp.1144–1146, 2006.

[36] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, Vol. 5, No. 4, pp.557–572, 2004.

[37] E. S. Venkatraman, and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, Vol. 23, No. 6, pp.657–663, 2007.

[38] Yuta Ishikawa, and Ryohei Nakano. Landscape of a Likelihood Surface for Gaussian Mixture and its Use for the EM Algorithm. In *Proceedings of International Joint Conference on Neural Networks*, pp. 1434–1440, 2006.

[39] A. Frank, and A. Asuncion. UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`. Irvine, CA: University of California, School of Information and Computer Science. 2010.

[40] P. Vlachos. StatLib data sets archive, `http://lib.stat.cmu.edu/datasets`. 2005.

[41] D. M. Rocke and J. Dai. Sampling and Subsampling for Cluster Analysis in Data Mining: With Applications to Sky Survey Data. *Data Mining and Knowledge Discovery*, Vol. 7, No. 2, pp. 215–232, 2003.

[42] Yuta Ishikawa, and Ryohei Nakano. Obtaining EM Initial Points by Using the Primitive Initial Point and Subsampling Strategy. In *Proceedings of International Joint Conference on Neural Networks*, pp. 1115–1120, 2007.

[43] Yuta Ishikawa, and Ryohei Nakano. EM Algorithm with PIP Initialization and Temperature-based Selection. In *Proceedings of Knowledge-Based Intelligent Information and Engineering Systems*, LNAI, Vol. 5179, pp. 58–66, 2008.

[44] Yuta Ishikawa, Ichiro Takeuchi, and Ryohei Nakano. Multi-directional search from the primitive initial point for Gaussian mixture estimation using variational Bayes method. *Neural Networks*, Vol. 23, Issue 3, pp. 356–364, 2010.

[45] Manfred Opper and David Saad (eds.). Advanced Mean Field Methods. *The MIT Press*, 2001.

[46] A. Corduneanu and C. M. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In *Artificial Intelligence and Statistics*, 2001.

[47] C. M. Bishop. Pattern Recognition and Machine Learning. *Springer*, 2006.

[48] J. D. Watson, and F. H. C. Crick. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171, pp.737-738, 1953.

[49] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, Vol. 20, pp.207–211, 1998.

[50] H. Tagawa, S. Tsuzuki, R. Suzuki, S. Karnan, A. Ota, Y. Kameoka, M. Suguro, K. Matsuo, M. Yamaguchi, M. Okamoto, Y. Morishima, S. Nakamura, and M. Seto. Genome-wide array-based comparative genomic hybridization of diffuse large-b-cell lymphoma: comparison between cd5-positive and cd5-negative cases. *Cancer Research*, Vol. 64, pp. 5948–5955, 2004.

[51] H. Tagawa, S. Karnan, R. Suzuki, K. Matsuo, X. Zhang, A. Ota, Y. Morishima, S. Nakamura, and M. Seto. Genome-wide array-based cgh for mantle cell lymphoma: identification of homozygous deletions of the proapoptotic gene bim. *Oncogene*, Vol. 24, pp.1348–1358, 2005.

[52] I. Takeuchi, H. Tagawa, A. Tsujikawa, M. Nakagawa, M. Katayama, Y. Guo, and M. Seto. The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of b-cell lymphomas and genetic regions involved in lymphomagenesis. *Haematologica*, Vol. 94, pp. 51–69, 2009.

[53] H. Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, Vol. 2, No. 3, pp. 360–378, 1931.

[54] J. H. Friedmann and L. C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample test. *Annals Statistics*, Vol. 7, No. 4, pp. 697-717, 1979.

[55] M. F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, Vol. 81, No. 395, pp. 799–806, 1986.

[56] N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics* Vol.16, No. 2, pp. 772–783, 1988.

[57] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. A method for calling gains and losses in array CGH data. *Biostatistics*, Vol. 6, No. 1, pp. 45–58, 2005.

[58] Y. Ge, S. Dudoit, and T. Speed. Resampling-based multiple testing for microarray data analysis. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, Vol. 12, No. 1, pp.1–77, 2003.

[59] S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological Knowledge. *Bioinformatics*, Vol. 22, No. 19, pp.2373–2380, 2006.

[60] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, Vol. 18, No. 1, pp.71–103, 2003.

[61] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp.65–70, 1979.

[62] Y. Benjamini, and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, Vol. 57, No. 1. pp.289–300, 1995.

[63] 石川 勇太, 磯部 浩太, 烏山 昌幸, 泉 泰介, 竹内 一郎. MST に基づく SVM パス追跡を用いた多重多変量 2 標本検定による遺伝子群解析に関する一考察. 電子情報通信学会, 信学技報, Vol. 110, No. 265, pp.211–220, 2010.

[64] X. Yang, Q. Song, and Y. Wang. A weighted support vector machine for data classification. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 21, No. 5, pp.961-976, 2007.

[65] M. Karasuyama, N. Harada, M. Sugiyama, and I. Takeuchi. Multi-parametric Solution-path Algorithm for Instance-weighted Support Vector Machines. arXiv:1009.4791, 2010.

[66] R. L. Graham, and P. Hell. On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing*, Vol. 7, No. 1, pp. 43–57, 1985.

[67] R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, Vol. 36, pp. 1389–1401, 1957.

[68] E. N. Gilbert, and H. O. Pollak. Steiner Minimal Trees. *Journal on Applied Mathematics*, Vol. 16, No. 1, 1968.

[69] V. K. Mootha, C. M. Lindgren, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature of Genetics*, Vol. 34, pp. 267–273, 2003.

[70] CC. Chang and CC. Lin. LIBSVM: a library for support vector machines. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[71] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, Vol. 102, No. 38, pp. 13544–13549, 2005.

[72] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, Vol. 97, No. 457, pp. 77–87, 2002.

[73] E. Fix, and J. L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. *Technical Report* 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[74] J. M. Boyett, and J. J. Shuster. Nonparametric one-side tests in multivariate analysis with medical applications. *Journal of the American Statistical Association*, Vol. 72, pp. 665–668, 1977.

[75] Yuta Ishikawa, and Ichiro Takeuchi. Differentially Aberrant Region Detection in Array CGH Data based on Nearest Neighbor Classification Performance. *IPSJ Transactions on Bioinformatics*, Vol. 3, 70–81, 2010.

[76] J. R. Schott. Matrix derivatives and related topics, In *Matrix Analysis for Statistics 2nd ed.*, (pp.351–393), John Wiley & Sons, Inc, 2005.

研究業績

学会誌掲載論文 (査読有り)

[1] **Yuta Ishikawa**, Ichiro Takeuchi, and Ryohei Nakano. Multi-directional search from the primitive initial point for Gaussian mixture extimation using variational Bayes method. *Neural Networks*, Vol. 23, Issue 3, pp. 356–364, 2010.

[2] M. Uchida, Y. Tsukamoto, T. Uchida, **Y. Ishikawa**, T. Nagai, N. Tung, C. Nakada, A. Kuroda, T. Okimoto, M. Kodama, K. Murakami, T. Noguchi, K. Matuura, M. Tanigawa, M. Seto, H. Ito, T. Fujioka, I. Takeuchi, M. Moriyama. Genomic profiling of gastric carcinoma in situ and adenomas by array-based comparative genomic hybridization. *Journal of Pathology*, Vol. 221, Issue 1, pp. 96–105, 2010.

[3] **Yuta Ishikawa**, and Ichiro Takeuchi. Differentially Aberrant Region Detection in Array CGH Data based on Nearest Neighbor Classification Performance. *IPSJ Transactions on Bioinformatics*, Vol. 3, 70–81, 2010.

国際学会発表論文 (査読有り)

[4] **Yuta Ishikawa**, and Ryohei Nakano. Landscape of a Likelihood Surface for Gaussian Mixture and its Use for the EM Algorithm. In *Proceedings of International Joint Conference on Neural Networks* (IJCNN '06), pp. 1434–1440, 2006.

[5] **Yuta Ishikawa**, and Ryohei Nakano. Obtaining EM Initial Points by Using the Primitive Initial Point and Subsampling Strategy. In *Proceedings of International Joint Conference on Neural Networks* (IJCNN '07), pp. 1115–1120, 2007.

[6] **Yuta Ishikawa**, and Ryohei Nakano. EM Algorithm with PIP Initialization and Temperature-based Selection. In *Proceedings of Knowledge-Based Intelligent Information and Engineering Systems* (KES '08), LNAI, Vol. 5179, pp. 58–66, 2008.

[7] **Yuta Ishikawa**, Ichiro Takeuchi, and Ryohei Nakano. Variational Bayes from the Primitive Initial Point for Gaussian Mixture Estimation. In *Proceedings of 16^{th} International Conference on Neural Information Processing* (ICONIP '09), LNCS, Vol. 5863, pp. 159–166, 2009.

[8] Naoyuki Harada, **Yuta Ishikawa**, Ichiro Takeuchi, and Ryohei Nakano. An Bayesian Graph Clustering Approach Using the Prior Based on Degree Distribution. In *Proceedings of 16^{th} International Conference on Neural Information Processing* (ICONIP '09), LNCS, Vol. 5863, pp. 167–174, 2009.

[9] **Yuta Ishikawa**, and Ichiro Takeuchi. Detecting Differenctially Aberrant Genomic Regions in Multi-Sample Array CGH Experiments Using Nearest-Neighbor Multivariate Test. In *Proceedings of International Joint Conference on Neural Networks* (IJCNN '10), pp. 1547–1554, 2010.

国内学会発表論文

[10] **石川 勇太**, 中野 良平. 原始初期点と Subsampling を用いた EM 初期値生成法. 電子情報通信学会, 信学技報, Vol. 107, No. 410, pp. 7–12, 2007.

[11] **石川 勇太**, 竹内 一郎, 中野 良平. 原始初期点を初期値とする変分ベイズ法に関する一考察. 電子情報通信学会, 信学技報, Vol. 108, No. 410, pp. 13–18, 2008.

[12] **石川 勇太**, 竹内 一郎. 最近傍多変量検定を用いたアレイ CGH のゲノム異常領域同定に関する一考察. 電子情報通信学会, 信学技報, Vol. 109, No. 53, pp. 31–36, 2009.

[13] **石川 勇太**, 竹内 一郎. 重み学習を用いた array CGH のゲノム異常領域同定に関する一考察. 第 7 回情報学ワークショップ (WiNF '09) 論文集, pp. 51–56, 2009.

[14] **石川 勇太**, 磯部 浩太, 竹内 一郎. サポートベクトルマシンのラベル並べ替え解のパス追跡とマイクロアレイデータ解析への応用に関する一考察. 電子情報通信学会, 信学技報, Vol. 109, No. 461, pp. 261–266, 2010.

[15] **石川 勇太**, 磯部 浩太, 烏山 昌幸, 泉 泰介, 竹内 一郎. MST に基づく SVM パス追跡を用いた多重多変量 2 標本検定による遺伝子群解析に関する一考察. 電子情報通信学会, 信学技報, Vol. 110, No. 265, pp.211–220, 2010.

受賞歴

平成 21 年度 電子情報通信学会 東海支部 学生研究奨励賞

付録

A 確率分布

本節では, 本稿で必要となる確率密度関数についてまとめる. それらは, 多変量正規分布, Gamma 分布, Dirichlet 分布, Wishart 分布, Gaussian-Wishart 分布である.

A.1 正規分布

平均 $\mu \in (-\infty, \infty)$, 分散 $\sigma^2 > 0$ の (単変量) 正規分布は, 以下で与えられる:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{1}{2\sigma^2}(x-\mu)^2\Big\}.$$

確率変数 X が, 上記の正規分布に従う場合, X の期待値, mode, 分散は,

$$\begin{aligned} \mathbb{E}[X] &= \mu, \\ \mathrm{var}[X] &= \sigma^2, \\ \mathrm{mode}[X] &= \mu, \end{aligned} \tag{A.1}$$

で与えられる.

平均ベクトル $\boldsymbol{m}$, 分散共分散行列 $\boldsymbol{\Sigma}$ の p 変量正規分布, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma})$ は以下で与えられる:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right].$$

確率変数 X が, 平均ベクトル $\boldsymbol{m}$, 分散共分散行列 $\boldsymbol{\Sigma}$ の多変量正規分布に従うとき, X の期待値, 共分散および mode は,

$$\begin{aligned} \mathbb{E}[X] &= \boldsymbol{m}, \\ \mathrm{cov}[X] &= \boldsymbol{\Sigma}, \\ \mathrm{mode}[X] &= \boldsymbol{m}, \end{aligned}$$

で与えられる.

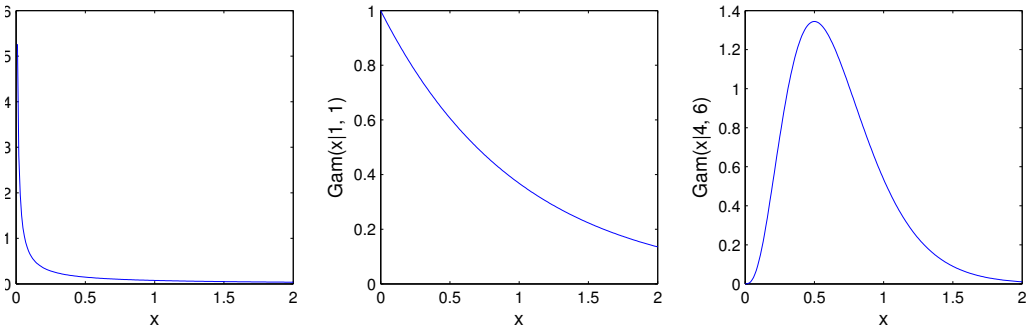


図 A.1: Gamma 分布の例: 左から $(a, b) = (10, 1),\ (1, 1),\ (4, 6)$

A.2 Gamma 分布

Gamma 分布はパラメータ a, b で特徴づけられる, 正の確率変数 x に対して定義される確率密度関数で, 以下で与えられる:

$$\mathrm{Gam}(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} \exp(-bx) \tag{A.2}$$

ここで, $\Gamma(a)$ は Gamma 関数で以下で与えられる.

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du \tag{A.3}$$

確率変数 X が Gamma 分布に従うとき, X の期待値および分散は次のようになる:

$$\begin{aligned} \mathbb{E}[X] &= \frac{a}{b}, \\ \mathrm{var}[X] &= \frac{a}{b^2}. \end{aligned}$$

図 A.1 にいくつかの a, b における Gamma 分布の形状を示す. なお, Gamma 関数の対数を微分したものは, digamma 関数と呼ばれ,

$$\psi(a) \equiv \frac{d}{da} \ln \Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)},$$

で与えられる.

A.3 Gaussian-Gamma 分布

Gaussian-Gamma 分布は, 正規分布のベイズ推定において, 平均, 分散ともに未知の場合に, それらの共役事前分布として用いられる分布である. Gaussian-Gamma 分布は, 正規分布と Gamma 分布の積として与えられ, 以下の形をとる:

$$p(\mu, \lambda|\mu_0, \eta_0, a_0, b_0) = \mathcal{N}(\mu|\mu_0, (\eta_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0, b_0).$$

ここで, μ_0, η_0, a_0, b_0 はハイパーパラメータである.

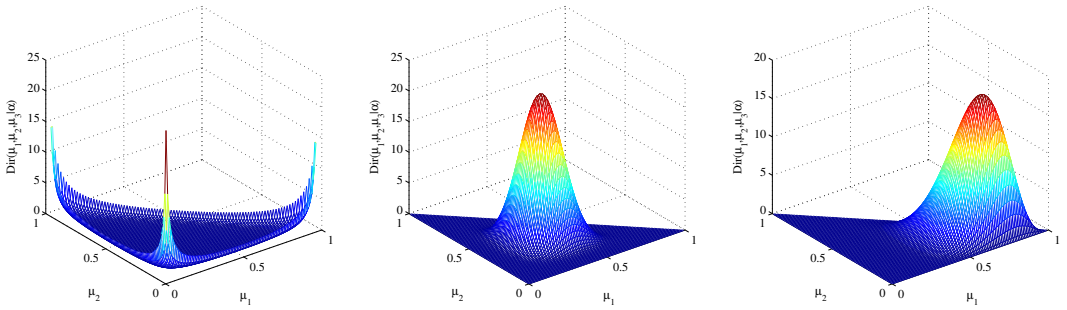


図 A.2: Dirichlet 分布 ($K = 3$) の例 (左から $\boldsymbol{\alpha} = [0.2, 0.2, 0.2]$, $[10, 10, 10]$, $[10, 3, 3]$): Dirichlet 分布を混合比の事前分布として用いる場合, 三角形の各頂点が混合要素を, そして, 三角形の内部の位置によりどの混合要素に重点がおかれるかが決定される. 図の場合, 三角形の各頂点 $(1, 0)$, $(0, 1)$, $(0, 0)$ が, それぞれ $k = 1$, 2, 3 に対応する. 図左の例では, 1 つもしくは 2 つの混合要素の混合比が大きくなり, 残りはほぼ 0 となる確率が高いことを意味している. 一方で, 真中の図では, 全ての混合要素が等しく現れうることを意味している. 右ではいずれの混合要素も現れうるが, やや $k = 1$ が現れる確率が高いことを意味する.

A.4 Dirichlet 分布

Dirichlet 分布は K 個の $0 \leq \mu_k \leq 1, k \in \mathbb{N}_K$ なる確率変数 ($\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^\top$ と表記する) の多次元分布であり,

$$\begin{aligned} \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \\ C(\boldsymbol{\alpha}) &= \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}, \\ \hat{\alpha} &= \sum_{k=1}^{K} \alpha_k, \end{aligned}$$

で与えられる. ただし, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^\top$ はハイパーパラメータであり, 確率変数 $\{\mu_k\}_{k \in \mathbb{N}_K}$ には以下の制約が課される:

$$0 \leq \mu_k \leq 1, \quad \sum_{k=1}^{K} \mu_k = 1,$$

Dirichlet 分布の期待値, mode は以下で与えられる:

$$\begin{aligned} \mathbb{E}[\mu_k] &= \frac{\alpha_k}{\hat{\alpha}}, \\ \mathrm{mode}[\mu_\mathrm{k}] &= \frac{\alpha_k - 1}{\hat{\alpha} - K}. \end{aligned}$$

図 A.2 に $K = 3$ の Dirichlet 分布の例を示す.

A.5 Wishart 分布

p 次元 Wishrt 分布 $\mathcal{W}(\mathbf{\Lambda}|\boldsymbol{W}_0,\nu)$ は以下で与えられる:

$$\begin{aligned}\mathcal{W}(\mathbf{\Lambda}|\boldsymbol{W}_0,\nu) &= B(\boldsymbol{W}_0,\nu)|\mathbf{\Lambda}|^{(\nu-p-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_0^{-1}\mathbf{\Lambda})\right),\\ B(\boldsymbol{W}_0,\nu) &= |\boldsymbol{W}_0|^{-\nu/2}\left(2^{\nu p/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}.\end{aligned}$$

Wishart 分布は Gamma 分布の多変量分布である. 平均, mode は以下で与えられる.

$$\begin{aligned}\mathbb{E}[\mathbf{\Lambda}] &= \nu\boldsymbol{W}_0,\\ \mathrm{mode}[\mathbf{\Lambda}] &= (\nu-p-1)\boldsymbol{W}_0 \quad \text{for} \quad \nu\geq p+1.\end{aligned}$$

A.6 Gaussian-Wishart 分布

Gaussian-Wishart 分布は多変量正規分布の平均ベクトル, 分散共分散行列が共に未知の場合に用いられる共役事前分布である. この分布は, 平均ベクトル $\boldsymbol{m}$ に対して Gauss 分布を, 精度行列 $\mathbf{\Lambda}$ に対して Wishart 分布を割り当て積を取った以下のような分布である:

$$p(\boldsymbol{m},\mathbf{\Lambda}|\boldsymbol{\mu}_0,\eta_0,\boldsymbol{W}_0,\nu_0)=\mathcal{N}(\boldsymbol{m}|\boldsymbol{\mu}_0,(\eta_0\mathbf{\Lambda})^{-1})\mathcal{W}(\mathbf{\Lambda}|\boldsymbol{W}_0,\nu_0).$$

ここで, $\boldsymbol{\mu}_0$, η_0, $\boldsymbol{W}_0$, ν_0 はハイパーパラメータである.

B EM アルゴリズムにおける Q 関数の停留点での Hesse 行列

混合正規分布推定での EM アルゴリズムの目的関数は PIP において停留点を持つ. ここでは, EM アルゴリズムでの目的関数である $\mathcal{Q}$ 関数

$$\mathcal{Q}(\Theta^{(t)},\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})\ln p(\boldsymbol{x}_i,k|\boldsymbol{\theta}_k)-\lambda\left[\sum_{k=1}^{K}\pi_k-1\right], \tag{B.4}$$

の停留点での Hesse 行列を導出する. ここで,

$$\begin{aligned}p(\boldsymbol{x}_i,k|\boldsymbol{\theta}_k) &= \pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k,\mathbf{\Sigma}_k),\\ P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)}) &= \frac{\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k^{(t)},\mathbf{\Sigma}_k^{(t)})}{\displaystyle\sum_{k=1}^{K}\pi_k g_k(\boldsymbol{x}_i|\boldsymbol{m}_k^{(t)},\mathbf{\Sigma}_k^{(t)})},\\ g_k(\boldsymbol{x}_i|\boldsymbol{m}_k,\mathbf{\Sigma}_k) &= \frac{1}{\sqrt{(2\pi)^K|\mathbf{\Sigma}_k|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{m}_k)^\top\mathbf{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{m}_k)\right\},\end{aligned}$$

である.

まず, 式 (B.4) を整理し, 以下を得る:

$$\begin{aligned}\mathcal{Q}(\Theta^{(t)},\Theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(t)})\Big\{\ln\pi_k - \frac{K}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_k^{-1}| \\ &\qquad -\frac{(\boldsymbol{x}_i-\boldsymbol{m}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{m}_k)}{2}\Big\} - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).\end{aligned}$$

混合正規分布推定におけるEMアルゴリズムの未知パラメータは$\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}_K} = \{\pi_k,\boldsymbol{m}_k,\boldsymbol{\Sigma}_k\}_{k\in\mathbb{N}_K}$であるから, これらのパラメータに関する$\mathcal{Q}(\Theta^{(t)},\Theta)$の2階微分を考える. なお, 本稿における実験では, 分散共分散行列は対角要素のみを考えることとする. 従って, 分散共分散行列による偏微分は,

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \begin{bmatrix} v_{k1} & & O \\ & \ddots & \\ O & & v_{kp}\end{bmatrix} \\ &= \mathrm{diag}(\boldsymbol{v}_k), \\ \boldsymbol{v}_k &= [v_{k1},\ldots,v_{kp}]^\top\end{aligned}$$

として, $\boldsymbol{v}_k$を用いる. 停留点であること, 即ち, $\partial\mathcal{Q}(\Theta^{(t)},\Theta)/\partial\Theta = 0$, と制約条件$\sum_{k=1}^{K}\pi_k = 1$を用いることで, 目的関数のHesse行列は以下のようになる:

$$\frac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_{k'}^\top} = \begin{bmatrix} \dfrac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\pi_k\partial\pi_{k'}} & \mathbf{0}_p^\top & \mathbf{0}_p^\top \\ \mathbf{0}_p & \dfrac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{m}_k\partial\boldsymbol{m}_{k'}^\top} & \dfrac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{m}_k\partial\boldsymbol{v}_{k'}^\top} \\ \mathbf{0}_p & \left(\dfrac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{m}_k\partial\boldsymbol{v}_{k'}^\top}\right)^\top & \dfrac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{v}_k\partial\boldsymbol{v}_{k'}^\top}\end{bmatrix}. \tag{B.5}$$

式 (B.5) における各要素は以下で与えられる:

$$\frac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\pi_k\partial\pi_{k'}} = \begin{cases} -\dfrac{n}{\pi_k} & \text{for} \quad k'=k, \\ 0 & \text{for} \quad k'\neq k,\end{cases}$$

$$\frac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{m}_k\partial\boldsymbol{m}_{k'}^\top} = \begin{cases} -\displaystyle\sum_{i=1}^{n} P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k)\boldsymbol{\Sigma}_k^{-1} & \text{for} \quad k'=k, \\ O_{p\times p} & \text{for} \quad k'\neq k,\end{cases}$$

$$\frac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{v}_k\partial\boldsymbol{v}_{k'}^\top} = \begin{cases} -\dfrac{1}{2}\displaystyle\sum_{i=1}^{n} P(k|\boldsymbol{x}_k,\boldsymbol{\theta}_k)\Big[\boldsymbol{\Sigma}_k^{-2} - 2\boldsymbol{\Sigma}_k^{-3}\big\{\mathrm{diag}(\boldsymbol{x}_i-\boldsymbol{m}_k)\big\}^2\Big] & \text{for} \quad k'=k, \\ \boldsymbol{O}_{p\times p} & \text{for} \quad k'\neq k,\end{cases}$$

$$\frac{\partial^2\mathcal{Q}(\Theta^{(t)},\Theta)}{\partial\boldsymbol{m}_k\partial\boldsymbol{v}_{k'}^\top} = \begin{cases} -\displaystyle\sum_{i=1}^{n} P(k|\boldsymbol{x}_i,\boldsymbol{\theta}_k)\mathrm{diag}(\boldsymbol{x}_i-\boldsymbol{m}_k)\boldsymbol{\Sigma}_k^{-2} & \text{for} \quad k'=k, \\ \boldsymbol{O}_{p\times p} & \text{for} \quad k'\neq k,\end{cases}$$

C 変分ベイズ法における負の自由エネルギーの停留点での Hesse 行列

ここでは, 負の自由エネルギー関数, 式 (4.26) の停留点における Hesse 行列を導出する. まず, Hesse 行列を計算するために用いる行列, ベクトル演算子および行列の偏微分に関する公式 [76] をまとめる. まず, 行列に対する 2 つの演算子を定義する. ある $M \times N$ 行列 A,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix},$$

に対して,

$$\mathrm{vec}(A) \equiv [a_{11}, \ldots, a_{1N}, \ldots, a_{M1}, \ldots, a_{MN}]^\top,$$

また, 任意の行列 B に対して

$$A \otimes B \equiv \begin{bmatrix} a_{11}B & \cdots & a_{1N}B \\ \vdots & \ddots & \vdots \\ a_{M1}B & \cdots & a_{MN}B \end{bmatrix},$$

と定義する. 次に, 行列, ベクトル関数に対する偏微分の以下の公式を導入する:

$$\begin{aligned} \frac{\partial}{\partial \mathrm{vec}(X)^\top} \mathrm{Tr}(X) &= \mathrm{vec}(\mathrm{I_M})^\top, \\ \frac{\partial}{\partial \mathrm{vec}(X)^\top} \ln|X| &= \mathrm{vec}\big[(X^{-1})^\top\big]^\top, \\ \frac{\partial}{\partial \mathrm{vec}(X)^\top} \mathrm{vec}(X^{-1}) &= -(X^{-1})^\top \otimes X^{-1}. \end{aligned}$$

ただし, X は $M \times M$ 実数行列である.

上記の演算子, 公式を用いて関数 $\mathcal{L}$ の停留点における Hesse 行列を導出する. まず, 式 (4.26) を再掲する:

$$\begin{aligned} \mathcal{L}(q) &= \sum_Z \int\int\int q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda}) \ln\left\{ \frac{p(X, Z, \boldsymbol{\pi}, \mu, \boldsymbol{\Lambda})}{q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})} \right\} \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{m} \mathrm{d}\boldsymbol{\Lambda} \\ &= \mathbb{E}[\ln p(X, Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(Z, \boldsymbol{\pi}, \boldsymbol{m}, \boldsymbol{\Lambda})] \\ &= \mathbb{E}[\ln p(X|Z, \boldsymbol{m}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(Z|\boldsymbol{\pi})] + \mathbb{E}[p(Z|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{m}, \boldsymbol{\Lambda})] \\ &\quad -\mathbb{E}[\ln q(Z)] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{m}, \boldsymbol{\Lambda})]. \end{aligned}$$

ここで, 上式中の各項は以下で与えられる:

$$
\begin{aligned}
\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{m},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{k=1}^{K} N_k\Big\{\ln\tilde{\boldsymbol{\Lambda}}_k - p\eta_k^{-1} - \nu_k \mathrm{Tr}(\boldsymbol{S}_k\boldsymbol{W}_k) \\
&\quad -\nu_k(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_k)^\top \boldsymbol{W}_k(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_k) - p\ln(2\pi)\Big\} \\
\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] &= \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}\ln\tilde{\pi}_k \\
\mathbb{E}[\ln p(\boldsymbol{\pi})] &= \ln C(\boldsymbol{\alpha}_0) + (\alpha_0-1)\sum_{k=1}^{K}\ln\tilde{\pi}_k \\
\mathbb{E}[\ln p(\boldsymbol{m},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{k=1}^{K}\Big\{p\ln\frac{\eta_0}{2\pi} + \ln\tilde{\boldsymbol{\Lambda}}_k - \frac{p\eta_0}{\eta_k} - \eta_0\nu_k(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)^\top\boldsymbol{W}_k(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)\Big\} \\
&\quad + K\ln B(\boldsymbol{W}_0,\nu_0) + \frac{\nu_0-p-1}{2}\sum_{k=1}^{K}\ln\tilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k) \\
\mathbb{E}[\ln q(\boldsymbol{Z})] &= \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}\ln r_{ik} \\
\mathbb{E}[\ln q(\boldsymbol{\pi})] &= \sum_{k=1}^{K}(\alpha_k-1)\ln\tilde{\pi}_k + \ln C(\boldsymbol{\alpha}) \\
\mathbb{E}[\ln q(\boldsymbol{m},\boldsymbol{\Lambda})] &= \sum_{k=1}^{K}\Big\{\frac{1}{2}\ln\tilde{\boldsymbol{\Lambda}}_k + \frac{p}{2}\ln\frac{\eta_k}{2\pi} - \frac{p}{2} - \mathrm{H}[\boldsymbol{\Lambda}_\mathrm{k}]\Big\}
\end{aligned}
$$

ただし,

$$
\begin{aligned}
r_{ik} &= \frac{\rho_{ik}}{\sum_{j=1}^{K}\rho_{ij}}, \\
\ln\rho_{ik} &= \ln\tilde{\pi}_k + \frac{1}{2}\tilde{\boldsymbol{\Lambda}}_k - \frac{p}{2}\ln(2\pi), \\
\mathbb{E}[z_{ik}] &= r_{ik} \\
\mathbb{E}_{\boldsymbol{m}_k\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_i-\boldsymbol{m}_k)^\top\boldsymbol{\Lambda}_k(\boldsymbol{x}_k-\boldsymbol{m}_k)] &= p\eta_k^{-1} + \nu_k(\boldsymbol{x}_k-\boldsymbol{\mu}_k)^\top\boldsymbol{W}_k(\boldsymbol{x}_k-\boldsymbol{\mu}_k) \\
\ln\tilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] &= \sum_{i=1}^{p}\psi\left(\frac{\nu_k+1-i}{2}\right) + p\ln 2 + \ln|\boldsymbol{W}_k| \\
\ln\tilde{\pi}_k \equiv \mathbb{E}[\ln\pi_k] &= \psi(\alpha_k) - \psi(\hat{\alpha}) \\
\mathrm{H}[\boldsymbol{\Lambda}_\mathrm{k}] &= -\ln B(\boldsymbol{W}_k,\nu_k) - \frac{\nu_k-p-1}{2}\ln\tilde{\boldsymbol{\Lambda}}_k + \frac{\nu_k p}{2}
\end{aligned}
$$

式 (4.26) を変形し, 整理することで以下を得る:

$$
\begin{aligned}
\mathcal{L} \;=\; & \frac{1}{2}\sum_{k=1}^{K}\Bigg[2(\alpha_0-\alpha_k)\{\psi(\alpha_k)-\psi(\hat{\alpha})\}-p\bigg(\frac{N_k+\eta_0}{\eta_k}+\ln\eta_k\bigg)\\
& +(\nu_0+N_k-\nu_k)\bigg\{\sum_{i=1}^{p}\psi\bigg(\frac{\nu_k+1-i}{2}\bigg)+p\ln 2+\ln|\boldsymbol{W}_k|\bigg\}\\
& -\nu_k\Bigg\{\mathrm{Tr}\Bigg[\bigg\{\boldsymbol{W}_0^{-1}+N_k\boldsymbol{S}_k+\frac{\eta_0 N_k}{\eta_0+N_k}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_0)^{\top}\\
& +(\eta_0+N_k)\bigg(\boldsymbol{\mu}_k-\frac{\eta_0\boldsymbol{\mu}_0+N_k\bar{\boldsymbol{x}}_k}{\eta_0+N_k}\bigg)\bigg(\boldsymbol{\mu}_k-\frac{\eta_0\boldsymbol{\mu}_0+N_k\bar{\boldsymbol{x}}_k}{\eta_0+N_k}\bigg)^{\top}\bigg\}\boldsymbol{W}_k\Bigg]\\
& -\ln|\boldsymbol{W}_k|-p-p\ln 2\Bigg\}+2\sum_{i=1}^{p}\ln\Gamma\bigg(\frac{\nu_k+1-i}{2}\bigg)+2\ln\Gamma(\alpha_k)\Bigg]\\
& +\sum_{k=1}^{K}N_k\{\psi(\alpha_k)-\psi(\hat{\alpha})\}-\ln\Gamma(\hat{\alpha}).
\end{aligned}
\tag{C.6}
$$

混合正規分布推定問題における変分ベイズ法の目的関数, 式 (4.26) は K 個の混合要素それぞれに対して, 5 種のハイパーパラメータ, $\{\boldsymbol{\theta}_k\}_{k\in\mathbb{N}_K}=\{\alpha_k,\eta_k,\boldsymbol{m}_k,\boldsymbol{W}_k,\nu_k\}_{k\in\mathbb{N}_K}$ により特徴づけられている. 従って, Hesse 行列を求めるために式 (C.6) をこれらのパラメータについて 2 階微分を考える. 停留点では, $\partial\mathcal{L}/\partial\Theta=0$, となるため, これを用いると Hesse 行列は以下の形を取る:

$$
\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_{k'}^{\top}}=\begin{bmatrix}
(C.8) & 0 & (\mathbf{0}_{p\times 1})^{\top} & (\mathbf{0}_{p^2\times 1})^{\top} & 0\\
0 & (C.9) & (\mathbf{0}_{p\times 1})^{\top} & (\mathbf{0}_{p^2\times 1})^{\top} & 0\\
\mathbf{0}_{p\times 1} & \mathbf{0}_{p\times 1} & (C.10) & (\boldsymbol{O}_{p^2\times p})^{\top} & (\mathbf{0}_{p\times 1})^{\top}\\
\mathbf{0}_{p^2\times 1} & \mathbf{0}_{p^2\times 1} & \boldsymbol{O}_{p^2\times p} & (C.11) & (C.13)\\
0 & 0 & \mathbf{0}_{p\times 1} & (C.13)^{\top} & (C.12)
\end{bmatrix}.
\tag{C.7}
$$

ここで, 式 (C.7) 中に (C.8), ..., (C.13) で示された要素は次の通りである:

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_k \partial \alpha_{k'}} = \begin{cases} -\sum_{n=0}^{\infty} \left\{ \frac{1}{(\alpha_k + n)^2} - \frac{1}{(\hat{\alpha} + n)^2} \right\} & \text{for} \quad k' = k, \\ -\sum_{n=0}^{\infty} \frac{2}{(\hat{\alpha} + n)^3} & \text{for} \quad k' \neq k, \end{cases} \tag{C.8}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_k \partial \eta_{k'}} = \begin{cases} -\frac{p}{2\eta_k^2} & \text{for} \quad k' = k, \\ 0 & \text{for} \quad k' \neq k, \end{cases} \tag{C.9}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{m}_k \partial \boldsymbol{m}_{k'}^\top} = \begin{cases} \eta_k \boldsymbol{W}_k & \text{for} \quad \mathrm{k}' = \mathrm{k}, \\ \boldsymbol{O}_{p \times p} & \text{for} \quad k' \neq k, \end{cases} \tag{C.10}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathrm{vec}(\boldsymbol{W}_k) \partial \mathrm{vec}(\boldsymbol{W}_{k'})^\top} = \begin{cases} -\frac{\nu_k}{2} (\boldsymbol{W}_k^{-1})^\top \otimes \boldsymbol{W}_k^{-1} & \text{for} \quad k' = k, \\ \boldsymbol{O}_{p^2 \times p^2} & \text{for} \quad k' \neq k, \end{cases} \tag{C.11}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \nu_k \partial \nu_{k'}} = \begin{cases} -\sum_{i=1}^{p} \sum_{n=0}^{\infty} \frac{1}{(\nu_k + 1 - i + 2n)^2} & \text{for} \quad k' = k, \\ 0 & \text{for} \quad k' \neq k, \end{cases} \tag{C.12}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathrm{vec}(\boldsymbol{W}_k) \partial \nu_{k'}} = \begin{cases} -\frac{1}{2} \mathrm{vec}(\boldsymbol{W}_k) & \text{for} \quad k' = k, \\ \boldsymbol{0}_{p^2 \times 1} & \text{for} \quad k' \neq k. \end{cases} \tag{C.13}$$