データ分布空間における距離構造の学習に関する研究

A Study on Distance Metric Learning of Data Distribution Spaces

2010年7月

早稲田大学大学院先進理工学研究科
電気・情報生命専攻 情報学習システム研究
日野英逸

目 次

第1章　序論

統計的機械学習やデータマイニング手法はその適用範囲を広げ続けており, 多種多様なデータから有用な情報を抽出する課題が日々生まれてきている. 多くの学習アルゴリズムの性能は, 入力データから抽出する情報の性質と, データの間に定義される距離構造に大きく依存している. 例えば, 胸部の細胞画像の半径, 色彩, 周径などからなる実ベクトルデータによる乳癌の診断という問題を考える. 癌細胞と正常な細胞を比較したとき, 半径の違いと色彩の違いとは質的に全く異なり, それぞれの違いが判別に同程度影響するようにそれぞれの尺度を調整する必要がある. また, 細胞の半径と周径には相関があると考えられるが, 半径が同程度の細胞同士の周径が大きく異なる場合には細胞の形状に異常があると考えらえるため, 両者の相関を考慮した尺度で比較すべきである. 正しい診断結果を得るためには, データベクトル間に適切な距離を定義する必要がある. また別な例として, 個人の嗜好に応じた推薦サービスを考える. 多くのオンラインショップサイトでは, 過去の購買履歴やアンケートなどから, ユーザが好みそうなアイテムを推薦するサービスを提供している. あるいは, ニュース記事のポータルサイトでも, 過去の閲覧履歴や傾向によって, ユーザが関心を持ちそうな記事を推薦するサービスがある. こうした情報推薦を実現するシステムは, 一般に推薦システムと呼ばれている. 特に, 過去の購買履歴に基づきユーザ同士の類似度を定義し, 類似度の高いユーザが高評価をしているアイテムを, そのアイテムをまだ購入していないユーザに推薦するシステムを, 協調フィルタリングと呼ぶ. 協調フィルタリングにおける推薦の良し悪しを決定づけるのは, アイテムの購買履歴という形式で表現されるユーザデータ同士の類似度である.

これまで多くの判別手法やアイテム推薦手法が提案され, 実際的なデータに適用されて成果をあげているが, 精度良く所望の結果を得られる手法は与えられた問題・データから有用な情報を抽出し, データ間に適切な距離あるいは類似度を定めることに成功している手法である. 所与のデータに基づき与えられた課題に応じてデータから情報を抽出し, データ同士

の距離構造を学習する手法は, 教師無し学習, 教師付き学習それぞれの枠組みで数多く提案されている [1].

教師無し学習とは, 学習用のデータとしてデータのクラスラベルや応答変数に関する情報が与えられずに, 説明変数のみから特徴量を抽出し, その特徴が顕著に現れるようにデータ間の距離構造を学習する枠組みである. 例えば, 多変量解析の分野でよく知られている主成分分析は教師無しの距離構造の学習手法の代表例である. 主成分分析においては, 所与のデータの分散が最も大きくなるという意味でデータの特徴的な構造を捉えた部分空間への射影を学習する. 主成分分析によって学習された部分空間にデータを射影した上で比較を行うことで, データの主たる特徴とは無関係なノイズを低減してデータの類似度を比較することが可能となる. 最近では, 生のデータに対する近傍関係からデータをグラフとして表現して, スペクトルグラフ理論 [2] に基づきデータを低次元空間に埋め込む手法が盛んに研究されている [3; 4]. こうした研究は, データの元の空間における近傍関係を保存するように低次元空間における距離構造を学習する手法として理解できる.

一方, 教師付き学習とは, 判別や回帰といった課題における入力データに対応する出力例が与えられた上で, 望ましい入出力関係を記述する写像を学習する枠組みである. 教師付き学習の最も代表的な例の一つは, Fisher の判別分析と呼ばれる手法である [5]. これは主成分分析と同様に分散構造に着目した手法であり, データをクラス別に見た場合のクラス内分散とデータ全体の分散の比を最小化することで, 各クラスのデータを分離するのに最も適した部分空間への射影を学習する. また, 距離構造の学習とはデータをある空間において最適配置する問題と捉えることも出来るため, 高度な最適化の手法を用いたアプローチも数多く提案されている [6; 7; 8; 9].

本論文では, 主に教師付き学習の枠組みで, データからの特徴抽出と距離構造の学習問題を扱う. 上述の距離構造が重要となる 2 つの例において, 前者は細胞の測定で得られる連続量, 後者は購買履歴という離散量を扱う問題である. 本論文では, 情報論的観点からの距離構造学習という立場に立ち, それぞれのタイプの問題に対して

1. データ同士の内積を目的に応じて適切に学習・定義するアプローチ
2. データが発生する分布 (生成モデル) を学習し, モデルに基づく自然な距離構造を学習するアプローチ

をとる.

データ同士の内積を学習するための統一的な手法として, 情報理論に基づく条件付きエントロピー最小化基準による方法を提案する. これにより, 従来の距離構造の学習問題では十分に論じられていなかった, 学習対象の情報論的な意味が明確になる. また, 離散データの生成モデルに基づき自然な距離構造を学習する研究として, 近年重要性を増している, 映画や書籍等のアイテム評価データの生成モデルを提案する. 提案モデルに基づくデータ間の類似度を導出し, 評価者とアイテムの関係性の解析に応用する.

本論文の構成を以下に示す. 第 2 章, 第 3 章では, 本論文で用いる情報理論及び統計的機械学習理論のレビューを行う. 第 2 章では, エントロピーや相互情報量, Kullback-Leibler ダイバージェンスといった情報理論における基本的な量を定義する. 特に本論文で重要となる, Shannon の微分エントロピーとその効率的な推定アルゴリズムを紹介する. 第 3 章では, 教師付きの距離構造学習に関する研究を紹介する. また, データ同士の距離の学習と特徴空間における内積の学習が等価であることを述べ, 特徴空間における内積を間接的に定めるカーネル関数を用いた手法の総称であるカーネル法の理論的な背景を説明する. ここで, 第 4 章で考察する Multiple Kernel Learning(MKL) による特徴空間の距離構造の最適化の理論的背景となる事実を示す.

第 4 章では文献 [10; 11] に従い, 条件付きエントロピー最小化という情報論的な基準に基づくデータ処理の枠組みを提案する. この枠組みの中で, データの低次元空間への線型変換による距離構造の学習手法, つまり次元削減手法を具体的に構成する. これは, データのクラス判別に適した部分空間への射影を求め, その部分空間において自然に定まる距離を用いて判別を行うというアプローチである. さらに, カーネル関数に付随する非線型特徴空間における距離構造の学習を条件付きエントロピー最小化基準の枠組みで行い, 近年盛んに研究されている MKL の一手法として定式化する. これは, MKL におけるカーネル関数の重ね合わせの係数を条件付きエントロピー最小化によって学習することで, 判別に適した特徴空間の距離構造を学習するというアプローチである. 提案する線型次元削減及び MKL 手法を, 人工データ及び実データに適用し, 既存手法との性能比較を行う.

第 5 章では, 離散観測データに対する距離構造の学習問題を考え, データの生成モデルに基づく類似度及びモデルのパラメタ空間へのデータ配

置手法を考察する. まず, 映画や書籍, レストランなどへの評価データの新しい生成モデルを提案する. 従来, 多数のアイテムに対して多数のユーザが比較をおこなったりランキングを与えたりすることで得られるデータは Bradley-Terry モデルあるいは Plackett-Luce モデルと呼ばれる確率モデルによってモデル化されてきた. 本章では, ランキングデータの生成モデルである Plackett-Luce モデルの自然な一般化として, グループ化ランキングモデルを提案する [12; 13; 14; 15; 16]. このモデルはアイテムの持つ価値パラメタによって特徴付けられるが, 尤度関数の直接評価が困難である. そこで, 効率的に評価可能な尤度関数の近似を与え, さらに情報幾何学的な考察を通してモデルのパラメタ推定方法を提案する. 提案する確率モデルを現実の映画評価データ及び書籍評価データに適用し, Fisher カーネルと呼ばれるカーネル関数を用いてユーザ同士の類似度を定義する. この類似度を, 協調フィルタリングにおけるユーザ間類似度として用いたアイテム推薦システムを提案する.

第 6 章では本論文の内容をまとめ, 今後の展望について述べる.

第2章　情報理論からの準備

本論文の第 4 章では確率変数のエントロピーが中心的な役割を果たす. また, 第 5 章では確率 (密度) 関数がなす統計多様体における擬似的な距離として Kullback-Leibler ダイバージェンスが用いられる. そこで, 本章ではエントロピー, 相互情報量, Kullback-Leibler ダイバージェンスと, 観測データを用いたエントロピーの推定手法を簡単に説明する.

2.1　エントロピー, 相互情報量, ダイバージェンス

簡単のため, ここでは 1 次元分布を考える. 確率変数 X は集合 $\mathcal{X} \subseteq \mathbb{R}$ に値を取る関数であり, その実現値を小文字 x で表す. また $|\mathcal{X}|$ で集合 $\mathcal{X}$ の要素数を表す. 本論文では, 確率変数が離散値を取る場合にはその確率関数を P で, 確率変数が連続値を取る場合にはその確率密度関数を p で表すものとする. 混乱が生じない限り, 確率関数あるいは確率密度関数は, 異なる分布の確率 (密度) 関数であっても全て P あるいは p で表現し, その引数である確率変数によって区別するものとする.

情報理論では, 「事象の不確かさ」をエントロピーという量で表し, ある情報による不確かさの減少分が, その情報の「情報量」であると考える. 離散確率変数 X の実現値 x を観測した時の情報量は,

$$I(x) = -\log P(x)$$

で定義され, その期待値が Shannon エントロピーである [17]:

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

また, 連続変数に対する Shannon エントロピーは Shannon の微分エントロピーと呼ばれ, 次式で定義される:

$$H(X) = -\int_{x \in \mathcal{X}} p(x) \log p(x) dx. \tag{2.1}$$

本論文では主に連続変数に対する Shannon の微分エントロピー (2.1) を考える. ここで, Shannon の微分エントロピーの単純な計算例として, 平均 0, 分散 σ^2 の 1 次元正規分布のエントロピーを示す. この分布の密度関数は

$$p(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \tag{2.2}$$

であり, エントロピー (2.1) を定義通りに計算すると,

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left(-\frac{1}{2}\log 2\pi\sigma^2 - \frac{x^2}{2\sigma^2} \right) dx \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2} \end{aligned} \tag{2.3}$$

となる.

多次元分布の同時エントロピーについては, 次の性質が知られている.

命題 2.1 *([18])*
n 次元確率変数 $\boldsymbol{X}$ の同時エントロピーは, その周辺エントロピーの総和以下である. つまり, n 次元確率変数 $\boldsymbol{X}$ の第 i 成分を X_i とすると,

$$H(\boldsymbol{X}) \leq \sum_{i=1}^{n} H(X_i) \tag{2.4}$$

が成り立つ.

なお, エントロピーには Shannon による定義の他にも幾つかあり, 例えば Renyi エントロピー

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{x\in\mathcal{X}} p(x)^\alpha dx, \quad \alpha > 0 \tag{2.5}$$

もよく用いられる [19]. 特に次の事実が有用である:

定理 2.2
$\alpha = 2$ の場合の *Renyi* の *2* 次エントロピー

$$H_{R^2}(X) = -\log \int p(x)^2 dx$$

は, *Shannon* エントロピーの下界を与える*:*

$$H(X) \geq H_{R^2}(X).$$

証明
対数関数の凸性と *Jensen* の不等式 $\int p(x)\log p(x)dx \leq \log \int p(x)^2 dx$ から従う.□

エントロピーと並んで情報理論において重要な量として, 相互情報量がある. 確率変数 X と Y の相互情報量とは,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

で定義される量であり, X に関する曖昧さから Y を知った後に残る X の曖昧さを除いたもの, あるいはそれを X, Y について交換したものである.

確率密度 $p(x;\theta)$ で表される分布を基準として, 密度関数 $p(x;\theta')$ で表される分布がどれくらい離れているかを測る尺度として Kullback-Leibler ダイバージェンス (KL ダイバージェンス) あるいは相対エントロピーと呼ばれる量がある. これは

$$KL(\theta,\theta') = KL(p(x;\theta), p(x;\theta')) = \int p(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta')}dx$$

で定義される量である. 対称性や三角不等式を満たさないため数学的には距離ではないが, 非負性を満たし, ピタゴラスの定理が成立するなど好ましい性質を多く持ち, 統計多様体上での擬似的な距離 (疑距離) として広く用いられている. 上述の相互情報量は, 同時分布と周辺分布の積との KL ダイバージェンスとして表現出来る:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dx \\ &= KL(p(x,y), p(x)p(y)). \end{aligned}$$

また, KL ダイバージェンスは統計における最尤推定とも関連があり, 最尤推定は経験分布からの KL ダイバージェンスを最小にするようにパラメタを定める推定方法である. 実際, q を経験分布とすると,

$$\begin{aligned} \arg\min_\theta KL(q, p(x;\theta)) &= \arg\min_\theta \int q(x)\log\frac{q(x)}{p(x;\theta)}dx \\ &= \arg\min_\theta \left\{ -H(q) - \int q(x)\log p(x;\theta)dx \right\} \\ &= \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N}\log p(x_i;\theta) = \hat{\theta}_{MLE} \end{aligned}$$

となる．ここで上式最右辺の $\hat{\theta}_{MLE}$ は最尤推定量 (Maximum Likelihood Estimator) のことである．

2.2 Shannon エントロピーの推定

観測したデータを用いてそのデータが従う分布のエントロピーを推定することは, 独立成分分析 [20], 画像分析 [21], 多様体学習 [22] など多くの分野で重要である．

多くのエントロピー推定手法が提案されており, 分布を仮定せずにエントロピーを推定するノンパラメトリック法と, 例えば混合正規分布などで分布を近似した上でエントロピーを推定するパラメトリック法に大別される．ここではその柔軟さからノンパラメトリックな手法のみを考え, 代表的な手法としてカーネル密度推定に基づく方法と, k 近傍法に基づく手法, そして k 近傍法を改良した効率的な手法として Mean Nearest Neighbor(MNN) 法を示す．

問題設定としては, n 次元確率変数 $\boldsymbol{X} \in \mathbb{R}^n$ を考え, 密度関数 $p(\boldsymbol{x})$ に従う確率変数 $\boldsymbol{X}$ の実現値として観測データ $D = \{\boldsymbol{x}_i\}_{i=1}^N$ が得られたとき, $\boldsymbol{X}$ のエントロピー

$$H(\boldsymbol{X}) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} = E[-\log p(\boldsymbol{X})]$$

を推定するというものである．

カーネル密度推定に基づくエントロピー推定

エントロピーは確率密度関数に基づき定義されていることから, 密度関数をカーネル密度推定法により推定した上で, 積分を和で近似するというアプローチである．カーネル関数の選択や, 積分の近似方法に応じて幾つかのバリエーションがあるが, その中で最も基本的な手法として密度推定にガウスカーネルを用い, 積分の近似には Leave-one-out(LOO) 法を用いたものを紹介する．データ $D = \{\boldsymbol{x}_i\}_{i=1}^N$ が与えられたとき, $\boldsymbol{x}$ の密度を

$$\hat{p}(\boldsymbol{x}; D, h) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-||\boldsymbol{x} - \boldsymbol{x}_i||^2 / 2h^2\right) \tag{2.6}$$

で推定する. 推定した密度を用いてエントロピーを

$$H(\boldsymbol{X}) \approx \tilde{H}(\boldsymbol{X}) = -E[\log \hat{p}(\boldsymbol{X}; D, h)] \tag{2.7}$$

のように近似し, さらに $\hat{p}(\boldsymbol{x}; D, h)$ を $\hat{p}(\boldsymbol{x}_j; D\backslash\{\boldsymbol{x}_j\}, h)$ で置き換えて, LOO 法によって期待値の計算を

$$\tilde{H}(\boldsymbol{X}) \approx \hat{H}(\boldsymbol{X}) = -\frac{1}{N}\sum_{j=1}^{N} \log \hat{p}(\boldsymbol{x}_j; D\backslash\{\boldsymbol{x}_j\}, h)$$

のように近似する. ここで, $D\backslash\{\boldsymbol{x}\}$ は, 集合 D から要素 $\boldsymbol{x}$ を除いて得られる集合を意味する.

k 近傍法に基づくエントロピー推定

エントロピー推定のもう一つの代表的な方法として, 各観測データ点の k 近傍を考えたときにそれらがどのくらい離れているか, という情報を用いるものがある. 基本的なアイディアは, $\log p(\boldsymbol{x}_i)$ を $\boldsymbol{x}_i$ とその k 近傍との距離 ϵ の密度関数 $p_{ik}(\epsilon)$ を介して推定するというものである.

点 $\boldsymbol{x}_i$ から $r \in [\epsilon, \epsilon + d\epsilon]$ の距離以内のところにデータが 1 点存在する確率を $p_{ik}(\epsilon)d\epsilon$ で表す. つまり, その他の $k-1$ 点は $\boldsymbol{x}_i$ から ϵ 未満の距離に存在し, その他の $N-k-1$ 点は $\epsilon + d\epsilon$ より遠くに存在する確率である.

データ点 $\boldsymbol{x}_i$ を中心とする ϵ-球内に点が存在する確率を $p_i(\epsilon)$ で表す:

$$p_i(\epsilon) = \int_{||\boldsymbol{x}-\boldsymbol{x}_i||<\epsilon} p(\boldsymbol{x})d\boldsymbol{x}.$$

上述の, $\boldsymbol{x}_i$ と他のデータ点との関係は,

1. 他のデータが 1 個, $r \in [\epsilon, \epsilon + d\epsilon]$ に存在する,
2. 他のデータが $k-1$ 個, ϵ 未満の距離に存在する,
3. 他のデータが $N-k-1$ 個, $\epsilon + d\epsilon$ より離れたところに存在する,

という 3 通りがあるので, $p_{ik}(\epsilon)$ は 3 項分布

$$p_{ik}(\epsilon) = \frac{(N-1)!}{1!(k-1)!(N-k-1)!}\frac{dp_i(\epsilon)}{d\epsilon}p_i^{k-1}(1-p_i)^{N-k-1}$$

で表される. ここで, $\log p_i(\epsilon)$ の $p_{ik}(\epsilon)$ に関する期待値をとると,

$$\begin{aligned} E_{p_{ik}}[\log p_i] &= \int_0^\infty p_{ik}(\epsilon) \log p_i(\epsilon) d\epsilon \\ &= k \binom{N-1}{k} \int_0^1 p_i^{k-1} (1-p_i)^{N-k-1} \log p_i dp_i \\ &= \psi(k) - \psi(N) \end{aligned}$$

となる. $\psi(x)$ は digamma 関数 $\phi(x) = \Gamma'(x)/\Gamma(x)$ である.

ここで, 密度関数 $p(\boldsymbol{x})$ が $\boldsymbol{x}_i$ 中心の ϵ 球内でほぼ定数であると仮定すると, c_n を n 次元単位球の体積 $c_n = \pi^{n/2}/\Gamma(1+n/2)$ として,

$$p_i(\epsilon) \sim c_n \epsilon^n p(\boldsymbol{x}_i) \tag{2.8}$$

であり, これを $\log p_i(\epsilon)$ に代入して

$$-\log p(\boldsymbol{x}_i) \sim \psi(N) - \psi(k) + \log(c_n) + n E_{p_{ik}}[\log \epsilon]$$

を得る. これを全てのデータ点に関して (経験分布を用いて) 期待値をとることで, エントロピーの推定量

$$H_k(\boldsymbol{X}) = \psi(N) - \psi(k) + \log(c_n) + \frac{n}{N} \sum_{i=1}^{N} \log \epsilon_i \tag{2.9}$$

を得る.

この推定量に現れる ϵ_i は, i 番目のデータ $\boldsymbol{x}_i$ から k 番目に近い他のデータまでの距離である. この推定方法は, k を適切に定めると高精度な推定が実現できることが知られている. また, 真の密度関数 $p(x)$ に対する条件

$$\int p(x)(\log p(x))^2 dx < \infty \tag{2.10}$$

の下で, H_k は平均二乗の意味で一致性を持つ [23]:

$$\lim_{n\to\infty} E[(H_k(\boldsymbol{X}) - H(\boldsymbol{X}))^2] = 0. \tag{2.11}$$

一方, 適切な k を設定しなければならないということと, 推定のためにデータのソートが必要になるという問題点がある.

次に, 文献 [24] において提案された k 近傍法の拡張に基づく効率的なエントロピー推定手法を紹介する.

Mean Nearest Neighbor(MNN) 法によるエントロピー推定

これは, 式 (2.9) を全ての k, $1 \leq k \leq N-1$ について平均するというシンプルなアイディアに基づく方法である. k 近傍法に基づくエントロピーの推定量において, N 個の観測データがある場合には k は 1 から $N-1$ まで動かすことができる. これらを全て加えて平均した

$$\begin{aligned} H_{\mathrm{MNN}}(\boldsymbol{X}) &= \frac{1}{N-1}\sum_{k=1}^{N-1} H_k(\boldsymbol{X}) \\ &= \log(c_n) + \psi(N) + \frac{1}{N-1}\sum_{k=1}^{N-1}\left(-\psi(k) + \frac{n}{N}\sum_{i=1}^{N}\log\epsilon_{i,k}\right) \\ &= \log(c_n) + \psi(N) \\ &\quad -\frac{1}{N-1}\sum_{k=1}^{N-1}\psi(k) + \frac{n}{N(N-1)}\sum_{i\neq j}\log||\boldsymbol{x}_i - \boldsymbol{x}_j|| \end{aligned}$$

で Mean Nearest Neighbor(MNN) エントロピー推定量を定義する. ここで, $\epsilon_{i,k} = ||\boldsymbol{x}_i - \boldsymbol{x}_{(i,k)}||$ であり, $\boldsymbol{x}_{(i,k)}$ はデータ $\boldsymbol{x}_i$ から k 番目の距離にある点である. 和の順序を入れ替えて, 先に k に関する和を計算すると, $\sum_{k=1}^{N-1}||\boldsymbol{x}_i - \boldsymbol{x}_{(i,k)}||$ は全ての k を考えることになるので, $\boldsymbol{x}_i$ 以外の点との距離を足し合わせることに他ならない. したがって,

$$H_{\mathrm{MNN}} = \frac{n}{N(N-1)}\sum_{i\neq j}\log||\boldsymbol{x}_i - \boldsymbol{x}_j|| + const. \tag{2.12}$$

を得る.

この推定量の一つの大きな利点は, カーネル密度推定や k 近傍法に基づく方法と違って, 調整すべきパラメタが全く存在しない点である. また, データのソートの必要もない. 一方, 式 (2.8) の仮定は k が大きな値の時には成立せず, それに伴い推定値の誤差が大きくなる可能性がある.

2.2.1 エントロピー推定実験

MNN 法は, カーネル密度推定に基づく手法及び従来の k 近傍に基づく手法と比較して計算効率の面で優れている. さらに, 推定量のバイアスは適切に近傍数 k を選択した場合の k 近傍法にわずかに劣るものの, 推定の分散は従来の k 近傍法よりも大幅に小さいことが確認されている. この性質は, エントロピーを高速に計算する必要がある場合や, エントロピー

表 2.1: LOO 法と MNN 法によるエントロピー推定の比較.

	LOO	MNN
絶対誤差の平均	0.1406625	0.1115902
絶対誤差の標準偏差	0.07178894	0.04636109
絶対誤差/理論値	0.3021533	0.2391081

を勾配法によって最適化する必要がある場合には望ましいものである. k 近傍法との比較は [24] において行われているため, ここでは LOO 法との簡単な比較実験の結果を示す.

LOO 法では, まず式 (2.6) によって確率密度関数を推定する. カーネルのバンド幅 h は, Silverman の経験則と呼ばれる方法で定めることにする [25]. この LOO 法による推定量と MNN 法による推定量を, 1 次元指数分布に従う確率変数の観測値からのエントロピー推定に適用して比較する. 指数分布の密度関数は $p(x;\mu) = \frac{1}{\mu}e^{-\frac{x}{\mu}},\ x \geq 0$ であり, そのエントロピーは $H(X) = \log\mu + 1$ という簡単な形をしている. この分布から $N = 500$ 個のサンプルを生成した. パラメタの値は $\mu = (0.2, 0.4, \ldots, 2.0)$ の 10 種類を用いて, それぞれのパラメタでの 10 回データを生成して理論的なエントロピーと推定エントロピーの値を比較したのが表 2.1 である. この表には, 理論値と推定値との誤差の絶対値の平均と標準偏差, 及び誤差の絶対値を理論値で割った値の平均が記してある. 表 2.1 から, MNN 法は LOO 法よりも正確であることがわかる. さらに, 標準偏差に大きな差があることもわかる. このように, MNN 法の大きな利点の一つは推定のばらつきが小さいということであり, 特に本論文第 4 章で提案する線型次元削減手法においては推定量の導関数を用いて最適化を行うため, 標準偏差が小さいことは望ましい性質である.

以上の考察から, 本論文におけるエントロピーの推定方法として MNN 法を採用する.

第3章　距離構造の学習問題とカーネル法

本章の前半では, 機械学習の分野における重要な課題である, データ分布空間における距離構造の学習問題について述べる. 一般に高次元の特徴空間の距離構造をカーネル関数によって非明示的に定めることで, 非線型特徴空間での学習を行う手法としてカーネル法がある. 本章の後半では, カーネル法に用いるカーネル関数の基本的な性質を述べる.

3.1　教師付きの距離構造の学習に関する研究

本論文で扱う教師付きの距離構造学習手法は数多く提案されている. 本節では, まず本論文で提案する手法と同様に情報理論に基づく手法を紹介し, 次に本論文で提案する手法と形式的に類似した手法を紹介する.

教師付き距離学習手法の分類は様々な観点から可能である. 例えば大域的な距離構造を学習するもの, 局所的な距離構造を学習するものという観点での分類や, 教師データとしてクラスラベルが与えられている場合と, データ対が類似あるいは非類似関係にあるという情報が与えられている場合という観点での分類ができる. あるいは, Support Vector Machine (SVM) のように学習の基準としてマージン最大化に基づくもの [26; 27; 28] と, Fisher の判別分析 (Fisher Discriminant Analysis: FDA [5]) や, FDA をデータの局所性を反映するように拡張した局所 Fisher 判別分析 (Local Fisher Discriminant Analysis: LFDA [29]) のように共分散構造に基づくものという分類ができる. 本論文では共分散構造に基づくものに注目する. FDA では, 各クラスで共分散構造が同一の正規分布が仮定される. エントロピーや相互情報量といった情報論的な量を考える事で, 共分散構造ベースの方法を一般化することができる. 例えば, 式 (2.3) ではエントロピーと正規分布の分散が直接結びついている. 正規分布は 2 次までのモーメントによって完全に規定される分布であり, エントロピーも 2 次のモー

メントを用いて記述される．正規分布以外の分布では一般にはエントロピーが分散のみで表現されることはなく, この意味でエントロピーを基準とする学習方法は分散ベースの手法の一般化であると考えられる．

Shannon の微分エントロピーに基づく距離構造学習手法はエントロピーか相互情報量の推定を必要とする．エントロピーは変換されたデータの密度関数から計算されるので, 密度推定に基づく種々の方法が提案されている．密度推定はパラメトリック手法とノンパラメトリック手法に分類される．パラメトリック手法の多くは, その扱いの容易さから Gaussian Mixture Model(GMM) がしばしば用いられる．線型変換 $A : \mathbb{R}^n \to \mathbb{R}^m$ により, データ $\boldsymbol{x}$ が $\boldsymbol{z} = A^T\boldsymbol{x}$ に変換されるとする．$\boldsymbol{Z}$ で変換後のデータ $\boldsymbol{z}$ を実現値とする確率変数を表すとして, $\boldsymbol{Z} = A^T\boldsymbol{X}$ の分布を複数のガウス分布の混合によって近似する．つまり, 混合比を表すパラメタと, 各混合要素の正規分布の平均, 共分散パラメタを学習することで分布を近似する．例えば, [30] では相互情報量 $I(\boldsymbol{Z}; Y)$ を GMM を介して計算し, 勾配法により $I(\boldsymbol{Z}; Y)$ を最大化することで判別的なデータ変換行列 A を学習する手法が提案されている．また, [31; 32; 33] では GMM により条件付き確率 $p(A^T\boldsymbol{x}|y)$ を近似してから Bayes の定理により $p(y|A^T\boldsymbol{x})$ を推定し, 条件付き尤度

$$L(A) = \sum_{i=1}^{N} p(y_i|A^T\boldsymbol{x}_i)$$

を勾配法で最適化する距離構造学習手法が提案されている．

一方, ノンパラメトリック手法では, データの分布に一切の仮定を設けず, 例えばカーネルバンド幅のようなごく少数のパラメタのみが事前に設定される．ノンパラメトリックな密度推定に基づく情報論的な距離学習の例として, 最近文献 [34] において提案された手法を紹介する．この文献では, 制約付きのエントロピー最大化問題

$$\max_A H(A^T\boldsymbol{X}) \quad s.t. \quad H(A^T\boldsymbol{X}|Y) = const.,\ A^TA = I_m$$

を解くことで距離構造の学習あるいは次元削減を行う方法を提案している．この手法は本論文第 4 章で提案する枠組みと類似している．この文献では変換後のデータの分布に強い仮定をおくことで制約付き最適化問題を一般化固有値問題に帰着して大域的最適解を近似的に求めている．さらに, エントロピーとしては Shannon のエントロピーではなく, 計算を容易に行うために 2 次の Renyi エントロピー (2.5) を用いている．これは, 最大化の目的関数が Shannon エントロピーであることから, 定理 2.2 に基づ

きその下界を最大化するというアプローチである. Renyi エントロピーは ICA などの教師無し学習の枠組みで Shannon エントロピーの代替としての利用が提案され [35], Shannon エントロピー推定の困難さを回避する手法として広く利用されている [36; 37; 38; 39]. しかし, 情報論的な学習手法の理論的背景は Shannon の微分エントロピーで記述されるものであり, Renyi エントロピーはあくまで本来の Shannon エントロピーの近似であることに注意しなければならない.

本節で述べたように, 情報論的な距離構造学習手法は既に数多く提案されている. しかし, 提案されているノンパラメトリックなアプローチは Shannon エントロピーではなく Renyi エントロピーの推定をしているものが多い. 第 2 章で示したように, 高速な Shannon エントロピーの推定手法である MNN 法を用い, また第 4 章で述べるように同時エントロピーを周辺エントロピーの和で近似することで, Shannon エントロピーを効率的に推定することができる.

次に, 情報論的な意味合いは薄いが, 本論文で提案する手法と類似した先行研究として k 近傍法を確率的に拡張した手法を 2 つ紹介する.

教師付きの線型距離構造学習問題は, 判別のために有効な距離行列 $W = AA^T$ を学習することが目的である. これは, データを変換する行列 A を学習し, A によって写像された空間におけるユークリッド距離を用いて判別処理を行うことと等価である. このとき, 変換行列 A としては, データが各クラスにおいて小さい領域に集中して分布し, 異なるクラスのデータ同士は遠く離れて分布することが望ましい. 文献 [6] では, データ $\boldsymbol{x}_i$ が他のデータ $\boldsymbol{x}_j$ を, 自らの近傍であるとする確率を

$$p_A(\boldsymbol{x}_j|\boldsymbol{x}_i) = \frac{\exp(-||A^T\boldsymbol{x}_j - A^T\boldsymbol{x}_i||^2)}{\sum_{k\neq i}\exp(-||A^T\boldsymbol{x}_k - A^T\boldsymbol{x}_i||^2)}, \quad p_A(\boldsymbol{x}_i|\boldsymbol{x}_i) = 0 \tag{3.1}$$

で定義した. そして, $\boldsymbol{x}_i$ と同じクラスに属するデータ集合を C_i で表し, 目的関数

$$f(A) = \sum_{i=1}^{N}\sum_{j\in C_i} p_A(\boldsymbol{x}_j|\boldsymbol{x}_i)$$

を勾配法により最大化することで変換行列 A を求めることを提案した. データ間の距離は, $\sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T AA^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}$ で計算される. この距離構造の学習手法を, Neighborhood Component Analysis (NCA) と呼ぶ.

文献 [7] では, NCA の教師付き学習としての拡張として, MCML (Maximally Collapsing Metric Learning) と呼ばれる手法が提案されている.

$p_0(\boldsymbol{x}_j|\boldsymbol{x}_i)$ を, クラスラベル情報を用いて

$$p_0(\boldsymbol{x}_j|\boldsymbol{x}_i) \propto \begin{cases} 1, & \boldsymbol{x}_j \in C_i, \\ 0, & \boldsymbol{x}_j \notin C_i \end{cases}$$

で定義される理想の分布とする. MCML では, p_A と理想の分布 p_0 との KL ダイバージェンス

$$\sum_{i=1}^{N} KL(p_0(\boldsymbol{x}|\boldsymbol{x}_i), p_A(\boldsymbol{x}|\boldsymbol{x}_i))$$

を最小化の目的関数とする. 距離行列 AA^T は半正定値行列なので, 目的関数は半正定値性の制約条件を与えた上で, 勾配法を用いて最小化できる. 具体的には, 勾配法を 1 ステップ行う毎に, 更新した行列 A を用いて AA^T の固有値を計算し, 負の固有値を 0 に置き換える処理を行う.

3.2 カーネル法の理論

実際の判別問題においては, 線型の判別平面で分離出来るデータは少なく, 何らかの方法で非線型の判別曲面を構成する判別器が必要となる. 非線型判別のための一つのアプローチとしては, k 近傍法やニューラルネットワークといった非線型判別器を用いる方法であるが, もう一つのアプローチとして, データを非線型な写像により特徴空間に写像した上で, 線型の判別器を適用するというものがある. この手法のメリットとしては, 線型の判別器は非線型の判別器と比較して実装が容易なことが多く, データを変換してしまえば既存のソフトウェアがそのまま使えることが多いという点がある. また, 線型判別器は理論的解析も比較的容易であり, 汎化性や収束性の議論という面でもメリットがある. 一方, 線型判別器により十分な性能が期待できるような特徴空間は一般に高次元であり, 無限次元空間にも成り得る. そのため, 全てのデータに対してこうした高次元写像を陽に計算することは現実的ではない. また, 近年の情報技術の発展によって様々なデータが電子的に管理されるようになり, テキストデータや時系列データからマルチメディアデータ, DNA 配列などの生物学的データまで, 処理の対象となるデータは多岐に及ぶ. こうしたデータの中には, 従来のデータのように実数ベクトルなどの一般的な表現を仮定出来ないものもあり, 主に実数ベクトル表現された入力データを仮定して構成されている多くの学習アルゴリズムの適用の範囲外となっている.

カーネル法は, 線型のモデルで非線型の問題を解くための上述のアプローチを, 特徴空間への写像を陽に計算することなく実行するための手法の総称である. カーネル法では, 利用する判別手法がデータの内積のみを用いて記述出来る場合, 特徴空間におけるデータの内積と同値な 2 変数関数を用いて判別手法に必要な種々の計算を実行する. 以下, カーネル関数に要求される性質と, 有限次元での計算を可能とする根拠であるリプレゼンター定理の説明をする. また, カーネル関数で定まる類似度と距離との関係を簡単に述べる. 次に, カーネル関数がその凸結合に関して閉じているという有用な性質を述べる. これは, 後述の Multiple Kernel Learning の根拠となる事実である. さらに, データの生成モデルがわかっている時に, モデルに基づきカーネル関数を構成する方法として Fisher カーネルを紹介する. Fisher カーネルは, 第 5 章で述べるデータの生成モデルに基づく距離構造の学習において利用する.

なお, カーネル法全般に関する成書としては [40] や [41] がある. また, 本論文では触れないが, カーネル法の理論的な背景となる再生核ヒルベルト空間の理論に関しては [42] が詳しい.

3.2.1 カーネル法の導入

説明変数 $\boldsymbol{x} \in \mathbb{R}^n$ が与えられたとき, 応答変数 $y \in \mathbb{R}$ を

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

の形で予測しようとするのが, 回帰分析における線型モデルの考え方である. この線型モデルは非常に単純であり, 説明変数 $\boldsymbol{x}$ と応答変数 y の直線的な関係しか捉えることができない. そこで, データを

$$\begin{aligned} \phi : \mathbb{R}^n &\rightarrow \mathcal{H} \\ \boldsymbol{x} &\mapsto \phi(\boldsymbol{x}) \end{aligned}$$

なる写像 ϕ により, ある空間 $\mathcal{H}$ に写像した上で線型モデルを考える. ここで, 変数 $\boldsymbol{x}$ が写像される空間 $\mathcal{H}$ を特徴空間と呼ぶ. 一般には特徴空間は実ベクトル空間と同型な空間でなくてもよいが, 簡単のためここでは $\mathcal{H} \subset \mathbb{R}^m$ として, $\mathcal{H}$ の元は $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ というベクトル表現ができるとする. すると, 特徴空間における線型モデルが

$$y = f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.2}$$

のように得られる. ここで, 関数 f はある関数空間 $\mathcal{F}$ の元であるとする. 線型モデルを考えるにあたって, もとの入力 $\boldsymbol{x}$ は実ベクトルである必要があった. 一方, 写像 ϕ によって特徴空間における線型モデルを考えるときは, 特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ が実ベクトルであれば元の $\boldsymbol{x}$ は実ベクトルである必要がない. つまり, 特徴空間への写像を考える事で, $\boldsymbol{x}$ としてベクトル以外の対象も考えることができるという利点がある. 例えば, 遺伝子解析やグラフ解析の分野では, 文字列やグラフなどを対象として処理を行うが, これらに直接内積を定めることは困難である. しかし, こうした対象に対しても何らかの手法で特徴量を抽出してベクトル表現が可能であれば, 上記のように特徴空間において内積を計算することが出来る.

ここで, $\mathcal{X} \times \mathcal{X}$ 上のカーネル関数を, 特徴ベクトルを用いて定義する.

定義 3.1
$\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ とする. この 2 つの元に対するカーネル関数 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ を, $\boldsymbol{x}_1, \boldsymbol{x}_2$ それぞれの特徴ベクトル同士の内積

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{\phi}(\boldsymbol{x}_1)^T \boldsymbol{\phi}(\boldsymbol{x}_2) \tag{3.3}$$

で定義する.

式 (3.2) により, $\mathcal{H}$ の元 $\boldsymbol{w}$ と $\boldsymbol{\phi}(\boldsymbol{x})$ の内積により関数 $f(\boldsymbol{x})$ の値が決まる.

上述のように, カーネル関数を特徴空間における特徴ベクトルの内積で定義した. こうして定義したカーネル関数は半正定値性を持つ. 実際, 任意の N 個のデータ $\{\boldsymbol{x}_i\}_{i=1}^N$ から計算されるグラム行列

$$\begin{aligned} K &= \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_1, \boldsymbol{x}_N) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \boldsymbol{\phi}(\boldsymbol{x}_1) & \cdots & \boldsymbol{\phi}(\boldsymbol{x}_N)^T \boldsymbol{\phi}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_1)^T \boldsymbol{\phi}(\boldsymbol{x}_N) & \cdots & \boldsymbol{\phi}(\boldsymbol{x}_N)^T \boldsymbol{\phi}(\boldsymbol{x}_N) \end{pmatrix} \end{aligned}$$

を用いて任意の N 次元実ベクトル $\boldsymbol{a} \in \mathbb{R}^N$ に関する二次形式を考えると,

$$\begin{aligned} \boldsymbol{a}^T K \boldsymbol{a} &= \boldsymbol{a}^T \begin{pmatrix} \boldsymbol{\phi}^T(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}^T(\boldsymbol{x}_N) \end{pmatrix} \Big(\boldsymbol{\phi}(\boldsymbol{x}_1) \quad \cdots \quad \boldsymbol{\phi}(\boldsymbol{x}_N) \Big) \boldsymbol{a} \\ &= \| \Big(\boldsymbol{\phi}(\boldsymbol{x}_1) \quad \cdots \quad \boldsymbol{\phi}(\boldsymbol{x}_N) \Big) \boldsymbol{a} \|^2 \geq 0 \end{aligned}$$

である. 逆に, 半正定値性を持つ任意の対称関数は, 何らかの特徴ベクトルの内積とみなすことが出来る. これは, Mercer の定理 ([41; 40]) と呼ばれる定理によって保証される.

カーネル法が広く用いられている理由の一つが, 一定の正則化条件の下で, 判別関数が学習サンプル点で評価したカーネル関数のみで記述できるという性質である.

定理 3.2 *(リプレゼンター定理 [41; 40])*
判別関数 *(3.2)* を, そのノルムに関する正則化項 $\lambda||f||_{\mathcal{F}}^2$ を含むコスト関数

$$R_{\mathrm{reg}}(f) = R(\{f(\boldsymbol{x}_i), y_i\}_{i=1}^N) + \lambda||f||_{\mathcal{F}}^2 \tag{3.4}$$

を最小化することで学習する問題を考える. ここで, $R : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R} \cup \{\infty\}$ は任意の損失関数であり, $\lambda > 0$ である. このとき, $R_{\mathrm{reg}}(f)$ を最小にする関数 $f \in \mathcal{F}$ は, 適当な $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)$ によって

$$f(\boldsymbol{x}) = \sum_{i=1}^N \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) \tag{3.5}$$

の形でかける.

リプレゼンター定理は, 判別関数 (3.2) における係数ベクトル $\boldsymbol{w}$ を,

$$\boldsymbol{w} = \sum_{i=1}^N \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) \tag{3.6}$$

の形に限って考えてもよいということを意味している.

カーネルで定まる類似度と距離との関係

カーネル関数は, 特徴空間におけるデータの内積を与えるものと理解できる. 内積は一種の類似度と考えられる. 通常, 空間においてその元同士の内積が定義されると, 自然にノルムが定まり, 同時にその空間における距離が定義される. ここで, カーネル関数で定まる類似度と距離との関係を簡単に述べる.

距離として最も馴染み深いものはユークリッド距離である. 点 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ の間のユークリッド距離は,

$$d_E(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2 = \boldsymbol{x}_1^T\boldsymbol{x}_1 + \boldsymbol{x}_2^T\boldsymbol{x}_2 - 2\boldsymbol{x}_1^T\boldsymbol{x}_2$$

で定義される. 一方, 特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ をユークリッド空間上の点とみなすと, その距離は

$$||\boldsymbol{\phi}(\boldsymbol{x}_1) - \boldsymbol{\phi}(\boldsymbol{x}_2)||^2 = ||\boldsymbol{\phi}(\boldsymbol{x}_1)||^2 + ||\boldsymbol{\phi}(\boldsymbol{x}_2)||^2 - 2\boldsymbol{\phi}(\boldsymbol{x}_1)^T\boldsymbol{\phi}(\boldsymbol{x}_2)$$

となる. 上式右辺の第 3 項は特徴ベクトルの内積で表現されていることから, 特徴ベクトルの距離とカーネル関数との関係は

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{2}\left(-||\boldsymbol{\phi}(\boldsymbol{x}_1) - \boldsymbol{\phi}(\boldsymbol{x}_2)||^2 + ||\boldsymbol{\phi}(\boldsymbol{x}_1)||^2 + ||\boldsymbol{\phi}(\boldsymbol{x}_2)||^2\right)$$

で与えられる. ここで, $||\boldsymbol{\phi}(\boldsymbol{x}_i)||^2 = k(\boldsymbol{x}_i, \boldsymbol{x}_i)$ に注意すると, カーネル関数値から距離を計算することができる:

$$D_{ij} = K_{ii} + K_{jj} - 2K_{ij}. \tag{3.7}$$

ここで, D_{ij} は特徴空間におけるデータ $\phi(\boldsymbol{x}_i)$ と $\phi(\boldsymbol{x}_j)$ の間の距離であり, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ である.

逆に, D_{ij} を (i, j) 成分とする距離行列 D からカーネル関数を計算する方法を考える. このとき, 特徴ベクトル全体を平行移動してもお互いの距離は不変であるが, 原点の位置が変わるために, 内積の値が変化することに注意する. つまり, 特徴ベクトル間の距離を定めるだけではカーネル関数値は一意に定まらないので, ここでは特徴ベクトルのサンプル平均が原点に一致するという制約を加える:

$$\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i) = \mathbf{0}.$$

この制約から,

$$\sum_{i=1}^{N} \boldsymbol{\phi}^T(\boldsymbol{x}_j)\boldsymbol{\phi}(\boldsymbol{x}_i) = \sum_{i=1}^{N} K_{ij} = 0 \tag{3.8}$$

が得られる. 式 (3.7) を i に付いて足し合わせると

$$\sum_{i=1}^{N} D_{ij} = \sum_{i=1}^{N} K_{ii} + NK_{jj}$$

となり, D の全ての成分の総和は

$$\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij} = 2N\sum_{i=1}^{N} K_{ii}$$

である. 以上より, 距離行列を用いてカーネル行列が

$$K_{ij} = -\frac{1}{2}D_{ij} + \frac{1}{2N}\sum_{l=1}^{N} D_{lj} + \frac{1}{2N}\sum_{l=1}^{N} D_{il} - \frac{1}{2N^2}\sum_{l=1}^{N}\sum_{m=1}^{N} D_{lm}$$

で得られる.

3.3 カーネル関数の設計

正定値かつ対称な 2 変数関数は, カーネル法におけるカーネル関数に用いることができる. カーネル関数は, カーネル関数同士の和, 積, テンソル積など, 広いクラスの演算について閉じており, 複数のカーネル関数を組み合わせて新しいカーネル関数を構成することが出来る. また, データの生成モデルがわかっている場合, 生成モデルを元にカーネル関数を設計する方法も考えられている. 本小節では, 本論文で扱うカーネルの凸結合の最適化と, 生成モデルに基づくカーネル関数の一つである Fisher カーネルについて簡単に述べる.

3.3.1 Multiple Kernel Learning

ここでは, 第 4 章で扱う Multiple Kernel Learning (MKL) の理論的根拠として, カーネル関数の凸結合として得られる関数が再びカーネル関数になることを述べ, MKL の代表的な先行研究を幾つか紹介する.

カーネル法は多くの問題に適用されて成果をあげているが, その適用にあたって扱う問題に応じて適切にカーネル関数を選択し, そのパラメタを選択しなければよい性能が得られないという難しさがある. カーネル関数を与えられたデータを用いて最適に設計する手法は数多く提案されており, MKL はその代表的な手法としてよく研究されている. MKL の研究は, Lanckriet らによる半正定値計画問題 (Semi-definite Programming; SDP) を用いた定式化をきっかけとして盛んに研究されている.

次のようなパラメトライズされたカーネル関数族を考える:

$$\mathcal{K} = \{k(\,\cdot\,,\,\cdot\,;\lambda); \lambda \in \Lambda\}.$$

ここで, λ はパラメタ空間 Λ に値をとるものとし, この値が $\mathcal{K}$ 内のカーネル関数を特徴付けるものとする. 例えば, ガウスカーネルの族

$$k(\boldsymbol{x}_j, \boldsymbol{x}_i; \lambda) = \exp\left(-\lambda||\boldsymbol{x}_j - \boldsymbol{x}_i||^2\right),$$

を考えるときは, λ は精度パラメタに対応し, $\Lambda = \{\lambda \in \mathbb{R}; \lambda > 0\}$ である. ここで, 族 $\mathcal{K}$ から取り出した S 個の要素カーネル関数 $k(\,\cdot\,,\,\cdot\,;\lambda_s),\ s = 1, \ldots, S$ の凸結合により,

$$k(\,\cdot\,,\,\cdot\,;\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{s=1}^{S} \beta_s k(\,\cdot\,,\,\cdot\,;\lambda_s), \quad \sum_{s=1}^{S} \beta_s = 1,\ \beta_s \geq 0,\ \ s = 1, \ldots, S \tag{3.9}$$

の形で新しい関数を定義する. こうして定義した新しい関数について, 次の命題が成り立つ:

命題 3.3
関数 *(3.9)* は半正定値対称なカーネル関数であり, ある特徴空間における特徴ベクトルの内積を定義する.

証明
対称性は明らかである. ある特徴空間における特徴ベクトルの内積を定めることを示す. $\boldsymbol{\phi}_s$ で s 番目の要素カーネル関数 k_s に対応する特徴空間の特徴ベクトルを表すとして, 任意のデータ $\boldsymbol{x}_i, \boldsymbol{x}_j$ に対して,

$$\begin{aligned}
&\beta_1 k_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + \cdots + \beta_S k_S(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&= \beta_1 \boldsymbol{\phi}_1(\boldsymbol{x}_i)^T \boldsymbol{\phi}_1(\boldsymbol{x}_j) + \cdots + \beta_S \boldsymbol{\phi}_S(\boldsymbol{x}_i)^T \boldsymbol{\phi}_S(\boldsymbol{x}_j) \\
&= \begin{pmatrix} \sqrt{\beta_1}\boldsymbol{\phi}_1(\boldsymbol{x}_i)^T & \cdots & \sqrt{\beta_S}\boldsymbol{\phi}_S(\boldsymbol{x}_i)^T \end{pmatrix} \begin{pmatrix} \sqrt{\beta_1}\boldsymbol{\phi}_1(\boldsymbol{x}_j) \\ \vdots \\ \sqrt{\beta_S}\boldsymbol{\phi}_S(\boldsymbol{x}_j) \end{pmatrix}
\end{aligned}$$

であることから, 要素カーネル関数の凸結合によって特徴ベクトル

$$\begin{pmatrix} \sqrt{\beta_1}\boldsymbol{\phi}_1(\boldsymbol{x}_j) \\ \vdots \\ \sqrt{\beta_S}\boldsymbol{\phi}_S(\boldsymbol{x}_j) \end{pmatrix}$$

に対応するカーネル関数が得られたことになる. □

Lanckriet らは文献 [9] において, SVM の判別関数

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

で用いるカーネル関数 k を複数のカーネル関数の凸結合で置き換えた判別関数

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i \sum_{s=1}^{S} \beta_s k_s(\boldsymbol{x}_i, \boldsymbol{x}) + b,$$

を考え, 半正定値計画問題を用いて判別器のマージンを $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^N$ と $\boldsymbol{\beta} = \{\beta_s\}_{s=1}^S$ 両方について最大化するという定式化を行った. 本論文ではこの MKL 手法を, SDP-MKL と呼ぶ.

近年, SVM の汎化誤差がマージンのみでなく, 特徴空間においてデータを包含する最小の球の半径 (radius) にも依存することと, カーネル関数の組み合わせにより特徴空間におけるデータ分布が変化することから, マージンと半径の両方を最適化する MKL の枠組みである R-MKL が提案された [43].

これまでに提案されている MKL 手法の殆どは, 上述のように判別器として SVM を利用しており, そのマージン最大化基準で判別器の係数 $\boldsymbol{\alpha}$ とカーネル関数の凸結合の係数 $\boldsymbol{\beta}$ を最適化するというアプローチを取っている. 一方, 少数ながら SVM 以外の判別器をベースとした MKL も存在する. 文献 [44] では, [9] にならい, Kernel Fisher Discriminatn Analysis (KFDA: [45]) のための MKL である KFDA-MKL を, 半正定値計画問題を用いて定式化している.

その他にも, 例えば SimpleMKL [46] や SILP [47] のように, 多数のカーネル関数の凸結合を効率的に最適化する手法が提案されているが, 得られる判別関数の性能は SDP-MKL による学習で得られる判別関数の性能と同程度である.

次節で, 本論文で提案する MCEM アルゴリズムと, SDP-MKL, R-MKL, KFDA-MKL とを実験的に比較する.

3.3.2 Fisher カーネル

観測されるデータを生成する確率モデル (生成モデル) が既知の場合に, その確率モデルをもとにカーネル関数を設計する方法が考えられている. ここではその代表例として, Fisher カーネル [48] を紹介する. データ $\boldsymbol{x}$ の生成モデルを確率分布 $p(\boldsymbol{x};\boldsymbol{\theta})$ とする. ここで $\boldsymbol{\theta} \in \mathbb{R}^m$ はモデルのパラメタベクトルである. 確率分布の対数 $\log p(\boldsymbol{x};\boldsymbol{\theta})$ をパラメタの各成分で偏微分して得られる関数

$$\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta}) = \left(\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_m} \right) \tag{3.10}$$

$$= \frac{1}{p(\boldsymbol{x};\theta)} \left(\frac{\partial p(\boldsymbol{x};\theta)}{\partial \theta_1}, \cdots, \frac{\partial p(\boldsymbol{x};\theta)}{\partial \theta_m} \right) \tag{3.11}$$

を, スコア関数と呼ぶ. また, $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})^T$ を $p(\boldsymbol{x};\boldsymbol{\theta})$ で平均した行列

$$G(\boldsymbol{\theta}) = E_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x};\boldsymbol{\theta})^T] \tag{3.12}$$

を Fisher 情報行列と呼ぶ. ここで, データ $\boldsymbol{x}_i, \boldsymbol{x}_j$ に対して

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j;\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{x}_i;\boldsymbol{\theta})^T G^{-1}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x}_j;\boldsymbol{\theta}) \tag{3.13}$$

で定義した関数は, 特徴ベクトルの内積の形をしているので正定値対称であり, Fisher カーネルと呼ばれる. Fisher カーネルはデータの生成モデルを利用できる場合のカーネル関数として有力な候補であり, その理論的な性質の解析や, 生成モデルに潜在変数を含む場合への拡張も行われている [49; 50]. なお, 理論的には Fisher カーネルは (3.13) で定義されるが, 実際上は Fisher 情報行列 (3.12) を計算することは困難な場合が多い. そこで, Fisher カーネル (3.13) において $G(\boldsymbol{\theta})$ を m 次元単位行列で置き換えたものを代用することが多い. また, スコア関数は対数尤度のパラメタに関する偏導関数であり, 尤度が非常に低いデータに対しては分母に現れる $p(\boldsymbol{x};\boldsymbol{\theta})$ が非常に小さくなり, 値が不安定になる可能性がある. そこで, 正規化 Fisher カーネル

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j;\boldsymbol{\theta}) = \frac{\boldsymbol{s}(\boldsymbol{x}_i;\boldsymbol{\theta})^T G^{-1}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{x}_j;\boldsymbol{\theta})}{||\boldsymbol{s}(\boldsymbol{x}_i;\boldsymbol{\theta})||_{G^{-1}} \cdot ||\boldsymbol{s}(\boldsymbol{x}_j;\boldsymbol{\theta})||_{G^{-1}}} \tag{3.14}$$

が提案されている [40]. 本論文第 5 章では, 特殊なデータに対する距離構造を理論的に妥当な方法で定義するために, その生成モデルを導出した上で, Fisher 情報行列を単位行列で置き換えた正規化 Fisher カーネルをデータ同士の類似度として利用する.

第4章　条件付きエントロピー最小化基準

教師付き学習による判別問題は, 判別に適した空間へのデータの写像 (変換) を学習することが目的であり, これは判別に適した距離構造を学習していることに他ならない.

本章では, 初めに低次元空間での距離構造を, データのクラス判別に適した形で学習する問題, すなわち次元削減問題を考える. そして, 教師付きの距離学習問題における目的関数として変換後のデータのクラス条件付きエントロピーを最小化することが理論的に妥当なものであることを示す. 次に, 条件付きエントロピー最小化基準に基づく線型次元削減手法を提案する. さらに, カーネル関数の最適化を, 条件付きエントロピー最小化基準を用いて行う方法を提案する. これは, カーネル関数に付随する特徴空間における距離構造を最適化していると理解できる.

4.1　距離構造の学習及び次元削減

データが有する本質的な情報を失わずにそのデータの次元を削減するという問題は, 情報処理における重要な課題の一つである. 学習データにクラスラベルが付随している教師付き次元削減手法としては, Fisher の判別分析 (Fisher Discriminant Analysis: FDA [5]) が広く用いられている. FDA は, 特徴データとそのクラスラベルが観測された状況で, クラス内のデータの分散を小さく保ちつつ, クラス間の分散が大きくなるような方向への特徴データの射影を求める手法である.

各クラスのデータが共分散構造の等しい正規分布に従っている時は, FDA は最適なクラス分離を与える方向を発見することができる. しかし, 多くの問題では正規性の仮定は成り立たず, 判別性の高い射影を得ることができないことがある. FDA の自然な拡張として, 同一のクラスに属するデータ同士の類似度を考慮した, Local Fisher Discriminant

Analysis(LFDA) が提案されている [29]. これはデータの局所性をデータ間の類似度行列という形で導入するものであり, 判別の前処理として用いた場合に FDA を大きく上回る判別精度が得られると報告されている.

本章では, 低次元空間での距離構造の学習問題として, 情報論的観点から次元削減問題にアプローチする. FDA における目的関数に着目し, 条件付きエントロピー最小化基準による距離構造の学習の枠組みを提案する. この枠組は, カーネル法に基づく特徴空間における距離構造の学習にも有効である. カーネル関数で定まる特徴空間の最適化問題である Multiple Kernel Learning(MKL) を提案する枠組みで捉え, 新たな MKL 手法を提案する.

4.1.1　次元削減の情報論的観点からの理解

次元削減問題とは, データ集合 $D = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^n$, が与えられたときに, これを $m \leq n$ なる m 次元ベクトルに写す写像 $f : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{x}_i \in \mathbb{R}^n$ を求める問題である. 線型の次元削減の場合には, 変換 f は行列 $A \in \mathbb{R}^{n\times m}$ を用いて

$$\boldsymbol{z}_i = A^T \boldsymbol{x}_i, \quad A \in \mathbb{R}^{n\times m} \tag{4.1}$$

で定義される. ここでは, 情報理論 [18] の観点から次元削減問題を考察する.

教師付きの次元削減手法としては, Fisher の判別分析 (Fisher Discriminant Analysis; FDA) が代表的である. データセット $D = \{\boldsymbol{x}_i\}_{i=1}^N$ とそのクラスラベル $\{y_i\}_{i=1}^N$, $y_i \in \{1, 2, \ldots, C\}$ が与えられたとする. D_y をクラスラベルが y であるようなデータ集合とし, そのデータの個数を $N_y = |D_y|$ で表す. 集合 D_y に属するデータの平均ベクトルと共分散行列をそれぞれ $\boldsymbol{\mu}_y$ 及び Σ_y として, 全データの平均ベクトルと共分散行列をそれぞれ $\boldsymbol{\mu}$, Σ とする. さらに, クラス内共分散行列 Σ_w とクラス間共分散行列 Σ_b をそれぞれ

$$\begin{aligned}
\Sigma_w &= \frac{1}{N}\sum_{y=1}^{C}\sum_{\boldsymbol{x}\in D_y}(\boldsymbol{x}-\boldsymbol{\mu}_y)(\boldsymbol{x}-\boldsymbol{\mu}_y)^T = \sum_{y=1}^{C}\frac{N_y}{N}\Sigma_y, \\
\Sigma_b &= \frac{1}{N}\sum_{y=1}^{C} N_y(\boldsymbol{\mu}_y-\boldsymbol{\mu})(\boldsymbol{\mu}_y-\boldsymbol{\mu})^T
\end{aligned}$$

で定義する. ここで, 文脈に応じて記号 D でデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^N$ と, データの添字集合 $\{1, 2, \ldots, N\}$ の両方を表すことに注意する.

FDA は, 変換された共分散行列の比を最小化することで最適な変換行列 A を求める. つまり, $|A^T\Sigma_w A|/|A^T\Sigma_b A|$ を最小化する. ここで, $|M|$ は正方行列 M の行列式である. 分子と分母の両方に非ゼロの定数をかけても目的関数の値は変わらないので, FDA は以下の最小化問題として定式化される:

$$\min_A \ |A^T\Sigma_w A| \quad \text{subject to} \quad |A^T\Sigma_b A| = const. \tag{4.2}$$

次に, FDA の最適化問題 (4.2) の情報論的な解釈を与える. 次の命題は, 条件付きエントロピーと FDA 基準 (4.2) との関係を表すものである.

命題 4.1
変換された確率変数 $A^T\boldsymbol{X}$ のクラス条件付きエントロピー

$$H(A^T\boldsymbol{X}|Y) = \sum_{y=1}^{C} \frac{N_y}{N} H(A^T\boldsymbol{X}|Y=y)$$

を考える. 記号 $H_G(\boldsymbol{X})$ で, $\boldsymbol{X}$ と同じ共分散構造を持つ正規分布のエントロピーを表す. この時, 次の不等式が成立する:

$$\begin{aligned} H(A^T\boldsymbol{X}|Y) &\leq H_G(A^T\boldsymbol{X}|Y) && (4.3)\\ &= \log(2\pi)^{m/2}e + \frac{1}{2}\sum_{y=1}^{C}\frac{N_y}{N}\log|A^T\Sigma_y A| && (4.4)\\ &\leq \log(2\pi)^{m/2}e + \frac{1}{2}\log|A^T\Sigma_w A|. && (4.5) \end{aligned}$$

ここで, e は自然対数の底である.

証明
初めの不等式 *(4.3)* は, 非有界サポートを持つ分布で共分散行列を固定したとき, 正規分布が最もエントロピーが高い分布であることから導かれる *[18]*. 二つ目の不等式 *(4.5)* は, Σ_w の定義と *Jensen* の不等式から得られる. □

この命題より, FDA は条件付きエントロピーの上界の最小化問題を解いていることが分かる. 次節では, 条件付きエントロピー最小化に基づく教師付きの距離構造学習の枠組みを提案する.

4.1.2　条件付きエントロピー最小化基準による次元削減

教師付き次元削減においては, 変換されたデータの低次元空間での表現は, 各クラスにおいてコンパクトに纏まっていることが望ましい. 直観的には, 確率変数の実現値であるデータを観測したとき, それらが狭い領域に集中して分布していればそのエントロピーは小さいといえる. FDA はクラス条件付きエントロピーの上界を最小化しているという事実に基づき, より直接的にクラス条件付きエントロピー $H(\boldsymbol{X}|Y)$ を最小化するような変換 $f : \boldsymbol{x} \mapsto \boldsymbol{z}$ を学習するような教師付きの次元削減の枠組みを提案する. ここで, 条件付きエントロピー $H(\boldsymbol{Z}|Y)$ は, 全てのデータ $\boldsymbol{x}$ を一点に写すような変換によって最小化されることに注意する. また, 変換 f の表現力が高すぎると, 与えられたデータに対する過適合が生じる可能性が高い. こうした自明な解や過適合を防ぐために, $H(\boldsymbol{Z}|Y)$ の最小化において何らかの正則化が必要である. 本論文では, 正則化の度合いをコントロールするパラメタ $\varepsilon > 0$ を導入し, 一般には変換 f とデータ D に依存するような非負正則化汎関数 $\Psi(f, D)$ により正則化を行う. つまり, 一般に正則化項を $\varepsilon\Psi(f, D)$ として, 最小化問題

$$\min_{f:\boldsymbol{x}\mapsto\boldsymbol{z}} H(\boldsymbol{Z}|Y) + \varepsilon\Psi(f, D) \tag{4.6}$$

による教師付き次元削減の枠組みを提案する. 正則化汎関数 $\Psi(f, D)$ は問題に応じて適切に定める必要がある. 例えば, 線型変換 (4.1) を考え, FDA のように変換後のデータのクラス間共分散行列の行列式が定数 (例えば 1) になるように制約を与える場合には, $\Psi(f, D) = \Psi(A, D) = (|A^T\Sigma_b A| - 1)^2$ とすれば良い. なお, 条件付きエントロピー最小化基準の理論的背景として次元削減手法である FDA の目的関数との関係を用いたが, 提案する条件付きエントロピー最小化基準は次元削減に限らず, 一般の距離構造学習の枠組みである.

エントロピーの推定手法

エントロピーの最小化問題 (4.6) を解くためには, エントロピーの推定が必要である. ここでは第 2 章で紹介した MNN 法 [24] を用いてエントロピーを推定するものとして, 以下では $H(\boldsymbol{X})$ と書いた場合には式 (2.12) の $H_{\mathrm{MNN}}(\boldsymbol{X})$ を表すものとする.

多次元分布の同時エントロピーを精度良く推定することは, MNN 法を用いても一般には困難である. そこで, 命題 2.1 から同時エントロピーが

周辺エントロピーの和によって上から抑えられることを用いて, 同時エントロピーの代わりに周辺エントロピーの和を用いる. 変換行列 A の l 番目の行ベクトルを $\boldsymbol{a}_l$ として, 変換後のベクトル $\boldsymbol{z}$ の l 番目の成分 $z_l = \boldsymbol{a}_l^T \boldsymbol{x}$ の周辺エントロピーは

$$H(\boldsymbol{a}_l^T \boldsymbol{X}) = H(Z_l) = -\int p(z_l) \log p(z_l) dz_l, \quad l = 1, \ldots, m$$

で定義される. この周辺エントロピーの和

$$\overline{H(\boldsymbol{Z})} = \sum_{l=1}^{m} H(Z_l) \geq H(\boldsymbol{Z})$$

は, $\boldsymbol{z} = (z_1, \ldots, z_m)$ の同時エントロピーの上界を与える. 同様に, クラス $Y = y$ のデータの条件付きエントロピーの上界も

$$\overline{H(\boldsymbol{Z}|Y=y)} = \sum_{l=1}^{m} H(Z_l|Y=y) \geq H(\boldsymbol{Z}|Y=y)$$

で計算できるので, $\overline{H(\boldsymbol{Z}|Y=y)}$ をクラス事前確率で重みをつけて加え合わせたものがクラス条件付きエントロピーの上界を与える:

$$\begin{aligned} \overline{H(\boldsymbol{Z}|Y)} &= \sum_{y=1}^{C} p(y) \overline{H(\boldsymbol{Z}|Y=y)} \\ &= \sum_{y=1}^{C} p(y) \sum_{l=1}^{m} H(Z_l|Y=y) \\ &\approx \sum_{y=1}^{C} \frac{N_y}{N} \sum_{l=1}^{m} H(Z_l|Y=y). \end{aligned}$$

ここで, クラス事前確率 $p(y)$ は N_y/N で推定した. 以下では, 多次元確率変数のエントロピーを扱うときは, 上述のようにその周辺エントロピーの和で定義される上界を扱うものとする.

勾配法に基づく最適化アルゴリズム

条件付きエントロピー最小化基準における目的関数 (4.6) を勾配法により最小化するアルゴリズムを具体的に与える.

線型次元削減を考えたとき, 最適化の対象は

$$\min_{A\in\mathbb{R}^{n\times m}} \overline{H(A^T\boldsymbol{X}|Y)} + \varepsilon\Psi(A, D), \tag{4.7}$$

である. ここで, 条件付きエントロピーは

$$\overline{H(A^T\boldsymbol{X}|Y)} = \sum_{y=1}^{C} \frac{N_y}{N} \overline{H(A^T\boldsymbol{X}|Y=y)} = \sum_{y=1}^{C} \frac{N_y}{N} \sum_{l=1}^{m} H(\boldsymbol{a}_l^T\boldsymbol{X}|Y=y)$$

で計算される. 周辺エントロピー $H(\boldsymbol{a}_l^T\boldsymbol{X}|Y=y)$ の MNN 法による推定量は

$$H(\boldsymbol{a}_l^T\boldsymbol{X}|Y=y) = \frac{1}{N(N-1)} \sum_{\substack{i,j\in D_y,\\ i\neq j}} \log||\boldsymbol{a}_l^T\boldsymbol{x}_i - \boldsymbol{a}_l^T\boldsymbol{x}_j|| + const.$$

で与えられるので, 変換行列の第 l 行 $\boldsymbol{a}_l$ に関する導関数は

$$\frac{\partial H(\boldsymbol{a}_l^T\boldsymbol{X}|Y)}{\partial \boldsymbol{a}_l^T} = \frac{2}{N(N-1)} \sum_{y=1}^{C} \sum_{\substack{i,j\in D_y,\\ i\neq j}} \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)}{(\boldsymbol{a}_l^T(\boldsymbol{x}_j - \boldsymbol{x}_i))^2},$$

である. これを用いて, $\overline{H(A^T\boldsymbol{X}|Y)} = \sum_{l=1}^{m} H(\boldsymbol{a}_l^T\boldsymbol{X}|Y)$ を勾配法により最小化することができる.

準直交化

高次元データの同時エントロピーを精度良く推定することが困難なことから, 本論文では周辺エントロピーの和により同時エントロピーを近似し, 最小化するというアプローチを取る. 変換行列 $A \in \mathbb{R}^{n\times m}$ は, 各列が n 次元から 1 次元空間への射影となっている. こうして周辺エントロピーを最小化するとき, 単純な最適化を行うと $\boldsymbol{a}_i$ が全ての i について同一になってしまう可能性がある. 勾配法の各ステップにおいて準直交化処理をおこない, 変換行列 A の各列の無相関化によりこの問題を解決する. これは, A^T の取りうる範囲を m 次元直交基底だけに限定することに対応する. こうした A^T の空間は Stiefel 多様体と呼ばれ, 独立成分分析などの最適化でよく用いられる [51]. 記述の簡単のためと, アルゴリズムの収束を早めるために, データを予め白色化しておく. 確率変数 $\boldsymbol{X}$ が白色であるとは, その共分散行列が単位行列であることをいう. 確率変数 $\boldsymbol{X}$ の共分

散行列の固有値分解を $E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T]=U\Lambda U^T$ とすると, 白色化処理は $\Lambda^{-\frac{1}{2}}U^T\boldsymbol{x}$ なる変数変換によって実現される.

準直交化とは, 所与のデータ $D=\{\boldsymbol{x}_i\}_{i=1}^N$ が白色化されているという仮定の下で, 周辺エントロピーの最小化の各ステップで得られる行列 $A\in\mathbb{R}^{n\times m}$ が近似的に $||A^TA-I_m||_F$ を満足するように修正を行うことである. ここで I_m は $m\times m$ の単位行列であり, $||\cdot||_F$ は Frobenius ノルムである. 準直交化は, 式 (4.7) における正則化項を $\Psi(A,D)=||A^TA-I_m||_F$ と設定したことと等価である.

補題 4.2
変換行列 A の準直交化は, 次の処理を収束するまで行うことで実現される *[20]:*

ステップ 1 行列 A を, A^TA の最大固有値の平方根で割る.

ステップ 2 $A\leftarrow\frac{3}{2}A-\frac{1}{2}AA^TA$.

ステップ 3 A の各列のノルムを 1 に正規化する.

証明
対称行列 A^TA の固有値分解を $A^TA=EDE^T$ とする. ここで $E\in\mathbb{R}^{m\times m}$ は直交行列であり, D は A^TA の固有値 $\{d_i\}_{i=1}^m$ を対角成分とする対角行列である. 上記の手続きのステップ *2* により, A^TA は

$$\begin{aligned} A^TA &\mapsto \frac{1}{4}(3A-AA^TA)^T(3A-AA^TA) \\ &= \frac{1}{4}E\left(9D-6D^2+D^3\right)E^T. \end{aligned}$$

のように変換される. ここで, A^TA の最大固有値がステップ *1* によって 1 に正規化されていることから, $d_i\in(0,1]$ である. この変換により A^TA の固有値は

$$h(d_i)=\frac{1}{4}(9d_i-6d_i^2+d_i^3),\quad i=1,\ldots,m$$

となる. ここで, $h(d_i)-d_i=\frac{d_i}{4}\{(d_i-3)^2-4\}\geq 0$ なので, この 3 ステップの繰り返しにより A^TA の固有値は 1 に収束する. □

以上より, 線型変換 $A^T:\boldsymbol{x}\mapsto\boldsymbol{z}$ に関するクラス条件付きエントロピー最小化アルゴリズムが得られる. アルゴリズムは図 4.1 にまとめた. このアルゴリズムを, **LCEM**(Linear dimensionality reduction algorithm based on Conditional Entropy Minimization) アルゴリズムと呼ぶことにする.

LCEM : Linear dimensionality reduction algorithm based on conditional entropy minimization.

入力: 学習データ $D=\{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^n$, クラスラベルデータ $\{y_i\}_{i=1}^N$, $y_i \in \{1,2,\ldots,C\}$. 変換後のデータの次元 $m\,(\leq n)$. 勾配法のステップ幅 $\xi > 0$.

初期化: 初期変換行列 $A \in \mathbb{R}^{n\times m}$ を, $\mathrm{rank}A=m$ となるように選ぶ. 学習データ $D=\{\boldsymbol{x}_i\}_{i=1}^N$ を, 経験共分散行列を用いて白色化する.

繰り返し: 収束するまで以下を繰り返す:

勾配ステップ: 変換行列の各行を更新する:

$$\boldsymbol{a}_l^T := \boldsymbol{a}_l^T - \xi \frac{\partial H(\boldsymbol{a}_l^T \boldsymbol{X}|Y)}{\partial \boldsymbol{a}_l^T}, \qquad l=1,\ldots,m.$$

準直交化ステップ: 以下の準直交化処理を収束するまで繰り返す:

1. 変換行列 A を, 行列 A^TA の最大固有値で割る.
2. $A := \frac{3}{2}A - \frac{1}{2}AA^TA$,
3. $\boldsymbol{a}_l := \boldsymbol{a}_l / ||\boldsymbol{a}_l||, \quad l=1,\ldots,m.$

出力: 収束した変換行列 A.

図 4.1: LCEM アルゴリズム. 各勾配ステップにおいて, 周辺化エントロピーは勾配法によって最小化される.

なお, 第3章で紹介した教師付き距離構造学習手法である MCML の目的関数と, LCEM アルゴリズムの目的関数は類似している. 実際, MCML

では, 以下のようにKullback-Leibler ダイバージェンスの和を最小化する:

$$\begin{aligned} & \arg\min_A \sum_{i=1}^{N} KL(p_0(\boldsymbol{x}_j|\boldsymbol{x}_i), p_A(\boldsymbol{x}_j|\boldsymbol{x}_i)) \\ = \; & \arg\min_A \left\{ -\sum_{i=1}^{N}\sum_{j=1}^{N} p_0(\boldsymbol{x}_j|\boldsymbol{x}_i) \log p_A(\boldsymbol{x}_j|\boldsymbol{x}_i) \right\} \\ = \; & \arg\max_A \sum_{i=1}^{N}\sum_{j\in C_i} \log p_A(\boldsymbol{x}_j|\boldsymbol{x}_i). \end{aligned}$$

一方 LCEM アルゴリズムにおいては, クラス条件付きエントロピーを最小化する:

$$\begin{aligned} & \arg\min_A H(A^T\boldsymbol{X}|Y) \\ = \; & \arg\min_A \left\{ -\sum_{y=1}^{C} p(y) \int p(A^T\boldsymbol{x}|Y=y) \log p(A^T\boldsymbol{x}|Y=y) d\boldsymbol{x} \right\} \\ \approx \; & \arg\max_A \sum_{y=1}^{C}\sum_{j\in C_y} \log p(A^T\boldsymbol{x}_i|Y=y). \end{aligned}$$

これらは類似しているように見えるが, MCML ではデータ $\boldsymbol{x}_i$ が他のデータ $\boldsymbol{x}_j$ をその近傍であるとみなす確率 $p_A(\boldsymbol{x}_j|\boldsymbol{x}_i)$ を用いており, これはデータそのものの分布とは異なる. さらに, 目的関数を単純な凸関数の形にするために, MCML では確率 $p_A(\boldsymbol{x}_j|\boldsymbol{x}_i)$ を Boltzmann 分布の形に仮定している.

4.2 特徴空間における距離の学習 –カーネル最適化 –

条件付きエントロピー基準により, カーネルにより誘導される特徴空間の距離構造を学習する方法を提案する.

条件付きエントロピー最小化による距離構造学習の枠組みは一般的な教師付き学習の枠組みであり, 非線型な距離構造学習への拡張も容易に行える. ここでは, カーネル Fisher 判別分析 (KFDA; [45]) に基づき, 提案する枠組みの非線型化と Multiple Kernel Learning (MKL) への拡張を行う.

4.2.1 カーネル Fisher 判別分析

Fisher の判別分析は, カーネル法を用いた非線型の判別分析に拡張されている. このカーネル Fisher 判別分析 (kernel Fisher Discriminant Analysis;KFDA [45]) は線型判別が不可能な幾つかのデータに対してうまく働くという結果が得られている. KFDA は多次元空間における距離構造学習の場合も考えることができるが, ここでは簡単のために 1 次元空間への次元削減のみを考える[1]. まず, $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ を $\mathbb{R}^n$ から $\mathbb{R}$ への線型変換とする. データ $\boldsymbol{x}$ はこの関数の値を用いて判別されるので, $f(\boldsymbol{x})$ を**判別関数**と呼ぶ. データ $\boldsymbol{x} \in \mathbb{R}^n$ が写像 $\phi : \mathbb{R}^n \to \mathbb{R}^{n'}$ によって n' 次元の特徴空間 $\mathbb{R}^{n'}$ に写像されるとする. このとき, 判別関数は n' 次元特徴空間から $\mathbb{R}$ への写像 $f(\boldsymbol{x}) = \boldsymbol{a}^T\phi(\boldsymbol{x})$ である. ここで, $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ における写像ベクトル $\boldsymbol{a}$ は n 次元ベクトルであるが, $f(\boldsymbol{x}) = \boldsymbol{a}^T\phi(\boldsymbol{x})$ における $\boldsymbol{a}$ は n' 次元ベクトルであることに注意する. ここで, 判別のコスト関数に $||\boldsymbol{a}||^2$ の形の正則化項を加えるとリプレゼンター定理が成立し ([40]), 実パラメタ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ を用いて $\boldsymbol{a} = \sum_{i=1}^N \alpha_i\phi(\boldsymbol{x}_i)$ とかけることを利用する. このとき, 特徴空間における内積は, カーネル関数を用いて $\langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ と表すことができ, 判別関数はカーネル関数を用いて

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) \tag{4.8}$$

となる.

与えられたデータ $D = \{\boldsymbol{x}_i\}_{i=1}^N$ のグラム行列を $K \in \mathbb{R}^{N\times N}$, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ として, $\boldsymbol{k}_i$ でその第 i 列を表すものとする. このグラム行列を用いて, 各クラスに属するデータのサンプル平均は

$$\bar{\boldsymbol{k}}^y = \frac{1}{N_y}\sum_{i\in D_y} \boldsymbol{k}_i,$$

全てのデータのサンプル平均は

$$\bar{\boldsymbol{k}} = \frac{1}{N}\sum_{i\in D} \boldsymbol{k}_i$$

[1]KFDA における判別関数はカーネル行列で表現された特徴空間におけるクラス内, クラス間共分散行列に関する一般化固有値問題の解として得られる後述の $\boldsymbol{\alpha}$ によるデータの射影と考えられる. 従って, 原理的にはクラス数 $C-1$ 次元空間に値を取る判別関数の構成が容易に行える

で計算できる. また, 特徴空間におけるクラス内共分散行列 V_w と, クラス間共分散行列 V_b はそれぞれ

$$V_w = \frac{1}{N}\sum_{y=1}^{C}\sum_{i\in D_y}(\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)(\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)^T,$$
$$V_b = \frac{1}{N}\sum_{y=1}^{C} N_y(\bar{\boldsymbol{k}}^y - \bar{\boldsymbol{k}})(\bar{\boldsymbol{k}}^y - \bar{\boldsymbol{k}})^T$$

である. KFDA の最小化目的関数は $\boldsymbol{\alpha}^T V_w \boldsymbol{\alpha}/\boldsymbol{\alpha}^T V_b \boldsymbol{\alpha}$ であり, 線型の FDA と同様に, $\boldsymbol{\alpha}^T V_b \boldsymbol{\alpha}$ が定数という制約条件の下での $\boldsymbol{\alpha}^T V_w \boldsymbol{\alpha}$ の最小化問題に帰着される. カーネル法を用いた場合, $\boldsymbol{\alpha}^T V_w \boldsymbol{\alpha}/\boldsymbol{\alpha}^T V_b \boldsymbol{\alpha}$ の最小化によりデータへの過適合が生じることが多い. ここでは, 正則化パラメタ $\zeta > 0$ を導入し, カーネル行列 K をクラス内共分散行列に加えたものでもとの共分散行列を置き換えることで過適合を防ぐ. つまり, KFDA は次のように定式化できる:

$$\min_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}^T(V_w + \zeta K)\boldsymbol{\alpha} \quad \text{subject to} \quad \boldsymbol{\alpha}^T V_b \boldsymbol{\alpha} = const. \tag{4.9}$$

本論文では, 記述の簡単のために 1 次元の判別関数のみを考えるが, 一般には判別関数を多次元にすることで判別精度の向上が期待できる. 例えば l 次元の判別関数を考えるときには, 判別軸への射影 $\boldsymbol{\alpha}$ を l 個考えて $A = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l) \in \mathbb{R}^{N\times l}$ なる行列を定義して, 最適化問題

$$\min_{A} \ |A^T(V_w + \zeta K)A| \quad \text{subject to} \quad |A^T V_b A| = const. \tag{4.10}$$

を解けば良い. ここで, $|\cdot|$ は行列の行列式を表す.

4.2.2 条件付きエントロピー基準に基づく Multiple Kernel Learning

本章では, 条件付きエントロピー最小化に基づく新しい MKL 手法を提案する.

第 3 章で述べたように, 一般にカーネル法では与えられたデータと課題に応じて適切なカーネル関数と適切なカーネルパラメタを選択しなければ十分な性能が得られず, このモデル選択の問題はカーネル法における重要な問題として認識されている. カーネルの最適選択問題への一つの解決法

として, 予め準備した複数のカーネル関数を, 与えられたデータに応じて適応的に組み合わせて用いる手法である Multiple Kernel Learning(MKL) がある.

本論文で提案する MKL 手法は, 以下の最適化問題によりカーネル関数の結合係数 $\boldsymbol{\beta}$ を求めるというものである:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & H(f(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta})|Y) \\ \text{subject to} \quad & H(f(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta})) = const., \\ & \sum_{s=1}^{S} \beta_s = 1,\ \beta_s \geq 0,\ s = 1,\ldots,S. \end{aligned} \tag{4.11}$$

ここで, 判別関数の $\boldsymbol{\alpha},\boldsymbol{\beta}$ への依存性を明示的に示すために, $f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x},\boldsymbol{x}_i;\boldsymbol{\beta},\boldsymbol{\lambda})$ とした. 形式的にはこの最適化問題は, 式 (4.6) で示した枠組みにおける正則化汎関数として

$$\begin{aligned} \Psi(f,D) &= \Psi(\boldsymbol{\alpha},\boldsymbol{\beta},D) \\ &= (H(f(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta})) - 1)^2 \\ &\quad + \left(\sum_{s=1}^{S} \beta_s - 1\right)^2 + \left(\sum_{s=1}^{S} (\beta_s - |\beta_s|)\right)^2 \end{aligned}$$

としたものと理解できる.

最適化問題 (4.11) を $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ 両方について同時に解くのは困難である. そこで, $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ に関する繰り返し最適化手法を用いる.

繰り返しアルゴリズムにおいて, t 回の繰り返し後に得られる $\boldsymbol{\alpha},\boldsymbol{\beta}$ の値をそれぞれ $\boldsymbol{\alpha}(t),\boldsymbol{\beta}(t)$ とする. まずは, $\boldsymbol{\beta}$ を固定したうえで, $\boldsymbol{\alpha}$ に関する最適化を考える. クラス内, クラス間共分散行列を, $\boldsymbol{\beta}$ への依存性を陽に表すために $V_w(\boldsymbol{\beta}), V_b(\boldsymbol{\beta})$ とし, 式 (4.9) における正則化項 ζK は記述の簡単化のために省略する. 不等式 (4.3), (4.5) で示したように, KFDA はクラス条件付きエントロピーの上界を最小化する. つまり, KFDA の目的関数とクラス条件付きエントロピーの関係は, 命題 4.1 と同様に

$$\begin{aligned} H(f(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta}(t-1))|Y) &\leq H_G(f(\boldsymbol{X};\boldsymbol{\alpha},\boldsymbol{\beta}(t-1))|Y) \qquad (4.12) \\ &= \log(2\pi)^{1/2} e + \frac{1}{2}\sum_{y=1}^{C} \frac{N_y}{N} \log \boldsymbol{\alpha}^T V_y(\boldsymbol{\beta}(t-1))\boldsymbol{\alpha} \\ &\leq \log(2\pi)^{1/2} e + \frac{1}{2} \log \boldsymbol{\alpha}^T V_w(\boldsymbol{\beta}(t-1))\boldsymbol{\alpha} \qquad (4.13) \end{aligned}$$

で表される. ここで, $V_y = \frac{1}{N_y}\sum_{i\in D_y}(\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)(\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)^T$ である. 式 (4.13) 右辺はクラス条件付きエントロピーの上界になっており, $\boldsymbol{\beta}$ を固定したとき $\boldsymbol{\alpha}$ に関する最小値は KFDA によって求めることが出来る.

次に, 前ステップで求めた $\boldsymbol{\alpha}$ を用いて, 条件付きエントロピーをカーネル結合係数 $\boldsymbol{\beta}$ に関して最小化する. 正則化項 $H(f(\boldsymbol{\alpha}, \boldsymbol{\beta}, D)) = const.$ は $\boldsymbol{\beta}$ を含むので, このエントロピー項を条件付きエントロピー項とまとめて最適化をする. パラメタ $\eta > 0$ を導入し, 新たに最小化の目的関数を次式で定義する:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & H(f(\boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\beta})|Y) - \eta H(f(\boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\beta})) \\ \text{subject to} \quad & \sum_{s=1}^{S} \beta_s = 1, \quad \beta_s \geq 0. \end{aligned} \tag{4.14}$$

この $\boldsymbol{\beta}$ の最適化の結果, 更新された係数 $\boldsymbol{\beta}$ を用いて新しいカーネル関数を得る. この新しいカーネルを用いて, 共分散行列 $V_w(\boldsymbol{\beta}), V_b(\boldsymbol{\beta})$ を計算し, 再び条件付きエントロピーを $\boldsymbol{\alpha}$ に関して KFDA で最小化する. この 2 ステップの最適化を, $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ が収束するか, 条件付きエントロピーの値が収束するまで繰り返す. このアルゴリズムを, **MCEM**(Multiple kernel learning algorithm based on Conditional Entropy Minimization) と呼び, 図 4.2 にまとめる.

MCEM アルゴリズムにおける $\boldsymbol{\beta}$ に関する最適化の方法は任意である. 本論文では 3 種類の最適化手法を考案した. 一つは, ランダムサーチに基づく方法であり, 後の二つは条件付きエントロピーを $\boldsymbol{\beta}$ に関する二次形式で近似した上で最小化するものである. ランダムサーチに基づく方法では, 前回の最適化結果として得られている $\boldsymbol{\beta}(t-1)$ を平均として, 単位行列を共分散行列とする正規分布から P 個のサンプル $\{\boldsymbol{\beta}_p\}_{p=1}^{P}$ を取り出し, それらを用いて条件付きエントロピー $H(f(\boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\beta}_p)|Y)$ を計算して最も小さい値を与えるサンプルを最適化結果として採用するというものである. この手法は非常に単純であるが, 実験の結果十分よい結果を与えることが確認できている. また, この手法はカーネルを凸結合以外の方法で組み合わせる場合にも適用可能な汎用的な方法である.

後の 2 つのアルゴリズムの詳細は付録 3 に詳しく記す. これら 2 つのアルゴリズムは目的関数を $\boldsymbol{\beta}$ の二次形式で近似するものであり, 片方は二次計画問題として, もう片方はさらに制約条件を緩和して固有値問題として $\boldsymbol{\beta}$ に関する最適化問題を定式化するものである. $\boldsymbol{\beta}$ の最適化手法に応じて, ランダムサーチに基づくアルゴリズムを **MCEM.R**, 二次計画問題

に基づくアルゴリズムを **MCEM.Q**, そして固有値問題に基づくアルゴリズムを **MCEM.E** と呼ぶ.

MCEM : Multiple kernel learning algorithm based on conditional entropy minimization.

入力: 学習データ $D = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^n$ とそのクラスラベルデータ $\{y_i\}_{i=1}^N$, $y_i \in \{1, 2, \ldots, C\}$. S 個の要素カーネル $\{k(\,\cdot\,,\,\cdot\,;\lambda_s)\}_{s=1}^S$ のカーネルパラメタ $\boldsymbol{\lambda}=\{\lambda_s\}_{s=1}^S$. KFDA のための正則化パラメタ $\zeta > 0$.

初期化: カーネル結合係数を初期化: $\boldsymbol{\beta}(0)=\{\beta_s(0)\}_{s=1}^S$.

繰り返し: 収束するまで以下を繰り返す:

$\boldsymbol{\alpha}$ の最適化ステップ: KFDA の最小化問題を, $\boldsymbol{\beta}(t-1)$ を固定して解き, $\boldsymbol{\alpha}(t)$ を得る:

$$\begin{array}{ll} \min_{\boldsymbol{\alpha}} & |\boldsymbol{\alpha}^T(V_w(\boldsymbol{\beta}(t-1)) + \zeta K)\boldsymbol{\alpha}| \\ \text{subject to} & |\boldsymbol{\alpha}^T V_b \boldsymbol{\alpha}| = const. \end{array}$$

$\boldsymbol{\beta}$ の最適化ステップ: 判別関数 $f(\boldsymbol{X}; \boldsymbol{\alpha}(t), \boldsymbol{\beta})$ の条件付きエントロピーを, $\boldsymbol{\alpha}(t)$ を固定した上で最適化して, $\boldsymbol{\beta}(t)$ を得る:

$$\begin{array}{ll} \min_{\boldsymbol{\beta}} & H(f(\boldsymbol{X}; \boldsymbol{\alpha}(t), \boldsymbol{\beta})|Y) \\ \text{subject to} & \displaystyle\sum_{s=1}^S \beta_s = 1, \;\; \beta_s \geq 0, \;\; s = 1, \ldots S. \end{array}$$

出力: 収束したパラメタ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$. これらのパラメタを用いて計算した判別関数 $f(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^N \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\lambda})$.

図 4.2: 一次元判別軸上の関数値の条件付きエントロピーが最小になるように, 判別関数のパラメタ $\boldsymbol{\alpha}$ とカーネル結合係数 $\boldsymbol{\beta}$ を繰り返し最適化する.

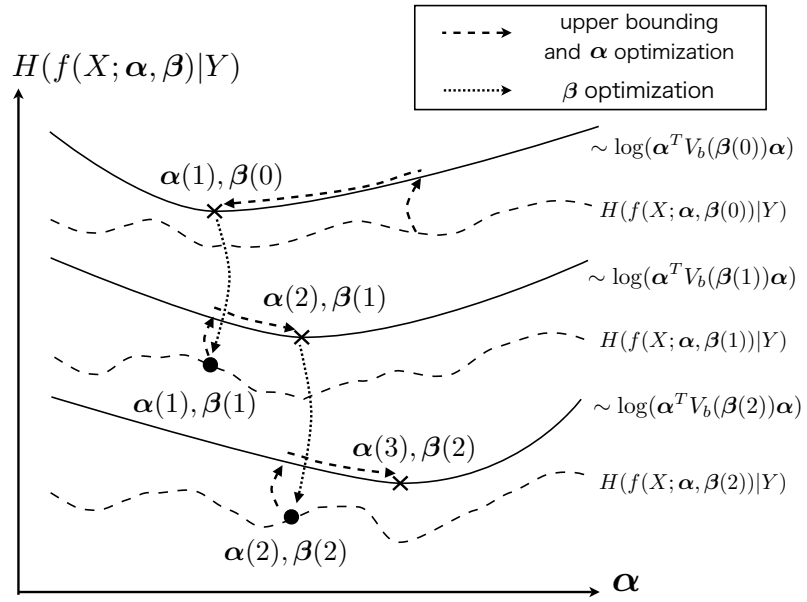

図 4.3: MCEM アルゴリズムの挙動の概念図．点線は, 条件付きエントロピーの等高線．実線は, 条件付きエントロピーの上界の等高線であり, KFDA の目的関数の値である．MCEM アルゴリズムは, $\boldsymbol{\beta}$ を固定した上で上界による近似と KFDA による最小化で $\boldsymbol{\alpha}$ に関する最適化を行う．次に, $\boldsymbol{\alpha}$ を固定した上で $\boldsymbol{\beta}$ に関する最適化を行う．

4.3 実験による評価

本節では, 本章で提案した条件付きエントロピー最小化基準に基づく線型次元削減手法 LCEM と, MKL 手法 MCEM の性能を, ベンチマークデータを用いた実験により評価する．

4.3.1 LCEM アルゴリズムの性能評価実験

提案した LCEM アルゴリズムの性能を, 実データを用いて評価する．ここでは次元削減手法を, 2 クラス判別問題の前処理として用いる．判別手法としては, ここでは実装が容易であることと, データ分布の影響が顕著に現れることから, 最近傍法 [52] を用いた．

評価に用いるデータとして, IDA データセットを用いた[2]. これは, 機械学習の分野で頻繁に用いられる 2 クラスのデータセットであり, 元々は文献 [53] で用いられた. このデータセットには 13 種類の 2 クラス判別問題が含まれており, それぞれのデータの次元やデータ数は表 4.1 に示した通りである. ここで, realization とは学習データとテストデータの対の数である.

表 4.1: IDA データの内容.

データ名	データ次元	学習データ数	テストデータ数	realization 数
banana	2	400	4900	100
breast-cancer	9	200	77	100
diabetes	8	468	300	100
flare-solar	9	666	400	100
german	20	700	300	100
heart	13	170	100	100
image	18	1300	1010	20
ringnorm	20	400	7000	100
splice	60	1000	2175	20
thyroid	5	140	75	100
titanic	3	150	2051	100
twonorm	20	400	7000	100
waveform	21	1000	1000	100

オリジナルのデータを, 主成分分析 (PCA), FDA, MCML, LFDA, LCEM それぞれを用いて低次元空間に写像する. 判別のための適切な次元は, 文献 [53] と同様に, 初めの 5-readlization 分の学習データを用いて 5-fold のクロスバリデーションを行い, 5 通りの削減次元の中央値として定めた. 推定された最適な次元 Dim は, 判別誤差の横に [Dim] のように示した. LCEM アルゴリズムの停止条件は, 第 t 回目の繰り返しの後の条件付きエントロピーを H_t として, $|H_t - H_{t-1}|/|H_{t-1}| < 10^{-4}$ とした. 表 4.2 に, 各次元削減手法に対応する判別結果の平均と標準偏差を示す. 値はパーセント表記である. 各データについて, 最良の判別結果と, その最良の結果と比較

[2]以下のページから取得可能
http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

して t 検定を行い 5 パーセントの有意水準で差がないとされた結果は太字で記してある.

表 4.2: 誤判別率の平均と標準偏差 (パーセント表記). 最も精度の良い結果と, 5%の有意水準の t テストによって同等と検定された結果は太字で表示してある.

Data name	PCA	FDA	MCML	LFDA	LCEM	Euclidean
banana	14.0(0.8)[2]	38.3(4.0)	39.6(1.3)[1]	**13.7**(0.8)[2]	**13.6**(0.8)[2]	13.6(0.8)
breast-cancer	40.7(7.1)[3]	34.9(5.1)	34.5(4.4)[4]	**33.3**(4.6)[6]	**33.6**(4.4)[4]	32.7(4.8)
diabetes	38.4(5.0)[4]	31.3(2.8)	31.3(1.9)[7]	32.3(2.6)[3]	**30.1**(2.1)[3]	30.1(2.1)
flare-solar	48.6(6.9)[5]	**36.4**(1.9)	**36.6**(2.0)[5]	36.8(1.9)[2]	**36.5**(1.9)[3]	36.5(1.9)
german	41.8(4.5)[2]	32.0(2.6)	31.4(2.4)[17]	**30.2**(2.47)[11]	31.2(2.6)[9]	29.5(2.5)
heart	46.3(23.9)[4]	22.9(4.1)	24.5(3.4)[10]	**21.6**(4.3)[5]	22.7(4.0)[3]	23.2(3.7)
image	37.3(9.5)[2]	22.1(0.9)	4.1(0.6)[15]	3.7(1.0)[13]	**3.4**(1.0)[16]	3.4(0.5)
ringnorm	28.0(5.1)[10]	31.7(1.0)	23.5(1.1)[8]	20.4(1.0)[6]	**19.7**(0.8)[8]	35.0(1.4)
splice	43.9(4.9)[2]	20.4(0.8)	27.0(0.7)[43]	**16.4**(0.8)[5]	20.6(0.6)[2]	28.8(1.5)
thyroid	9.1(4.4)[2]	17.9(4.9)	4.9(2.1)[4]	**4.3**(2.3)[3]	**4.4**(2.2)[4]	4.4(2.2)
titanic	26.4(8.4)[1]	**22.5**(1.1)	**22.5**(1.1)[1]	**22.6**(1.5)[1]	**22.5**(1.1)[1]	22.5(1.1)
twonorm	7.6(18.8)[3]	**3.5**(0.5)	8.0(0.7)[19]	**3.5**(0.4)[6]	**3.6**(0.4)[2]	6.7(0.7)
waveform	31.7(18.7)[9]	18.6(1.2)	17.8(0.7)[17]	**11.7**(0.7)[2]	16.3(1.0)[17]	15.8(0.7)

表 4.2 から, 多くのデータセットに対して, PCA, FDA, MCML といった従来手法よりも LCEM アルゴリズムは優れた判別結果を示すことが分かる. また, LFDA とはほぼ同等の判別結果であるといえる. 表 4.2 の列”Euclidean” には, 次元削減をしない状態, つまり通常のユークリッド空間での最近傍法による判別結果を示した. ユークリッド空間における判別結果と比較して, LCEM アルゴリズムは多くのデータセットに対して判別精度をよく保存し, さらにいくつかのデータセットに対しては精度の向上が見られる. この結果から, 今回実験した線型次元削減手法には以下のような傾向があることが見て取れる. IDA データセットの中で,“banana”, “thyroid” 及び “waveform” はクラス内でデータが多峰性の分布に従い, 他のデータは各クラス内でデータは単峰性の分布に従うことがわかっている. FDA は, 文献 [29] で指摘されたように, 多峰性のデータには良い結果を示していない. Maximum Collapsing Metric Learning(MCML) は, 最近傍法による判別の前処理としての次元削減手法としてはあまり性能が良くないことが分かる. LFDA と LCEM は, 類似した判別性能を示している. 文献 [29] では, LFDA は多峰性のデータに適していると主張している. LCEM アルゴリズムは, 多峰性のデータである “banana” と “thyroid” に対してはよい判別結果を示しているが, もうひとつの多峰性データである “waveform” に対しては誤り率がやや高い. 現時点では, どの手法がどのデータに対して適しているかという一般的な結論を導くことは困難であり, これは明らかにすべき重要な課題の一つである.

データを 1 次元に削減したときの性能評価と, 削減する次元を変化させた時の精度のグラフなど, より詳しい実験結果は付録 2 に示す.

4.3.2 MCEM アルゴリズムの性能評価実験

線型次元削減手法 LCEM アルゴリズムの実験と同様に, 最近傍判別器を用いた判別実験を行う. また, 先行研究において利用された, タンパク質の機能推定問題にも提案手法を適用し, その有効性を示す.

表 4.3 に, KFDA [45], KFDA-MKL [44], SDP-MKL [9], R-MKL [43] 及び MCEM アルゴリズムを IDA データセットに適用した判別結果を示す[3]. ここで, KLFDA は LFDA のカーネル版である [29]. データを削減する次元 m は任意であるが, 簡単のためここでは全て 1 次元への写像のみを考

[3]SDP-MKL は大規模データに対しては現実的な時間内での実行が困難であるため, “banana”, “image”, “splice” については, SDP-MKL を実行する際にデータを 1/10 程度に削減した.

える. 組み合わせるカーネル関数としては, 20種類のガウスカーネルを用いる. そのパラメタとしては,

$$\boldsymbol{\lambda} = (10, 9, \ldots, 1, 0.75, 0.5, 0.25, 0.1, 0.075, 0.05, 0.025, 0.01, 0.005, 0.001).$$

を用いた. KFDAについては, カーネルパラメタを2種類の方法で定めた. 一つは, Jaakkolaのヒューリスティクスと呼ばれる方法である. これは, 片方のクラスのデータともう片方のクラスのデータとのユークリッド距離で最も小さい値の中央値をガウスカーネルのパラメタとして用いるというものである [54]. 表 4.3 ではKFDA(H)と記した. もう一つは, 先行研究 [53; 45] で行われたように, 初めの5つのrealizationの学習用データを用いた5-foldのクロスバリデーションでカーネルパラメタを定めるものである. 表 4.3 ではKFDA(C)と記した. SDP-MKLとR-MKLはソフトマージンSVMを用いるため, ソフトマージンパラメタの決定が必要である. このパラメタも, 初めの5つのrealizationを用いた5-foldクロスバリデーションで定めた. また, MCEMアルゴリズムにおける正則化パラメタηも同様である. KFDA (4.9) におけるパラメタζは全ての実験において$\zeta = 0.001$で固定した.

表 4.3: KFDA, KLFDA, MCEM アルゴリズムによる誤判別率の平均と分散. 最も精度の良い結果と, 5%の有意水準の t テストによって同等と検定された結果は太字で表示してある. また, ユークリッド空間における最近傍法の結果と比較して 5%の有意水準の t テストによって有意に精度が向上したと検定された結果には, ◦ を記してある

Data name	KFDA(CV)	KFDA(H)	KFDA-MKL	SDP-MKL	R-MKL	MCEM.R	MCEM.Q	MCEM.E
banana	31.26(3.40)	**15.00**(0.98)	**14.97**(0.85)	17.05(4.75)	17.01(1.31)	15.83(1.19)	15.66(1.37)	26.44(12.89)
breast-cancer	31.76(4.84)	31.8(4.91)	37.05(5.48)	47.08(5.87)	35.01(5.63)	**31.36**(4.86)	**31.06**(5.68)	32.32(4.66)
diabetes	30.23(2.44)	29.44(2.21)	28.56(2.20)	31.88(3.99)	30.73(2.78)	27.38(2.48)	**25.73**(2.27)	29.54(2.59)
flare-solar	**35.48**(2.09)	**35.66**(2.18)	36.10(1.82)	37.37(1.92)	36.78(1.87)	36.02(1.96)	36.42(1.99)	**35.58**(2.03)
german	28.91(2.88)	29.31(2.67)	28.17(2.59)	28.82(2.25)	30.34(2.63)	27.24(2.52)	**23.56**(2.45)	28.46(2.63)
heart	21.12(3.72)	21.39(2.67)	21.00(4.05)	**17.45**(4.17)	22.36(3.82)	20.29(4.34)	20.86(4.68)	22.99(7.04)
image	12.86(1.23)	11.8(1.4)	20.69(2.05)	12.81(5.05)	**9.99**(0.63)	12.80(1.60)	13.14(3.78)	19.71(9.05)
ringnorm	2.06(0.45)	2.06(0.38)	3.76(2.36)	**1.70**(0.41)	2.24(0.46)	2.31(0.87)	2.51(1.08)	4.68(3.43)
splice	18.14(0.76)	20.16(1.13)	18.00(2.21)	24.86(6.99)	18.04(1.14)	**17.40**(1.34)	23.82(10.19)	21.24(5.72)
thyroid	**5.45**(2.27)	5.93(2.39)	5.88(2.46)	**5.14**(2.72)	6.20(2.50)	6.99(2.90)	8.62(4.86)	19.07(4.42)
titanic	22.61(1.05)	**22.37**(1.06)	**22.21**(1.05)	22.50(1.08)	**22.48**(1.03)	**22.22**(1.05)	**22.47**(1.09)	**22.32**(1.05)
twonorm	3.21(0.45)	3.21(0.45)	3.56(1.00)	**2.54**(0.25)	3.31(0.46)	3.26(0.63)	3.27(0.75)	3.89(6.15)
waveform	11.67(0.74)	12.03(0.82)	13.65(1.42)	12.69(2.00)	12.77(1.05)	**11.03**(1.08)	**11.04**(0.97)	12.70(8.62)

表 4.3 から, カーネル法による非線型化により, 多くのデータに対して線型次元削減手法を前処理とした判別結果 (表 4.2) を大きく上回る精度が達成できることが分かる.

幾つかのデータに対しては, カーネル法を用いた場合に線型手法より判別精度が劣る. これは, 元のデータが線型手法によって十分に判別可能であることが理由であると考えられる. カーネル法はデータを非線型写像により特徴空間に写像した上で判別を行う方法である. 従って, 元の空間において十分線型判別が可能なデータにカーネル法を適用することで, 判別性能が低下する可能性がある. 表 4.3 から, “banana”, “image” 及び”thyroid” データに対しては, カーネル法はユークリッド空間における最近傍法よりも精度が低いことが分かる. これらのデータに対してはユークリッド空間においても比較的良好な判別精度が得られており, このことがカーネル法による精度低下の原因であると考えられる.

提案した MCEM アルゴリズムでは SDP-MKL や R-MKL と同様に正則化パラメタを定める必要があるが, 経験的には η の値は精度に大きな影響を及ぼさず, $\eta = 1.5$ 程度に固定してよいことがわかっている.

次に, 単一カーネル関数を用いた SVM, SDP-MKL, KFDA-MKL, R-MKL, MCEM.R, MCEM.Q, MCEM.E を, 酵母タンパク質の同定問題に適用する. このデータは, 論文 [55] のサポートウェブサイトから入手可能である. IDA データセットを用いた実験では, パラメタの異なるガウスカーネルの凸結合の最適化を考えたが, ここでは酵母タンパク質の性質を異なる手法で評価して得られる 3 種類のカーネル行列の凸結合を考える. 3 種類のカーネル行列はそれぞれ,

1. 遺伝子発現情報 (実数値データ)
2. タンパク質内の Pfam 領域[4]の有無に関する E-value(実数値データ)
3. タンパク質配列類似度 (可変長配列データ)

である. このように, 異なる形式のデータを同一の枠組みで扱えるのがカーネル法の大きな利点の一つであった. タンパク質の持つ 12 の機能に着目し, 各タンパク質がそれらの機能を有するか否かの判定を行う 12 個の判別問題を考えた. つまり, 各機能に対応して 12 個の 2 値判別器を学習した. もとのデータは 3588 種のタンパク質に関する大規模なカーネル行

[4]タンパク質の各機能が共通して持つ配列のデータベースで, 各機能を持つタンパク質群の配列を並べ, 共通配列を抽出したのが Pfam ドメイン

列である. 特に半正定値計画問題に基づく手法は, 大規模なデータに適用することは困難である. ここでは, 先行研究 [56] で行われた方法に従い, ランダムに 500 個のタンパク質をサンプリングし, さらに 1 対 2 にデータを分割して 2/3 のデータで判別器を学習し, 残りのデータでテストをするという 3-fold のクロスバリデーションを行った. 500 個のタンパク質のサンプリングとクロスバリデーションを, 5 回繰り返した. 表 4.4 に, この 15 回の実験で学習した判別器の ROC 曲線下面積 (AUC: area under the curve) の平均を示した. SDP-MKL と R-MKL については, 内部で利用するソフトマージン SVM のソフトマージンパラメタを色々と変えて実験し, 最良の結果が得られるパラメタを採用した. MCEM アルゴリズムにおける η パラメタも同様にして定めた.

提案する MCEM アルゴリズムは, 所与のデータの分布が特徴空間においてクラス毎にコンパクトに纏まるようにカーネル関数を学習するものである. これは必ずしも KFDA のような特定の判別器に特化した方法ではない. つまり, KFDA 以外の判別器に用いるためのカーネル関数の学習にも利用可能である. そこで, MCEM アルゴリズムによって学習したカーネル行列を用いて, SVM によって判別を行った結果も示す. なお, ここで用いた SVM のソフトマージンパラメタ及び η は簡単のため 1 に固定した. 表 4.4 から, 提案する MCEM アルゴリズムは既存の MKL 手法と同等の性能を示すことが分る.

なお, 今回の実験では MCEM.E アルゴリズムの判別性能は MCEM.R, MCEM.Q アルゴリズムと比較して優れたものではなかった. しかし, MCEM.E アルゴリズムは固有値問題に基づく方法であり, 計算速度は非常に早い. そこで, 大規模なデータに対する MCEM.R や MCEM.Q アルゴリズムの初期値を定めるための前処理としての利用が効果的と考えられる.

表 4.4: 酵母タンパク質機能同定問題に対する MKL 手法の比較．タンパク質の各機能 (1 から 12 番まで) に対し, 各判別器による AUC の平均値が記してある．最大の AUC は太字で記した．初めの 3 行は, 単一のカーネル行列を用いた SVM による結果である (それぞれ遺伝子発現, タンパク質-タンパク質相互作用及び配列の類似度).

Function	Exp	Dom	Seq	SDP-MKL	R-MKL	KFDA-MKL	MCEM.R	MCEM.Q	MCEM.E	MCEM.R +SVM	MCEM.Q +SVM	MCEM.E +SVM
1	0.682	0.767	0.774	0.778	0.778	0.784	0.784	0.766	0.712	**0.796**	0.776	0.718
2	0.708	0.676	0.689	0.737	0.725	0.749	0.736	**0.748**	0.712	0.713	0.728	0.722
3	0.619	0.689	0.688	0.683	**0.699**	0.693	**0.699**	0.692	0.647	0.697	0.695	0.664
4	0.706	0.733	0.758	**0.786**	0.776	0.786	0.769	0.771	0.734	0.769	0.770	0.734
5	0.854	0.789	0.777	0.856	**0.874**	0.807	0.804	0.817	0.845	0.803	0.834	0.852
6	0.590	0.655	0.688	**0.692**	0.680	0.694	0.690	0.682	0.626	**0.692**	0.688	0.621
7	0.570	0.678	0.708	**0.714**	0.703	0.697	0.710	0.695	0.625	**0.714**	0.704	0.610
8	0.612	0.635	0.669	0.711	0.716	0.710	0.700	**0.746**	0.703	0.684	0.726	0.677
9	0.686	0.744	0.741	0.783	0.775	0.804	0.752	0.768	0.777	0.750	**0.784**	0.768
10	0.622	0.658	0.701	0.698	0.660	0.669	**0.705**	0.673	0.626	0.703	0.674	0.642
11	0.612	0.585	0.608	0.586	0.593	0.599	0.613	0.611	0.619	0.597	0.582	**0.617**
12	0.657	**0.911**	0.883	0.875	0.895	0.900	0.885	0.832	0.841	0.886	0.848	0.842

4.4 条件付きエントロピー最小化基準による距離構造の学習に関するまとめ

本章では距離構造の学習問題として, 特に次元削減と MKL 手法を考察した. 距離構造の学習問題を情報論的な最適化問題として取り扱い, 教師付き距離構造学習のための一般的な枠組みとして, 条件付きエントロピー最小化基準を提案した.

近年, データ分布の局所性に着目した次元削減や多様体学習の手法が盛んに研究されている. 本論文で LCEM アルゴリズムと比較した LFDA の他にも, 多様体学習の手法として locality preserving projection (LPP; [4]) や Laplacian eigenmap (LE; [3]) などが, LFDA と同様に局所的な類似度行列を陽に用いる手法として知られている. LPP や LE ではデータが低次元空間に写像されるが, このとき元の空間において近くに存在するデータは再び近くにまとまって分布するような写像を学習する. LPP や LE の最適化問題は, データ同士の類似度行列から計算されるグラフラプラシアンを用いて表現され, 一般化固有値問題として定式化される. 一方, 本論文で提案した枠組みは, 写像したデータのクラス条件付きエントロピーが最適化の目的関数であり, データの局所的な性質を直接的にモデル化するものではない. しかし, エントロピー推定 [24] において k 近傍法に基づく方法を用いており, これを介してデータの局所性が自然に目的関数に反映されたために LFDA のような局所性を積極的に利用した手法と同等の精度が得られていると考えられる. なお, 古典的な FDA ではデータに正規分布に従っているという仮定をおいていることになる. 一方, 上述の LFDA ではデータの生成モデルとしてどのような分布を仮定しているかは明らかでない. こうした次元削減の諸手法の背後にある確率モデルを明らかにし, LCEM アルゴリズムとの関係を考察することは重要な研究課題である.

また, 条件付きエントロピー最小化基準から, 新しい MKL 手法を提案した. 2004 年の Lackriet らによる SDP に基づく MKL の提案以来, 多くの MKL 手法が提案されているが, 条件付きエントロピー最小化基準に基づく手法は本研究において初めて提案されたものである. 提案した MCEM アルゴリズムは新しいのみではなく, 実データに対して優れた性能を示した. また, 既存の他の MKL 手法と比較しても同等の性能が確認された. さらに, MCEM アルゴリズムは, 例えば SVM など他の判別器に用いるカーネル関数の学習のための前処理としても利用が可能である.

第5章　生成モデルに基づく距離の学習

本章では, グループ化ランキング観測データという特殊な, しかし近年重要性が増して来ているデータに対して新しい確率モデルを提案する. このモデルのパラメタを最尤推定で直接求めることが困難なので, 尤度の近似を導出し, 確率分布関数がなす統計多様体における距離構造に着目し, モデルのパラメタ推定手法を情報幾何学的観点から導出する.

5.1　ランキングモデルの研究

多数のユーザが多数のアイテムに与えた評価データの生成モデルに関する研究は古くから行われており, 賭け事の予想, 心理学における感性データ解析, 官能検査や, 経済学における効用理論もこうした研究の例として理解できる. 特に評価データとしてランキングデータを考えたとき, その生成モデルは, 以下の 2 種類に分類出来る:

1. アイテムの順序そのものに着目したモデル.
2. 個々のアイテム価値に着目したモデル.

前者のモデルとしては, Mallow のモデル [57] や Fligner のモデル [58] が有名である. アイテムの順序の分布に関する研究は [59] に詳しく, また, 最近でも多くの研究者によって盛んに研究されている [60; 61; 62; 63; 64].

後者の研究としては, Bradley と Terry による確率モデル [65] が代表的である. Bradley と Terry は, 各アイテム I_i に対してその総和が 1 になるように正規化された正値のパラメタ θ_i を割り当て, アイテム I_i, I_j の比較においてアイテム I_i が選択される確率を,

$$P(I_i \succ I_j) = \frac{\theta_i}{\theta_i + \theta_j}$$

で定義した.

Bradley-Terry モデルは機械学習分野においてよく知られた解析ツールとなっている [66; 67; 68]. Bradley-Terry モデルの自然な拡張として Plackett-Luce モデルがある [69]. このモデルは, N 個のアイテムに順序をつけるプロセスを, アイテムの選好度パラメタ $\theta_i,\ i=1,\ldots,N$ に基づく逐次的なアイテム選択として表現し, ランキングデータ $(I_{a(1)} \succ I_{a(2)} \succ \cdots \succ I_{a(N)})$ を観測する確率を

$$\begin{aligned} P(I_{a(1)} \succ I_{a(2)} \succ \cdots \succ I_{a(N)}) &= \frac{\theta_{a(1)}}{\sum_{j=1}^{N}\theta_{a(j)}} \frac{\theta_{a(2)}}{\sum_{j=2}^{N}\theta_{a(j)}} \cdots \frac{\theta_{a(N-1)}}{\theta_{a(N-1)}+\theta_{a(N)}} \\ &= \prod_{i=1}^{N-1} \frac{\theta_{a(i)}}{\sum_{j=i}^{N}\theta_{a(j)}} \ , \end{aligned} \tag{5.1}$$

で定めるものである. ここで, $a(j)$ はこのランキングデータの中で j 番目の位置を占めるアイテムの添字を表す. このモデルは, 選好度パラメタの値が高いアイテムほど先に選ばれる傾向があるという仮定を反映している. この Plackett-Luce モデルに対しては, Hunter [70] が尤度の下界を与え, その下界を最大化する繰り返しアルゴリズムを提案している.

次に, 複数の評価者 (ここではユーザ) が, 複数のアイテムに対して, 離散値で評価を与える状況を考える. 厳密な定義は後で与えるとして, 本論文ではこうした形式のデータを**グループ化ランキング観測データ**と呼ぶ. 例えば映画や書籍, レストランなどへのユーザによる評価はグループ化ランキング観測データとして表現される. 近年, こうしたデータはインターネットの普及に伴い大量に蓄積されており, その解析手法は重要な研究課題である. 本章では, このグループ化ランキング観測データを扱うため, Plackett-Luce を一般化した**グループ化ランキングモデル**を提案する. このモデルは, N 個のアイテムそれぞれに対して, M 段階の離散値の評価が与えられる確率を表現するものである. このモデルの基本的な考え方は, 観測可能なのは各アイテムがどの評価を得られたかという情報のみであるが, 実際には同じ値の評価を得たアイテム間にも観測不可能な順序が存在する, という仮定である. 単純な例として, 7 個のアイテム $I=\{I_1,\ldots,I_7\}$ が $M=3$ 段階の評価を与えられるとする. このとき, 一人のユーザ u から, 次のような形でグループ化ランキング観測データが得られる: $D^u=\{G_1^u,G_2^u,G_3^u\}, G_1^u=\{3,5\}, G_2^u=\{2,6,7\}, G_3^u=\{1,4\}$. ここで, G_m^u はユーザ u によって評価値 m を与えられたアイテムの添字集合である. これらの観測データが U 人のユーザから独立に得られるものとする.

つまり, 観測データは $\{D^u\}_{u=1}^U$ である. ここで, N 個のアイテムはそれぞれ θ_i という選好度パラメタを持っているものとする. 選好度パラメタは, 多項分布のパラメタと同様に $\theta_i > 0$, $\sum_{i=1}^N \theta_i = 1$ なる性質を満たしているものとする. グループ化ランキング観測データ $\{D^u\}_{u=1}^U$ のみを用いて, パラメタ $\theta = (\theta_1, \ldots, \theta_N)$ を推定するのが目的である.

文献 [71] によると, ランキングデータの形式的なモデル化は次の 2 種類に分類される:

1. ランキングプロセスのモデル化.
2. ランキングを与えるユーザのモデル化.

単一の Plackett-Luce モデルのみによって, ユーザそれぞれで異なりうるランキングプロセスを全て説明することができるとは考えにくい. そこで, ランキングモデルの混合モデルを考えるのが自然である. 本論文では, 提案するグループ化ランキングモデルの混合モデルを提案し, アイテム選好度パラメタと混合パラメタの推定方法を与える. また, 学習した混合モデルを, アイテムとユーザの可視化及び協調フィルタリングに応用した結果を示す.

5.2 グループ化ランキングモデル

まず, 本章で考察するグループ化ランキングデータを定義し, このデータに対する生成モデルを導出する.

5.2.1 モデルの定義と尤度関数

U 人のユーザが独立に, 1 から M までの離散評価値を N 個のアイテム $I_1, \ldots, I_N$ に与えるとする. G_1^u を, ユーザ u が最も良いと評価したアイテムの添字集合, G_2^u をその次に良いと評価したアイテムの添字集合とし, 以下同様に $G_3^u, \ldots, G_M^u$ とする. なお, モデルの導出の簡単のため, ここではユーザは全てのアイテムに評価を与えるとする. 実際は全てのユーザが全てのアイテムに評価を与えることは考えにくく, 未評価アイテムの取り扱いは重要である. 未評価アイテムがある場合の取り扱いについては付録 7 にて詳述する.

ユーザ u による評価アイテムの添字集合を M 個集めたものを $D^u = \{G_1^u, \ldots, G_M^u\}$, $G_m^u = \{\, i \mid I_i \in m$ 番目のグループ $\}$ とする. また, $\gamma_m^u = |G_m^u|$ とする. 本論文では, D^u をユーザ u による**グループ化ランキング観測データ**と呼ぶことにする. 同一の評価値を与えられたアイテム同士の優劣は観測出来ないが, 実際には各アイテムグループの中に順序がつけられるものと仮定する. グループ G_m^u の中の添字に順序を考える必要がある場合には, この集合 G_m^u に作用する置換群 π_m^u によりそれを表現する. $\pi_m^u(i)$ で, 順序付きの添字集合 $\pi_m^u(G_m^u)$ の中の i 番目のアイテムの添字を表すものとする. 例えば, グループ $G_m^u = \{2,6,7\}$ への置換 $\pi_m^u = \left(\begin{smallmatrix}1&2&3\\2&3&1\end{smallmatrix}\right)$ の作用により, 順序付きグループ $\pi_m^u(G_m^u) = (7,2,6)$ を得る. ここで, $\pi_m^u(1) = 7$, $\pi_m^u(2) = 2$, $\pi_m^u(3) = 6$ である.

第1節で用いた例を用いて, 提案するモデルの生成プロセスを詳述し, 尤度関数を導出する. 今, 7個のアイテムに対する一つのグループ化ランキング観測データ $D^u = \{G_1^u = \{3,5\}, G_2^u = \{2,6,7\}, G_3^u = \{1,4\}\ \}$ を観測したとする. グループ G_1^u に含まれる任意のアイテムは, G_2^u に含まれる任意のアイテムよりこのユーザ u から高評価を得たということは分かるが, グループ内部のアイテムに関する順序の情報は有していない. ここで, グループ G_2^u においてユーザは実際には $I_7 \succ I_2 \succ I_6$ なる順序でアイテムを評価したと仮定する. つまり, $\pi_2^u = \left(\begin{smallmatrix}1&2&3\\2&3&1\end{smallmatrix}\right)$ ということである. ここで, グループ G_m^u に添字が属する選好度パラメタ θ_i の総和を $\Theta_m^u = \sum_{i \in G_m^u} \theta_i$ で定義し, D^u におけるグループ G_m^u のグループパラメタと呼ぶことにする.

グループ G_1^u に含まれるアイテムは既に選択されて取り除かれていると考え, G_2^u に含まれるアイテムが上で仮定した順序で選択される確率は次式で与えられる:

$$
\begin{aligned}
&P((\pi_2^u, G_2^u)|G_1^u) \\
&= \left(\frac{\theta_7}{\sum_{n=2}^3 \Theta_n^u}\right)\left(\frac{\theta_2}{\sum_{n=2}^3 \Theta_n^u - \theta_7}\right)\left(\frac{\theta_6}{\sum_{n=2}^3 \Theta_n^u - (\theta_7+\theta_2)}\right) \\
&= \prod_{i=1}^{\gamma_2^u} \frac{\theta_{\pi_2^u(i)}}{\sum_{n=2}^3 \Theta_n^u - \sum_{j<i} \theta_{\pi_2^u(j)}}.
\end{aligned}
$$

ここで, $P((\pi_m^u, G_m^u)|G_1^u, \ldots, G_{m-1}^u)$ はグループ $\{G_1^u, \ldots, G_{m-1}^u\}$ のアイテムが既に選択されているという状況で, グループ G_m^u のアイテムが π_m^u で定まる順序で選択されるという条件付き確率である. なお, 既に選ばれたグループ $G_1^u, \ldots, G_{m-1}^u$ 内部のアイテムの順序は, m 番目のグループ π_m^u 内の順序に影響しないことに注意する.

ここで, 上記の例を一般化して提案モデルを導く. 完全な観測データを $\{(\pi_1^u, G_1^u), \ldots, (\pi_M^u, G_M^u)\}$ とすると, 置換 π_m^u は観測出来ない隠れ変数とみなせる. この完全な観測データを次のように分解する:

$$\begin{aligned} &P((\pi_1^u, G_1^u), \ldots, (\pi_M^u, G_M^u)) \\ &= P((\pi_1^u, G_1^u))P((\pi_2^u, G_2^u)|G_1^u) \cdots P((\pi_M^u, G_M^u)|G_1^u, \ldots, G_{M-1}^u). \end{aligned}$$

ここで, $\{G_1^u, \ldots, G_{m-1}^u\}$ を既に観測した状況で, グループ G_m^u を順序 π_m^u 付きで観測する確率を

$$P((\pi_m^u, G_m^u)|G_1^u, \ldots, G_{m-1}^u) = \prod_{i=1}^{\gamma_m^u} \frac{\theta_{\pi_m^u(i)}}{\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^u(j)}} \tag{5.2}$$

で定める.

各グループの隠れた順序は, G_m^u に作用する置換 π_m^u によって表現されることに注意して, $\mathcal{S}(G_m^u)$ で $\gamma_m^u!$ 通りの取りうる全ての置換を表す. 以上より, グループ G_m^u を観測する確率は, 取りうる全ての隠れた順序に対応する確率の総和として,

$$P(G_m^u|G_1^u, \ldots, G_{m-1}^u) = \sum_{\pi_m^u \in \mathcal{S}(G_m^u)} P((\pi_m^u, G_m^u)|G_1^u, \ldots, G_{m-1}^u) \tag{5.3}$$

で定義出来る. これは γ_m^u 個のアイテムに対する Plackett-Luce モデルにおいて, アイテムの置換に関する周辺化を施したものに他ならない. また, もし $G_m^u = \emptyset$, つまり評価値 m を与えられたアイテムが存在しない時には, 上記の和において対応する m の和を除けばよい. こうして定義された m 番目の評価値を得たアイテムグループの観測確率から, データ $D^u = \{G_1^u, \ldots, G_M^u\}$ の尤度は

$$P(D^u) = \prod_{m=1}^{M} \sum_{\pi_m^u \in \mathcal{S}(G_m^u)} P((\pi_m^u, G_m^u)|G_1^u, \ldots, G_{m-1}^u)$$

で得られる. これを, **グループ化ランキングモデル**と呼ぶ. このモデルにおけるデータ生成プロセスに関する詳しい説明と, このモデルが確率モデルとして妥当なものであるという議論は付録 4 に記す. 式 (5.2) と (5.3) から, グループ G_m^u に関する対数尤度

$$l(\theta; m, u) = \log \left(\sum_{\pi_m^u \in \mathcal{S}(G_m^u)} \prod_{i=1}^{\gamma_m^u} \frac{\theta_{\pi_m^u(i)}}{\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^u(j)}} \right) \tag{5.4}$$

を得る. これを全てのグループと全てのユーザについて加えることで, 与えられたデータ $\{D^u\}_{u=1}^U$ の対数尤度

$$L(\theta) = \sum_{u=1}^{U}\sum_{m=1}^{M} \log \left(\sum_{\pi_m^u \in \mathcal{S}(G_m^u)} \prod_{i=1}^{\gamma_m^u} \frac{\theta_{\pi_m^u(i)}}{\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^u(j)}} \right) \tag{5.5}$$

が得られる.

このモデルの尤度の最大化は明らかに困難な問題である. 困難さの主な原因は, モデルに含まれる隠れた順序に関する周辺化処理である. 周辺化の計算のためには考えうる全ての順序を列挙する必要があり, この順序の列挙という操作は比較的少数のアイテムを扱うときでも計算量的に困難である.

5.2.2　対数尤度の近似

対数尤度関数 (5.5) を, 置換に関する周辺化を含まないように近似する.

まず, 式 (5.4) の分母が, 連続的なアイテム選択におけるパラメタ θ_i の正規化という意味を持つことに着目する. 式 (5.4) の分母を, $\sum_{n=m}^{M} \Theta_n^u$ で置き換えることで, (5.4) の下界

$$\begin{aligned} \underline{l}(\theta; m, u) &= \log \left(\sum_{\pi \in \mathcal{S}(G_m^u)} \prod_{i=1}^{\gamma_m^u} \frac{\theta_{\pi_m^u(i)}}{\sum_{n=m}^{M} \Theta_n^u} \right) \\ &= \log \left(\gamma_m^u! \prod_{i \in G_m^u} \frac{\theta_i}{\sum_{n=m}^{M} \Theta_n^u} \right) \leq l(\theta; m, u) \end{aligned}$$

を得る. この分母の置換えにより, 式 (5.4) における周辺化は単純な定数 $\log(\gamma_m^u!)$ で置き換わることに注意する. グループ数 M が大きくなると一つ一つのグループに含まれるアイテム数は減少し, 項 $\sum_{j<i} \theta_{\pi_m^u(j)}$ を無視することの影響は小さくなる.

次に, $\underline{l}(\theta; m, u)$ の上界を与える. 相加・相乗平均の関係

$$\left(\prod_{i \in G_m^u} \theta_i \right)^{1/\gamma_m^u} \leq \frac{1}{\gamma_m^u} \sum_{i \in G_m^u} \theta_i = \frac{1}{\gamma_m^u} \Theta_m^u$$

を用いると, 次の不等式を得る:

$$\gamma_m^u! \prod_{i \in G_m^u} \frac{\theta_i}{\sum_{n=m}^M \Theta_n^u} = \gamma_m^u! \frac{\prod_{i \in G_m^u} \theta_i}{\left(\sum_{n=m}^M \Theta_n^u\right)^{\gamma_m^u}} \leq \gamma_m^u! \left(\frac{\Theta_m^u/\gamma_m^u}{\sum_{n=m}^M \Theta_n^u}\right)^{\gamma_m^u}.$$

従って, $\underline{l}(\theta; m, u)$ の上界

$$\begin{aligned} \tilde{l}(\theta; m, u) &= \log \left\{ \gamma_m^u! \left(\frac{\Theta_m^u/\gamma_m^u}{\sum_{n=m}^M \Theta_n^u}\right)^{\gamma_m^u} \right\} \qquad (5.6) \\ &= \log \gamma_m^u! + \gamma_m^u \left(\log \Theta_m^u - \log \gamma_m^u - \log \left(\sum_{n=m}^M \Theta_n^u \right) \right) \\ &= \gamma_m^u \left\{ \log \Theta_m^u - \log \left(\sum_{n=m}^M \Theta_n^u \right) \right\} + \log \gamma_m^u! - \gamma_m^u \log \gamma_m^u \end{aligned}$$

が得られる. $\tilde{l}(\theta; m, u)$ は対数尤度 $l(\theta; m, u)$ の厳密な下界になっているわけではないが, 準備的な実験により多くの場合 $\tilde{L}(\theta) = \sum_{u=1}^U \sum_{m=1}^M \tilde{l}(\theta; m, u)$ は正確な対数尤度 $L(\theta)$ よりも小さな値をとることが確認されている. 従って, 次の近似式を得る:

$$\underline{l}(\theta; m, u) \leq \tilde{l}(\theta; m, u) \lesssim l(\theta; m, u). \qquad (5.7)$$

この近似の正当性は, 後に実験的に示す.

全てのグループとユーザに関して $\tilde{l}(\theta; m, u)$ を足し合わせることで,

$$\tilde{L}(\theta) = \sum_{u=1}^U \sum_{m=1}^M \gamma_m^u \left(\log \frac{\Theta_m^u}{\sum_{n=m}^M \Theta_n^u} \right) + const. \qquad (5.8)$$

なる対数尤度の近似式が得られる. この関数 $\tilde{L}(\theta)$ は置換の列挙を含んでいない. また, $L(\theta) \gtrsim \tilde{L}(\theta)$ が成立するので, $\tilde{L}(\theta)$ の最大化は間接的に対数尤度 $L(\theta)$ の最大化につながることが期待できる.

近似した対数尤度 $\tilde{L}(\theta)$ を θ に関して最大化する方法は任意である. しかし, $\tilde{L}(\theta)$ を直接最大化するという問題には 2 つの困難さが残る. 一つ目は, 計算量の問題である. 近似尤度 $\tilde{L}(\theta)$ は $\theta = (\theta_1, \ldots, \theta_N)$ に関する非線型関数であり, その最大化に要する計算時間は N の増加に伴い大幅に増加する. 準備的な実験では, $\tilde{L}(\theta)$ を Nelder-Mead 法によって θ に関して最大化したところ, アイテムの数 N に対して線型以上のオーダーで

計算時間が増加した. 二つ目の問題は, 特にオンライン処理が必要な場合には深刻な問題である. 尤度関数を通常の方法で最大化した場合は, 新しいユーザがシステムに参加する毎に $\tilde{L}(\theta)$ を最大化し直さなければならない. 次節で, 近似尤度を最大化するパラメタを求める効率的な方法を提案する.

5.3 パラメタ推定のアルゴリズム

本節では, 情報幾何学 [72] の観点からパラメタ推定手法を導出する. 尤度関数を最大化するのは, 全ての観測データと可能な限り矛盾が小さくなるようなパラメタ $\theta = \{\theta_i\}_{i=1}^N$ を求めるのが目的である. そのために, (5.8) の第一項を, Θ_m^u に関する U 個の独立な最適化問題に分解する.

$$\max_{\{\Theta_m^u\}} \sum_{m=1}^{M} \gamma_m^u \log\left(\frac{\Theta_m^u}{\sum_{n=m}^{M}\Theta_n^u}\right), \quad \text{subject to} \quad \Theta_m^u > 0, \ \sum_{m=1}^{M}\Theta_m^u = 1. \tag{5.9}$$

これらの問題は, 線型制約を持つ M 変数の最適化問題であり, 非線型最適化問題を扱うことができる任意のソルバーによって簡単に解くことが出来る. なお, パラメタの正値性条件 $\Theta_m^u > 0$ を課すために, 対数障壁関数 $\sum_{m=1}^{M}\frac{1}{M}\log\Theta_m^u$ を (5.9) の目的関数に加え, 最適化問題

$$\max_{\{\Theta_m^u\}} \sum_{m=1}^{M} \gamma_m^u \log\left(\frac{\Theta_m^u}{\sum_{n=m}^{M}\Theta_n^u}\right) + \sum_{m=1}^{M}\frac{1}{M}\log\Theta_m^u, \quad \text{subject to} \quad \sum_{m=1}^{M}\Theta_m^u = 1 \tag{5.10}$$

を考えることにする. この障壁関数は目的関数である分布に事前分布として一様分布を与えることに相当する. なぜなら, 一様分布 $\{\Theta_m^u\}_{m=1}^M = \{\frac{1}{M},\ldots,\frac{1}{M}\}$ がこの障壁関数を最大化するからである. ランキングモデルの最適化における障壁関数及び正則化に関する詳細な解析は, 例えば [70] や [62] などで詳しく議論されている.

上記の U 個の小規模な最適化問題を解くことで, U 個のグループパラメタ $\{\Theta_m^u\}_{m=1}^M,\ u=1,\ldots,U$ が得られる. 問題は, これら U 個の解と最も整合性のとれたパラメタ θ を求めることである. 本論文では, この θ の推定問題を em アルゴリズムの枠組みで解く方法を与える [73]. このアルゴリズムは, 確率モデルがなす空間と, 観測データがなす空間の間で射影を繰り返すことで, θ の局所最適解を与えるものである.

最適化問題 (5.10) の解は, 選好度パラメタに関する不完全な観測データとみなすことが出来る. 選好度パラメタ θ には, $^{\forall}i,\ \theta_i > 0$, $\sum_{i=1}^{N} \theta_i = 1$ という制約があるので, $(N-1)$-確率単体と呼ばれる多様体 Δ_{N-1} を構成する. 最適化問題 (5.10) の解 $\{\hat{\Theta}^u_m\}_{m=1}^M$ は, 確率単体 Δ_{N-1} の部分多様体を構成する (図 1a):

$$\mathcal{D}_u = \{\, \theta \mid \sum_{i \in G^u_m} \theta_i = \hat{\Theta}^u_m,\ m = 1, \ldots, M\} \subset \Delta_{N-1}.$$

選好度パラメタ $\theta = \{\theta_i\}$ を離散確率測度と同一視することで, 最適なパラメタ θ は Δ_{N-1} 上の一点で, 全ての部分多様体 $\{\mathcal{D}_u\}_{u=1}^U$ と Kullback-Leibler (KL) ダイバージェンス $KL(\theta, \theta') = \sum_{i=1}^{N} \theta_i \log(\theta_i/\theta'_i)$ の意味で最も近い点として求められる. つまり, 目的関数

$$L_{em}(\theta) = \sum_{u=1}^{U} KL(\mathcal{D}_u, \theta) = \sum_{u=1}^{U} \min_{\theta^u \in \mathcal{D}_u} KL(\theta^u, \theta) \tag{5.11}$$

を $\theta \in \Delta_{N-1}$ に関して最適化することが問題となる. この em アルゴリズムは, 目的関数 (5.11) を e ステップと m ステップという処理を繰り返すことで最小化する方法である.

いま, em アルゴリズムの t 回目の繰り返しの後で, パラメタ推定値 $\theta(t)$ を得ているとする. e ステップでは, 部分多様体 $\mathcal{D}_u$ 上において $\theta(t)$ から KL ダイバージェンスの意味で最も近い点 $\hat{\theta}^u(t)$ を求める (図 1b). つまり,

$$\hat{\theta}^u(t) = \arg\min_{\theta \in \mathcal{D}_u} KL(\theta, \theta(t)) \tag{5.12}$$

を計算する. この手続きは e 射影と呼ばれる.

m ステップでは, 各部分多様体 $\mathcal{D}_u$ 上の点 $\hat{\theta}^u(t)$ から KL ダイバージェンスの意味で最も近い Δ_{N-1} 上の一点 $\theta(t+1)$ を求める (図 1c). つまり,

$$\theta(t+1) = \arg\min_{\theta} \sum_{u=1}^{U} KL(\hat{\theta}^u(t), \theta)$$

を計算する. この手続きは m 射影と呼ばれる.

本節で考えているモデルにおける e 射影, m 射影によるパラメタ更新は, より具体的に書き下すことが出来る:

命題 5.1
e 射影は

$$\hat{\theta}^u_i(t) = \frac{\theta_i(t)}{\sum_{j \in G^u_{m|i}} \theta_j(t)} \hat{\Theta}^u_{m|i}\ , \quad i = 1, \ldots, N,\ u = 1, \ldots, U \tag{5.13}$$

と書ける．ここで, $G^u_{m|i}$ はアイテム I_i が属するグループであり, $\hat{\Theta}^u_{m|i}$ は対応するグループパラメタである．m 射影は

$$\theta_i(t+1) = \frac{1}{U}\sum_{u=1}^{U} \hat{\theta}^u_i(t)\ , \quad i = 1, \ldots, N \tag{5.14}$$

と書ける．

証明

付録 *6* に示す．□

これら e ステップと m ステップを収束するまで繰り返すことで, θ の局所最適推定値を得ることができる．図 5.1 に, $N=3$, $U=3$, $M=2$ の時の em アルゴリズムの挙動の模式図を示す．提案モデルに対する em アルゴリズムを, 図 5.2 にまとめる．

本節の最後に, 提案するパラメタ推定のアプローチをまとめておく．対数尤度関数 $L(\theta)$ の評価は, 各グループ内で取りうる全ての順列を列挙した上で周辺化する必要があり, 計算量的に困難である．そこで, 順列の周辺化を含まない形の近似尤度 $\tilde{L}(\theta)$ を導出した．アイテム数 N やユーザ数 U が大きい時は, $\tilde{L}(\theta)$ を θ に関して直接最大化するのも困難である．そこで, $\tilde{L}(\theta)$ を最大化する代わりに, 各ユーザのグループパラメタ $\{\Theta^u_m\}_{m=1}^M$ に関する小規模な U 個の最適化問題 (5.10) を考えた．これらの問題の解として得られるグループパラメタは, パラメタ θ が属する確率単体上で観測部分多様体を構成する．観測データと最も整合性のとれたパラメタ θ は, 確率単体における em アルゴリズムによって推定される．

5.4　グループ化ランキングモデルの混合モデル

本節では, 様々な傾向を持つユーザ集団を表現するため, ランキングモデルの混合モデルを考える．K 個の混合モデルは,

$$P(x) = \sum_{k=1}^{K} \omega_k P(x; \theta^k), \tag{5.15}$$

と表される．ここで, x は一つのデータ, ω_k はデータが k 番目の要素モデルから生成されるという事象の事前確率, θ^k は k 番目のモデルのパラメタである．原理的には, 混合モデル (5.15) のパラメタ $\{\omega_k, \theta^k\}_{k=1}^K$ は EM アルゴリズム [74] によって推定することが出来る．しかし, グループ化ラン

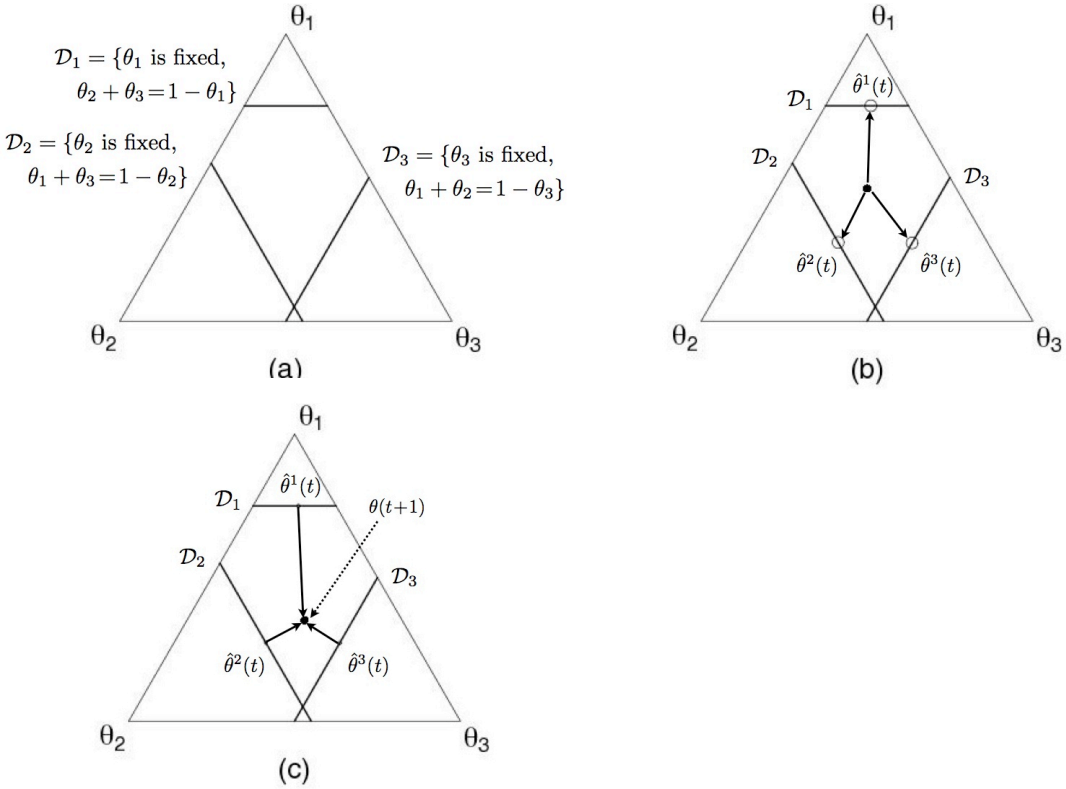

図 5.1: (a): 各観測データは, 確率単体 Δ_{N-1} の部分多様体 $\mathcal{D}_u$ を定義する. (b): 前回の推定値 $\theta(t)$ から各部分多様体に向けて最も近い点 $\hat{\theta}^u(t)$ を求める (e ステップ). (c): 各部分多様体上の点 $\hat{\theta}^u(t)$ から最も近い一点 $\theta(t+1)$ を求める (m ステップ).

キングモデルの混合を考える場合には, 後述するように EM アルゴリズムによる混合モデルのパラメタ推定は困難である. そこで, エントロピー正則化ソフトクラスタリング [75] と呼ばれる手法を採用する.

5.4.1 エントロピー正則化ソフトクラスタリング

ソフトクラスタリングとは, データ $x^u, u=1,\ldots,U$ をクラスタに分割する際に, 各データ x^u が確定的にあるクラスタに属して他のクラスタには

グループ化ランキングモデルのための em アルゴリズム

入力: グループ化ランキング観測データ $\{D^u = \{G^u_m\}_{m=1}^M\}_{u=1}^U$.

初期化: 初期パラメタ $\theta(0)$ を選択し, 次の最適化問題を解いて U 個のグループパラメタ値 $\{\hat{\Theta}^u_m\}_{m=1}^M$ を得る:

$$\max_{\{\Theta^u_m\}} \sum_{m=1}^{M} \gamma^u_m \log\left(\frac{\Theta^u_m}{\sum_{n=m}^{M}\Theta^u_n}\right) + \sum_{m=1}^{M}\frac{1}{M}\log\Theta^u_m,$$
$$\text{subject to} \quad \sum_{m=1}^{M}\Theta^u_m = 1.$$

繰り返し: 収束するまで以下を繰り返す:

e **ステップ:** $\hat{\theta}^u(t)$ を e 射影により更新する:

$$\hat{\theta}^u_i(t) := \frac{\theta_i(t)}{\sum_{j\in G^u_{m|i}}\theta_j(t)}\hat{\Theta}^u_{m|i}\ , \quad i=1,\ldots,N,\ u=1,\ldots,U.$$

m **ステップ:** $\theta(t)$ を m 射影により更新する:

$$\theta_i(t+1) := \frac{1}{U}\sum_{u=1}^{U}\hat{\theta}^u_i(t)\ , \quad i=1,\ldots,N.$$

出力: 収束したパラメタ θ.

図 5.2: 観測データと最も整合性のとれた点を求めるアルゴリズム. 整合性の尺度は, 各観測に対応する部分多様体との KL ダイバージェンスで評価される.

属さないというように分類するのではなく, 各クラスタへのメンバシップ $g_{uk}, k=1,\ldots,K$ を各データに x^u に与える手法である. このメンバシップ g_{uk} は, データ x^u が k 番目のクラスタに" どの程度" 属するかを定めるものである. データ x^u のメンバシップベクトルを $g_u=(g_{u1},\ldots,g_{uK})\in\mathbb{R}^K$ で

定義する. このベクトルは $\sum_{k=1}^{K} g_{uk}=1$, $u=1,\ldots,U$ を満たすように正規化されているものとする. 各クラスタは, クラスタの代表点 $\xi^k, k=1,\ldots,K$ を有しており, データ x^u とクラスタ k の代表点 ξ^k との距離 d_{uk} は適当な距離関数 $d_{uk}=d(x^u,\xi^k)$ で定義する. ソフトクラスタリング手法は数多く提案されているが, その中でもエントロピー正則化は精度の良い方法の一つとして知られている [75].

メンバシップベクトルのエントロピーを $H(g_u)=-\sum_{k=1}^{K} g_{uk}\log g_{uk}$ で定義する. エントロピー正則化ソフトクラスタリングは, 目的関数

$$\begin{aligned} J_\lambda(\{g_{uk}\},\{\xi^k\}) &= \sum_{u=1}^{U}\sum_{k=1}^{K} g_{uk}d_{uk} - \frac{1}{\lambda}\sum_{u=1}^{U} H(g_u) \\ &= \sum_{u=1}^{U}\sum_{k=1}^{K} g_{uk}d_{uk} + \frac{1}{\lambda}\sum_{u=1}^{U}\sum_{k=1}^{K} g_{uk}\log g_{uk} \end{aligned} \tag{5.16}$$

の g_{uk} と ξ^k に関する最小化を, 制約 $\sum_{k=1}^{K} g_{uk}=1$, $u=1,\ldots,U$ のもとで行う. 式 (5.16) の第一項は各データ x^u をクラスタに割り当てた際のコストに相当するものである. データ x^u がクラスタ代表点 ξ^k から遠くはなれている時には, そのクラスタへのメンバシップ g_{uk} は小さくなければならない. 第二項はエントロピー正則化項であり, クラスタの確定的な分割を防ぐ. つまり, 各データが単一のクラスタに属するような状況では, 第二項の正則化項は大きな値をとることになる. この最小化問題は, EM アルゴリズムと良く似た繰り返し演算により局所最適解を得ることが出来る. 以下, エントロピー正則化ソフトクラスタリングの解法を記す. t 回の繰り返しによって得られたメンバシップ g_{uk}, クラスタ代表点 ξ^k 及び距離 d_{uk} をそれぞれ $g_{uk}(t)$, $\xi^k(t)$ 及び $d_{uk}(t)$ とする. 制約 $\sum_{k=1}^{K} g_{uk}=1$, $u=1,\ldots,U$ を考慮すると, メンバーシップの更新式は, ラグランジュの未定乗数法を用いて次式で得られる:

$$g_{uk}(t) = \frac{\exp\left(-\lambda d_{uk}(t-1)\right)}{\sum_{l=1}^{K}\exp\left(-\lambda d_{ul}(t-1)\right)}. \tag{5.17}$$

式 (5.17) によるクラスタへのメンバシップ更新式について考察する. これは, 各ユーザ u のクラスタ k に対する事後確率 $P(k|D^u)$ を定めるものと考えられる. これは

$$g_{uk} \sim P(k|D^u) = \frac{P(D^u|k)P(k)}{P(D^u)} \propto P(D^u|k)P(k) \tag{5.18}$$

であり, 特に k の事前分布を一様分布とすると $g_{uk} \sim P(D^u|k)$ と考えられる. 今, 確率変数 X が値 $(v_1, \ldots, v_n)$ を確率分布 $p = (p_1, \ldots, p_n)$ に従って取るものとする. N 回の試行で v_l という結果を N_l 回得たとする. $\sum_{l=1}^{n} N_l = N$ である. 試行の結果に基づく相対頻度を

$$x_l = \frac{N_l}{N}, \quad (l = 1, \ldots, n) \tag{5.19}$$

として, 経験分布を $\tilde{p} = (x_1, \ldots, x_n)$ とする. 大偏差統計の理論によると, 多項分布から $x = (x_1, \ldots, x_n)$ が観測される確率は

$$p(x_1, \ldots, x_n) \propto \exp\left(NC(x)\right) \tag{5.20}$$

とかける. ここで $C(x)$ は一般にクラメル関数と呼ばれ, ここでは経験分布 $\tilde{p}$ と, X が従う分布 p との KL ダイバージェンスを用いて

$$C(x) = -KL(\tilde{p}, p) \tag{5.21}$$

と表すことができる. 今, 各ユーザデータ D^u は, 多項分布のパラメタがなす単体上に不完全なモデルを定める. このモデルと, クラスタ代表点 ξ^k に対応するモデルとの KL ダイバージェンスの指数によって $P(D^u|k)$ を定める式 (5.17) は, 観測データ D^u がモデル ξ^k から生成された多数のデータ (N 個) の平均統計量であるとして, 大偏差統計の理論により $P(D^u|k)$ を表現したものとみなすことができる. このとき, パラメタ λ は実質的に生成されたデータ数 N に対応しており, λ が大きいほど $P(D^u|k)$ は強いピークを持つ分布になることが分かる. 例えば λ をクロスバリデーションによって定めた場合には, 多項分布に従うサンプリングにより実質的に生成するデータ数を定めたという理解が可能である.

次に, グループ化ランキングモデルの混合モデルのパラメタ推定アルゴリズムを与える. グループ化ランキングモデルにおけるパラメタは $N-1$ 確率単体上の点であり, データ D^u とクラスタ代表点 ξ^k との距離 d_{uk} を

$$d_{uk} = KL(D^u, \xi^k) = \min_{\theta \in \mathcal{D}_u} KL(\theta, \xi^k)$$

で定義する. メンバシップの更新式は (5.17) と同様である. クラスタ代表点の更新はメンバシップの更新のように単純ではなく, 再び em アルゴリズムを用いることになる.

e ステップ: クラスタ代表点 ξ^k からデータ部分多様体 $\mathcal{D}_u$ へ e 射影を行う:

$$\hat{\theta}^{uk} = \arg\min_{\hat{\theta}^u \in \mathcal{D}_u} KL(\hat{\theta}^u, \xi^k), \quad u = 1, \ldots, U.$$

m **ステップ:** データ部分多様体 $\mathcal{D}_u$ からクラスタ代表点 ξ^k へ m 射影を行う:

$$\hat{\xi}^k = \arg \min_{\xi^k \in \Delta_{N-1}} \sum_{u=1}^{U} g_{uk} KL(\hat{\theta}^{uk}, \xi^k), \quad k = 1, \ldots, K.$$

ラグランジュの未定乗数法を用いて, 上記の e, m ステップをより具体的に書き下す. 部分多様体 $\mathcal{D}_u$ 上で, クラスタ代表点 ξ^k に最も近い点の i 番目の成分は,

$$\hat{\theta}_i^{uk} = \frac{\xi_i^k}{\sum_{j \in G_{m|i}^u} \xi_j^k} \Theta_{m|i}^u$$

で得られる. また, クラスタ代表点 ξ^k の i 番目の要素は,

$$\hat{\xi}_i^k = \frac{\sum_{u=1}^{U} g_{uk} \hat{\theta}_i^{uk}}{\sum_{u=1}^{U} g_{uk}}$$

で更新する.

このエントロピー正則化ソフトクラスタリングアルゴリズムによるグループ化ランキングモデルの混合モデルパラメタの推定アルゴリズムを図 5.3 にまとめる.

5.4.2 EM アルゴリズム適用の困難性

本小節では, グループ化ランキングモデルの混合モデルのパラメタ推定を, 混合モデルのパラメタ推定手法としては標準的な方法である EM アルゴリズムで行うことが困難であることの原因を考察する.

まず, EM アルゴリズムをグループ化ランキングモデルの混合モデル

$$P(D^u) = \sum_{k=1}^{K} \omega_k P(D^u; \theta^k) \tag{5.22}$$

のパラメタ推定に適用してみる. 第 k 番目のモデルが選ばれ, このモデルから観測 D^u が得られる同時確率を

$$P(D^u, k; \theta) = \omega_k P(D^u; \theta^k) = \omega_k \prod_{m=1}^{M} P(G_m^u | G_1^u, \ldots, G_{m-1}^u; \theta^k) \tag{5.23}$$

グループ化ランキングモデルの混合モデルのパラメタ推定のためのソフトクラスタリングアルゴリズム.

入力: グループ化ランキング観測データ $\{D^u = \{G^u_m\}_{m=1}^M\}_{u=1}^U$, クラスタ数 K, エントロピー正則化パラメタ λ.

初期化: 最適化問題

$$\max_{\{\Theta^u_m\}} \sum_{m=1}^M \gamma^u_m \log\left(\frac{\Theta^u_m}{\sum_{n=m}^M \Theta^u_n}\right) + \sum_{m=1}^M \frac{1}{M}\log\Theta^u_m, \quad \text{subject to} \quad \sum_{m=1}^M \Theta^u_m = 1$$

を解き, U 個のグループパラメタ値 $\{\hat{\Theta}^u_m\}_{m=1}^M$ を得る. メンバシップを初期化する:$g_{uk}(0) = \frac{1}{K}, k = 1, \ldots, K, u = 1, \ldots, U$. クラス代表点の初期値 $\xi^k(0),\ k = 1, \ldots, K$ を Δ_{N-1} 上でランダムに選ぶ.

繰り返し: 収束するまで以下を繰り返す:

メンバシップの更新: $g_{uk}(t)$ を次式で更新する:

$$g_{uk}(t) := \frac{\exp(-\lambda d_{uk}(t-1))}{\sum_{l=1}^K \exp(-\lambda d_{ul}(t-1))},$$
$$d_{uk} = KL(D^u, \xi^k) = \min_{\theta \in \mathcal{D}_u} KL(\theta, \xi^k).$$

クラスタ代表点の更新: ξ^k を次のように更新する:
em アルゴリズムのループを表す添字 s を導入する.
$\xi^k(t_0) := \xi^k(t-1)$ として, $\hat{\theta}^{uk}(t_0)$ を $\mathcal{D}_u$ からランダムに選ぶ.

繰り返し: $s = 0$ から初めて収束するまで以下を繰り返す:

e **ステップ:** $\hat{\theta}^{uk}(t_s)$ を e 射影によって更新する:

$$\hat{\theta}^{uk}_i(t_s) := \frac{\xi^k_i(t_{s-1})}{\sum_{j \in G^u_{m|i}} \xi^k_j(t_{s-1})} \hat{\Theta}^u_{m|i},$$
$$i = 1, \ldots, N,\ u = 1, \ldots, U,\ k = 1, \ldots, K.$$

m **ステップ:** $\xi^k(t_s)$ を m 射影によって更新する:

$$\hat{\xi}^k_i(t_s) := \frac{\sum_{u=1}^U g_{uk}(t)\hat{\theta}^{uk}_i(t_s)}{\sum_{u=1}^U g_{uk}(t)}, \quad i = 1, \ldots, N,\ k = 1, \ldots, K.$$

$\xi^k(t) := \hat{\xi}^k(t_{s^*})$, ここで s^* は収束した添字 s.

出力: 収束したパラメタ $\{g_{uk}\}, \{\xi^k\}$.

図 5.3: グループ化ランキングモデルの混合モデルのパラメタを, ソフトクラスタリングと em アルゴリズムによって求めるアルゴリズム.

とすると, EM アルゴリズムにおける Q 関数は次のように書ける:

$$Q(\theta|\theta(t)) = \sum_{u=1}^{U}\sum_{k=1}^{K} P(k|D^u;\theta(t)) \log P(D^u, k;\theta). \tag{5.24}$$

ここで, θ で全てのパラメタ $\{\theta^k\}_{k=1}^K, \{\omega_k\}_{k=1}^K$ を表した. $\theta(t)$ は t 回の EM アルゴリズムの繰り返しによって得られているパラメタ推定値である.

ユーザ u によってデータ D^u が与えられたとして, D^u が k 番目のモデルから生成されたという確率は

$$P(k|D^u;\theta(t)) = \frac{P(D^u, k;\theta(t))}{\sum_{l=1}^K P(D^u, l;\theta(t))} = \frac{\omega_k P(D^u;\theta^k(t))}{\sum_{l=1}^K \omega_l P(D^u;\theta^l(t))} \tag{5.25}$$

で計算される. 式 (5.23) と (5.25) を Q 関数 (5.24) に代入し, (5.24) を制約 $\sum_{k=1}^K \omega_k = 1$ のもとで E-, M-ステップの繰り返しによって最大化する.

クラスタ k の事前分布 ω_k の更新式は容易に求められるが, $\theta^k, k=1,\ldots,K$ の計算は周辺化操作を含むため困難である. 単一のグループ化ランキングモデルにおいては, この困難な点を尤度の近似によって回避した. 同様にして, $P(D^u;\theta^k)$ を $\tilde{P}(D^u;\theta^k)$ で置き換えることで, 修正 Q 関数

$$\tilde{Q}(\theta|\theta(t)) = \sum_{u=1}^{U}\sum_{k=1}^{K} \frac{\omega_k(t)\tilde{P}(D^u;\theta^k(t))}{\sum_{l=1}^K \omega_l(t)\tilde{P}(D^u;\theta^l(t))} \left(\log\omega_k + \log\tilde{P}(D^u;\theta^k)\right) \tag{5.26}$$

を得る. ここで

$$\tilde{P}(D^u;\theta^k) = \prod_{m=1}^{M} \gamma_m^u! \frac{\left(\frac{\Theta_n^{uk}}{\gamma_m^u}\right)^{\gamma_m^u}}{\left(\sum_{n=m}^M \Theta_n^{uk}\right)^{\gamma_m^u}}$$

である. 修正 Q 関数 (5.26) を θ_i^k に関して直接最大化するのはやはり困難であるため, ここでも em アルゴリズムを用いる. つまり, $\tilde{Q}(\theta|\theta(t))$ を各ユーザ u に対応する部分に分割して観測部分多様体を構成してから, モデル多様体と観測部分多様体の間で e-,m-射影を繰り返す. しかし, 混合モデルの場合にはこの近似アプローチはうまく働かない. なぜなら, $P(D^u;\theta^k) \gtrsim \tilde{P}(D^u;\theta^k)$ が成立しても, 不等式 $Q(\theta|\theta(t)) \geq \tilde{Q}(\theta|\theta(t))$ は多くの場合成立しないからである. 修正 Q 関数 (5.26) は項

$$\tilde{P}(k|D^u;\theta(t)) = \frac{\omega_k(t)\tilde{P}(D^u;\theta^k(t))}{\sum_{l=1}^K \omega_l(t)\tilde{P}(D^u;\theta^l(t))} \tag{5.27}$$

を含み, この項はしばしば真の値

$$P(k|D^u;\theta(t)) = \frac{\omega_k(t)P(D^u;\theta^k(t))}{\sum_{l=1}^K \omega_l(t)P(D^u;\theta^l(t))} \tag{5.28}$$

よりも大きくなってしまう. ゆえに, $\tilde{Q}$ の最大化は Q の最大化を保証しない. 従って, 単一のグループ化ランキングモデルの場合には有効であった近似手法が混合モデルの場合には適用出来ないことが分かる.

準備的な実験の結果, EM アルゴリズムによって推定した確率 (5.25) はほとんど全ての場合に, 単一の k に対して 1 になり, その他のクラスに対しては 0 になった. この現象は, 尤度の近似 $\tilde{P}(D^u;\theta) \lesssim P(D^u;\theta)$ の性質から説明出来る. 簡単のため事前分布 $\{\omega_k\}_{k=1}^K$ を無視し, $\alpha_k = P(D^u;\theta^k)$, $\varepsilon_k = P(D^u;\theta^k) - \tilde{P}(D^u;\theta^k)$ とする. すると, (5.28) と (5.27) の差は

$$\begin{aligned}
&\frac{P(D^u;\theta^k(t))}{\sum_{l=1}^K P(D^u;\theta^l(t))} - \frac{\tilde{P}(D^u;\theta^k(t))}{\sum_{l=1}^K \tilde{P}(D^u;\theta^l(t))} \\
&= \frac{\alpha_k}{\sum_{l=1}^K \alpha_l} - \frac{\alpha_k - \varepsilon_k}{\sum_{l=1}^K (\alpha_l - \varepsilon_l)} = \frac{\sum_{l=1}^K (\alpha_l\varepsilon_k - \alpha_k\varepsilon_l)}{\left(\sum_{l=1}^K \alpha_l\right)\left(\sum_{l=1}^K (\alpha_l - \varepsilon_l)\right)}
\end{aligned} \tag{5.29}$$

とかける. 差 (5.29) は α_k が比較的小さいとき正になり, α_k が大きい時に負になる. 今, $0 \le P(k|D^u;\theta) \le 1$ なので, $\tilde{P}(k|D^u;\theta)$ は大きな $P(k|D^u;\theta)$ に対してはより大きく計算され, 小さな $P(k|D^u;\theta)$ に対してはより小さく計算される. この効果が, 条件付き分布が $\{0,1\}$ に値をとってしまうことの一つの原因と考えられる.

5.5 混合モデルの応用

本節では, グループ化ランキングモデルの混合モデルの 2 種類の応用を考える. 一つ目は, データ解析と可視化のツールとしての応用であり, 二つ目は協調フィルタリングとしての応用である.

5.5.1 データ可視化

市場分析においては, ユーザ (消費者) とアイテム (製品) との関係性をモデル化し解釈を与えることは非常に重要な課題である. つまり, 価格やデザインなどといったアイテムの表層的な属性の他に, どのアイテム同士

が類似しているか (同じようなユーザに同じような評価を受けるか) と, どのアイテム同士が類似していないか (全く異なるタイプのユーザ集団に受け入れられるか) の解析が重要である.

各ユーザの好みの傾向を, K 個の異なるランキングモデルの混合によって表現することを試みる. この時, 各混合の要素となるモデルは, “典型的なユーザによるアイテム評価モデル” を表していると考えられる. つまり, K 個の理想的なユーザが存在し, それらの選好パターンは $\theta^k = (\theta^k_1, \ldots, \theta^k_N)$, $k = 1, \ldots, K$ で表現されると考える. 従って, 各アイテムは K 通りの理想的なユーザに対応するモデルの選好度パラメタによって表現され, K 次元空間の一点

$$(\theta^1_i, \ldots, \theta^K_i) \in \mathbb{R}^K \tag{5.30}$$

で表すことができる. 一方, 各ユーザは k 番目のクラスタに対するメンバシップ, つまりクラス事後確率 $P(k|D^u)$ を有する. このメンバシップは, K 次元ベクトルの成分とみなすことができる. このベクトル

$$\boldsymbol{u} = (P(1|D^u), \ldots, P(K|D^u)) \tag{5.31}$$

はユーザ u に対応する方向ベクトルで, $\boldsymbol{u}$ の k 番目の成分はユーザ u が k 番目の典型的なユーザグループに属する確率と考えられる. このグループ化ランキングの混合を考えるとき, 実際にはエントロピー正則化ソフトクラスタリングを用いていることから, 確率 $P(k|D^u)$ を g_{uk} に置き換えることに注意する.

式 (5.30) と (5.31) との対応から, アイテムを $\mathbb{R}^K$ の第一象限にマップし, ユーザを $\mathbb{R}^K$ における半直線 (ユーザ u による “評価軸”) としてマップする. こうしてマップされたアイテムを, さらにユーザ u の評価軸に射影する:

$$I_i \mapsto \theta^u_i, \quad \theta^u_i = \sum_{k=1}^{K} \theta^k_i P(k|D^u). \tag{5.32}$$

ここで, $\theta^u = \{\theta^u_i\}_{i=1}^N$ はこのユーザ u 特有のアイテム選好度パラメタと解釈出来る.

この同時マッピングにより, アイテムとユーザの関連が明らかになる. 類似したユーザに対応する軸は近くに表示され, 類似したアイテムに対応する点は近くに表示される. このような, 好みに関する可視化に関する研究は既に幾つか存在する. 文献 [76] では, 例えば性別, 年齢などのユーザに関する付加情報と, 例えばアクションやホラー, コメディなどのアイテ

ム (ここでは映画) のジャンルに関する付加情報を用いる. 各ジャンルにおいて, 各ユーザに対する映画の選好度が $\mathbb{R}^3$ 空間にマップされる. この研究で用いられている例では, ユーザの年齢と性別が (X, Y) 軸に対応し, 映画の選好度が Z に対応する. こうした3次元への映画のマップが, ジャンル毎に作られる. 一方, 文献 [77] ではユーザとアイテムは同一のユークリッド空間に写像される. そして, ユーザによるアイテムへの評価値は, ある評価関数 $f(||u - I_i||)$ の値であるとみなす. この評価関数は, ユークリッド空間におけるユーザ u とアイテム I_i の距離の関数である. 評価関数は学習用の評価値データから学習し, アイテムとユーザをユークリッド空間に埋め込むのに用いる. つまり, この評価関数の値と実際の評価値との差が最小になるような配置でアイテム, ユーザが埋め込まれる. こうした先行研究と本研究の最も顕著な違いは, アイテムとユーザが埋め込まれる空間が異なるという点である. 本研究におけるアイテム, ユーザ可視化手法では, アイテムは $\mathbb{R}^K$ に埋め込まれる一方で, ユーザは式 (5.32) で定義されるアイテム集合から実数値への射影がなす空間に写像される. ユーザが写像される射影は, $\mathbb{R}^K$ において半直線として表現される. 映画の評価データなど, ある種のデータについては, アイテムはユーザによって評価される対象として存在しており, ユーザとアイテムを全く同一の空間のオブジェクトとして扱うことは適切でないことがある. こうしたデータに対しては, アイテムを点で, それを評価するユーザは評価軸という意味で線で表示するのは自然である.

5.5.2 協調フィルタリング

協調フィルタリングシステムは, アイテムの評価履歴に基づきユーザ同士の類似度を定義し, 類似度の高いユーザが高評価をしているアイテムを, そのアイテムをまだ評価していないユーザに推薦するシステムである. つまり アイテム I_i はユーザ u にまだ評価されていないとして, u が既に評価したアイテムへの評価値を用いて I_i への評価値 r_i を推定する問題である. ここで推定の良し悪しを決定づけるのは, アイテムの評価履歴という形式で表現されるユーザデータ同士の類似度である. 協調フィルタリングを含む推薦システムについては [78] が詳しい. 協調フィルタリングに関する研究は非常に多く行われており, 提案されている多くの手法には適用される局面やデータに依存して向き不向きがある. また, 各手法を実際のシステムに組み込む際には, その状況に応じた作り込みが不可欠である.

数多くある協調フィルタリング手法の網羅的な比較は現実的ではないため, 本論文では古典的な協調フィルタリング手法とその手法で用いられる代表的な類似度の尺度を 2 種類用い, 提案するグループ化ランキングモデルに基づく協調フィルタリングと比較する.

通常, 協調フィルタリングは注目するユーザ u によるアイテム I_i への評価値 $r_{u,i}$ を予測する問題として定式化される. 多くの協調フィルタリングシステムで, 評価値の予測のために次式が採用されている:

$$r_{u,i} = \bar{r}_u + \frac{1}{C} \sum_{v \in ne(i)} \mathrm{sim}(u,v)(r_{v,i} - \bar{r}_v). \tag{5.33}$$

ここで, $\bar{r}_u$ は注目するユーザ u が今まで評価したアイテムへの評価値の平均であり, 近傍 $ne(i)$ はアイテム I_i を既に評価した他のユーザの集合, C は $C = \sum_{v \in ne(i)} |\mathrm{sim}(u,v)|$ で定義される正規化因子である. この予測式による予測結果は, ユーザ同士の類似度の計算方法に大きく依存する. 最も有名な協調フィルタリングシステムは GroupLens [79] と呼ばれるシステムであり, GroupLens では Pearson の相関係数を類似度として採用している:

$$\mathrm{sim}_{Pe}(u,v) = \frac{\sum_{i \in S_{u,v}} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in S_{u,v}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in S_{u,v}} (r_{v,i} - \bar{r}_v)^2}}. \tag{5.34}$$

ここで, $S_{u,v}$ はユーザ u, v の両方が評価したアイテム集合である.

一方, 次で定義されるコサイン類似度の方が Peason 相関係数による推定よりもよい結果を与えるデータも報告されている:

$$\mathrm{sim}_{cos}(u,v) = \frac{\sum_{i \in S_{u,v}} r_{u,i} r_{v,i}}{\sqrt{\sum_{i \in S_{u,v}} r_{u,i}^2} \sqrt{\sum_{i \in S_{u,v}} r_{v,i}^2}}. \tag{5.35}$$

グループ化ランキングモデルの混合モデルに基づき, 幾つかのユーザ間類似度を考案し, 準備的な実験を行った. その結果, 正規化 Fisher カーネルを類似度として用いると, 従来の Peason 相関係数やコサイン類似度に基づく手法と同等の推薦精度が得られることが分かった. 提案する類似度の計算には, 近似したグループ化ランキングモデル $\tilde{P}(D^u; \theta^u)$ に, ユーザ u の選好度パラメタ $\theta^u = (\theta^u_1, \ldots, \theta^u_N)$ を代入したものを用いる. ここで, このユーザ u のパラメタは混合モデルの推定で得られるパラメタ $\{\theta^k_i\}$ を用いて

$$\theta^u_i = \sum_{k=1}^{K} \theta^k_i P(k|D^u) \sim \sum_{k=1}^{K} \theta^k_i g_{uk}, \quad i = 1, \ldots, N$$

で計算できる．こうして各ユーザに関する生成モデルが得られたので, ユーザ間の類似度をこの生成モデルに基づき定めることができる．第3章で述べたように, Fisher カーネル [48] によって, 背後にあるデータ生成モデルに基づいて類似度を定めることが出来る．特徴ベクトルとしてモデルのスコアベクトル $\boldsymbol{s}_u = \partial_\theta \log \tilde{P}(D^u; \theta^u)$ を用い [80], この Fisher スコアを正規化したものとしてユーザ間の類似度を

$$\mathrm{sim}_{NF}(u, v) = \frac{\boldsymbol{s}_u^T \boldsymbol{s}_v}{||\boldsymbol{s}_u|| \cdot ||\boldsymbol{s}_v||} \tag{5.36}$$

で定める [40]．本来 Fisher カーネルでは Fisher 情報行列を用いてスコアの内積を計算するが, 実用上よく行われているように Fisher 情報行列の代わりに単位行列を用いた．

5.6 実験による評価

本節では, 本章で提案したモデルのパラメタ推定手法の妥当性と計算効率性を人工データを用いて評価する．また, グループ化ランキングモデルの混合モデルの応用として提案したアイテム, ユーザの可視化について実データを用いた例を示し, 協調フィルタリングについては既存手法との比較を行う．

5.6.1 グループ化ランキングモデルのパラメタ推定実験

本小節では, 提案したパラメタ推定アルゴリズムの妥当性を示すために, 人工データを用いた簡単な実験の結果を示す．

提案アルゴリズムが正確な尤度関数を増加させることの検証

まず, 近似尤度の最大化により, 本来の尤度関数が最大化されることを確認する．ここではグループ数を $M = 3$ とし, アイテム数は $N = 7$ とする[1]．ユーザ数は $U = 100$ とし, Plackett-Luce モデルに従い完全なランキングデータを生成した後に, その順序を変えることなくランダムに3つのグループに分割してグループを作成した．この Plackett-Luce モデルにおいて, アイテム選好度パラメタは 100 通りランダムに生成した．

[1]効率的に正確な尤度関数を計算するために, アイテム数は少なく設定した

提案アルゴリズムによってパラメタ θ を推定し, $L(\theta)$ と $\tilde{L}(\theta)$ の値を 100 回計算した. 図 5.4(左) が, 正確な尤度関数と近似尤度関数の値の平均値を縦軸に, アルゴリズムの繰り返し回数を横軸にとったグラフである. なお, $L(\theta)$ と $\tilde{L}(\theta)$ の値を同じグラフに示すために, $\tilde{L}(\theta)$ には定数を加えている. また, 同じ図に真のパラメタと推定されたパラメタの間の KL ダイバージェンスも示した. 図 5.4 から, $\tilde{L}(\theta)$ の増加に伴い $L(\theta)$ も増加することが見て取れる. また, 提案アルゴリズムにより推定パラメタは真のパラメタに単調に近づくことも分かる.

尤度の近似の妥当性の検証

次に, 正確な尤度関数 $l(\theta; m, u)$ の値と近似尤度関数 $\tilde{l}(\theta; m, u)$ の値が近くなるのはどのような状況であるかを調べる. 近似尤度関数 $\tilde{l}(\theta; m, u)$ を導出するために, 二つの不等式を用いた. 初めの不等式は, 式 (5.4) の分母に現れる項 $\sum_{j<i} \theta_{\pi^u_m(j)}$ を省略することで得られた. グループのサイズ γ^u_m が十分小さければそのグループ内で取りうる順列の数も小さく, 項 $\sum_{j<i} \theta_{\pi^u_m(j)}$ を省略する効果は小さいと考えられる. 二つ目の不等式は相加・相乗平均の不等式であり, 等号が成立するための必要十分条件はグループ G^u_m 内の全ての $\theta_{\pi^u_m(i)}$ が同じ値をとることである. グループ内の全ての $\theta_{\pi^u_m(i)}$ が同じ値をとるということは現実的には考えにくい. しかし, グループサイズ γ^u_m が小さければ小さいほど, この状況が近似的に成立する可能性は高くなる. M が大きければ, グループの殆どは少数のアイテムしか含まないことになり, その結果相加・相乗平均の不等式は多くの場合等号に近くなることが期待される. この議論の正当性を確認するため, グループの数を変えて生成したデータとパラメタを用いて $|L(\theta)/\tilde{L}(\theta)|$ の値を計算する. ここで, $L(\theta)$ と $\tilde{L}(\theta)$ は常に負値なので, $L(\theta) \geq \tilde{L}(\theta)$ ならば $|L(\theta)/\tilde{L}(\theta)| \leq 1$ が成立する. アイテム数を $N = 7$, ユーザ数を $U = 100$ として, グループ数 M を 2 から 7 まで変えて実験を行う. 各 M の値に対して, 比 $|L(\theta)/\tilde{L}(\theta)|$ を, ランダムに選んだパラメタとそのパラメタに基づいて生成した観測データを用いて 100 回計算した. 図 5.4(右) に, 比 $|L(\theta)/\tilde{L}(\theta)|$ の平均をエラーバー付きで示した. この図から, $L(\theta)$ は $\tilde{L}(\theta)$ によって下から抑えられており, 比はグループ数の増加に従い 1 に近づくことが分かる. 従って, グループに含まれているアイテム数が比較的少数の時には, 近似尤度 $\tilde{L}(\theta)$ の最大化によりパラメタ θ のよい推定値が得られると考えられる.

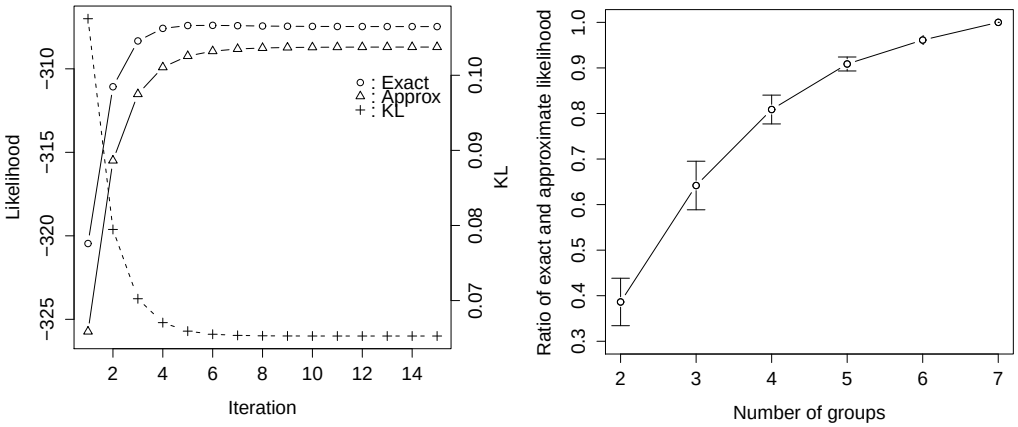

図 5.4: 左: 厳密な尤度と近似尤度の平均値, 及び真のパラメタと推定パラメタの KL ダイバージェンスの平均値. 右: グループの数を変えて計算した尤度の比 $|L(\theta)/\tilde{L}(\theta)|$ の平均と, 標準偏差 (エラーバー表示).

上述の議論は, 近似尤度が厳密な尤度に近づく理想的な状況を論じたものである. 付録 5 では, 厳密な尤度と近似尤度の差を解析的なアプローチで調べる.

計算コストに関する考察

最後に, アイテム数 N あるいはグループ数 M と, 提案手法の計算コストに関する実験を行う. なお, ユーザ数 U の影響は, 基本的には解くべき最適化問題 (5.10) の数と, em アルゴリズムにおける射影を計算する数が線型に増加するのみであることは明らかである. 従って, この実験ではユーザ数は $U = 100$ で固定し, N と M のみを考える.

まず, 提案手法によるパラメタ推定に要する計算時間を計測した. 図 5.5 にその平均時間 (秒) を示す. なお, 標準偏差が平均時間のオーダーに比べて非常に小さかったため, このグラフではエラーバーは省略した. 提案アルゴリズムはプログラミング環境 R(バージョン 2.9.1 [81]) で実装し, Intel プロセッサマシン[2] で実行した. システム I/O に要した時間は除いてある. アルゴリズムの初期化の部分では U 個の最適化問題 (5.10) を解く必要がある. この最適化は, 準ニュートン法 (BFGS 法) を用いて行った. 図 5.5

[2] 2.4 GHz デュアルコアプロセッサ, 主メモリ 4,096MB, Mac OS X version 10.5.8.

から, 提案アルゴリズムはアイテム数とグループ数の増加に伴いほぼ線型オーダーで計算量が増加することが分かる.

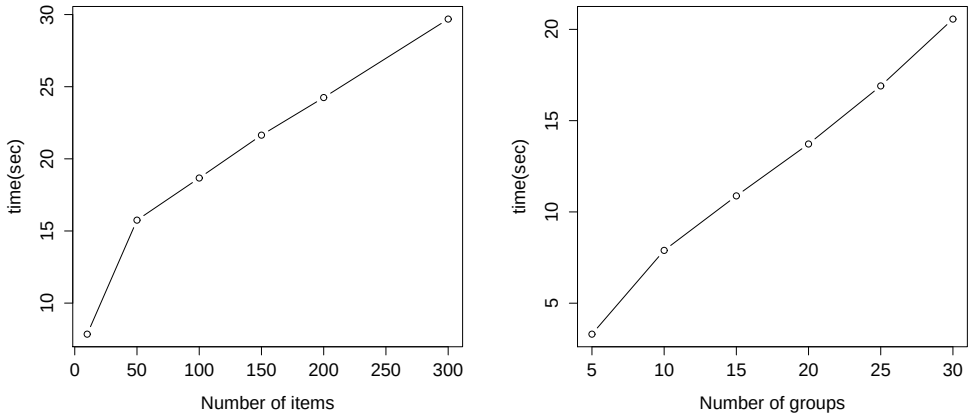

図 5.5: 平均 CPU 消費時間. アイテム数とグループ数を変えて実験した結果.

5.6.2 アイテム, ユーザ可視化実験

本章の第 5.5.1 節で提案した可視化手法を, MovieLens データセット [82] に実際に適用する. このデータセットは 100,000 個の評価データを含む. 評価値は 1 から 5 の 5 段階であり, 評価対象となる映画は 1,682 本, 評価をしているユーザ数は 943 人である. ここでは, 可視化の効果の実証が目的であるため, もとのデータの一部を用いて評価する. まず, 頻繁に評価されている $N=100$ 本の映画を取り出す. 次に, その 100 本の映画のうち 20 本以上の映画に評価を与えているユーザを取り出す. こうしたユーザは, $U=554$ 人であった. 混合モデル (5.22) の混合数は $K=2$ として, モデル

$$P(x) = \omega P(x;\theta^1) + (1-\omega)P(x;\theta^2),$$

を考え, パラメタ $\{\omega, \theta^1, \theta^2\}$ は図 5.3 に示したアルゴリズムで推定した. 図 5.6 はアイテム可視化の例である. 2 次元空間に映画が点として表示してある. この空間において映画 I_i が表現されている点に対応する座標は, θ_i^1 と θ_i^2 である. 100 本の映画タイトルのうち, $(\theta_i^1)^2 + (\theta_i^2)^2$ の大きい上位 5 本の映画にはタイトルも付記した. この可視化手法を用いることで, 例

えば映画 “Star Wars” と “Titanic” は多くのユーザに好まれるが, これらの映画を好むユーザの集団は異なるということが分かる.

図 5.7 は, ユーザを 2 次元空間における半直線として示した例である. 各ユーザは $\boldsymbol{u} = (P(1|D^u), P(2|D^u))$ で定義される方向ベクトルとしてこの空間に表現されている. この半直線はユーザの評価基準を表しているとみなせる. この図からは, ユーザ 4 とユーザ 6 の興味は似通っており, ユーザ 1 の興味とは大きく異なることが読み取れる.

次に, アイテムとユーザ両方を可視化した例を図 5.8 に示す. ユーザ 114 によって既に評価されている映画は丸で, 評価されていない映画は菱形でプロットされている. 式 (5.32) で定義される射影を用いることで, アイテムはユーザを表す評価軸に射影される. この軸上で原点から遠ければ遠いほど, このユーザによる評価値の高いアイテムであるということになる. 図 5.8 から, このユーザ 114 は未評価の映画 “The Shawshank Redemption” は好むが, “Alien” は好まないことが予想される. この可視

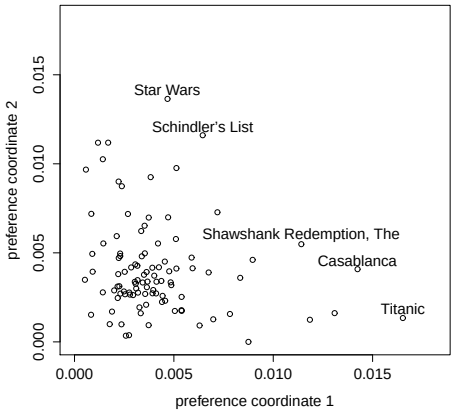

図 5.6: アイテムマッピングの例. アイテムが各混合成分におけるアイテム選好度パラメタの値を座標とする 2 次元空間にマッピングされている. 高い選好度パラメタを有する映画はタイトルもあわせて表示してある.

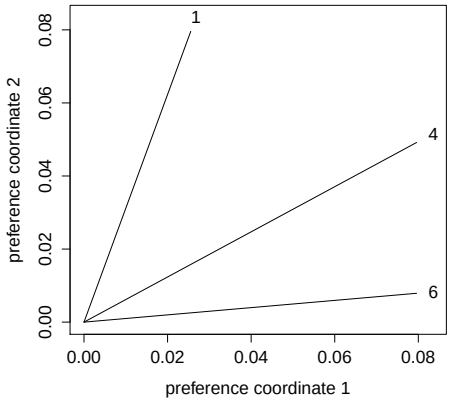

図 5.7: ユーザマッピングの例. 式 (5.31) により, ユーザは 2 次元空間における半直線 (評価軸) として表現される.

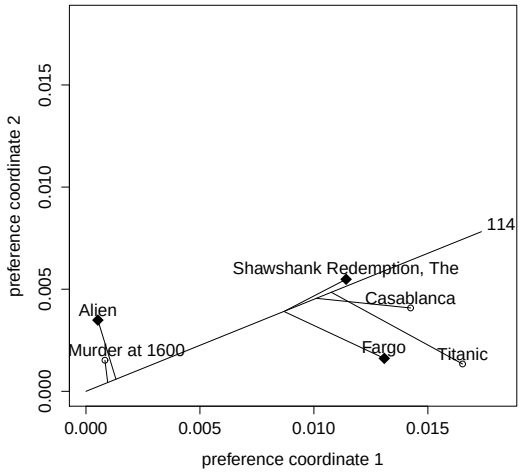

図 5.8: 2 次元空間へのユーザ/アイテムマッピングの例. 114 番のユーザによって既に評価されているアイテムは丸で, 未評価のアイテムは菱形で表示されている. ユーザ 114 にとってのアイテム選好度パラメタは, アイテムを射影 (5.32) によってこのユーザを表す軸に射影したものである.

化手法により, アイテム同士の関係, ユーザ同士の関係, そしてアイテムとユーザとの関係を見てとることができるようになる.

5.6.3 協調フィルタリング実験

実データを用いた協調フィルタリングの実験を行い, 提案する類似度が既存手法と同等の精度を与えることを示す. データとしては, MovieLens データセットと BookCrossing データセット [83] を用いる. GroupLens データセットは, 5-fold のクロスバリデーション用に分割した形で提供されている. つまり, 全体の 80%を学習用部分セット (80,000 の評価データ), 残りの 20%をテスト用データセット (20,000 の評価データ) としたデータが 5 組与えられている. 一方 BookCrossing データセットは, 278,858 人のユーザによる 271,379 冊の書籍に対する 1,149,780 の評価データから構成されている. 非常にサイズが大きいデータなので, 8 冊以上の書籍に評価を与えたユーザのみを取り出す. そして, 12 人以上のユーザに評価された書籍のみを取り出す. これにより, 残ったデータは 663 人のユーザによ

る 1,191 冊の書籍に対する 48,484 の評価データとなる．この残ったデータを, GroupLens データセットと同様に 5-fold クロスバリデーション向けに分割した．

これらのデータセットを用いて, Pearson 相関係数による類似度 (5.34), コサイン類似度 (5.35) 及び正規化 Fisher カーネルによる類似度 (5.36) を用いた評価値予測式 (5.33) による予測精度の評価実験を行った．グループ化ランキングモデルの混合モデルにおける混合数は, 本来ならばクロスバリデーションなどによって定めるべきであるが, ここでは簡単のために 5 で固定した．推薦精度の評価方法としては, 絶対誤差の平均 (Mean Absolute Error:MAE) を用いた．これは次式で計算される:

$$MAE = \frac{1}{|T|} \sum_{(u,i) \in T} |r^*_{u,i} - r_{u,i}|.$$

ここで, T はテスト用データ集合であり, $r^*_{u,i}$ はユーザ u によるアイテム I_i への真の評価値である．表 5.1 に, MovieLens と BookCrossing データセットに対する協調フィルタリングによる推薦精度の値 (平均及び標準偏差) を示す．この結果から, MovieLens データに対しては Pearson の相関係数に

表 5.1: 協調フィルタリングの精度

	Pearson	cosine	Fisher cosine
MovieLens	$\mathbf{0.712} \pm 0.0069$	0.720 ± 0.0055	0.721 ± 0.0056
BookCrossing	1.283 ± 0.0145	1.254 ± 0.0151	$\mathbf{1.253} \pm 0.0133$

基づく類似度を用いることで最良の結果が得られるが, BookCrossing データに対してはグループ化ランキングモデルに基づく正規化 Fisher カーネル類似度を用いた場合が最良の結果を与えることが分かる．以上より, 提案する類似度は評価値の推定精度の観点から, 既存手法と同等の性能を示すことが分かった．

5.7　ランキングデータ生成モデルに基づく計量の学習に関するまとめ

本章では, グループ化ランキング観測データに対する確率モデルを提案した．データに対して適切な生成モデルを定義することで, グループ化ラ

ンキング観測データのような特殊なデータに対しても本論文のテーマである距離構造の情報論的観点からの学習を行うことができる.

本研究で扱ったような, タイ (同点) が存在する状況でのランキングデータから, アイテムがある順序に並べられる確率を求める研究も存在する. 文献 [63] では, Mallows による順列の生起確率を代数学的手法を用いて部分ランキングに拡張し, そのパラメタの推定を行う方法を与えている. 本論文で扱ったような選好度パラメタに基づくモデルと, Mallows のモデルのようなランキングそのものに確率を与えるモデルで, どちらがどのような状況に適しているのかの検討は興味深い課題である.

本論文では, グループ化ランキングの混合モデルを考え, 二つの応用を提案した. ひとつはデータの可視化であり, もうひとつは協調フィルタリングである. 先行研究として, ユーザの多様性を表現するために, ランキングモデルの混合モデルを考えたものは幾つかある. 文献 [84] では Bradley-Terry モデルの混合が用いられている. 一方, 文献 [60] と [64] では Mallow のモデルのような距離ベースのランキングモデルの混合モデルが提案されている. また, 文献 [85] でも順序データの効率的なクラスタリング手法が開発されており, 協調フィルタリングへの応用が提案されている. 本論文で提案された混合モデルに基づく可視化などの応用は, アイテムに選好度パラメタを割り当てて評価値を説明するような任意のモデルに適用可能である [14].

本論文で提案した協調フィルタリングによる推薦の精度は, 従来の手法と同程度である. 協調フィルタリングによる推薦精度は問題に応じたチューニングに大きく依存するため, 単純に精度の向上のみを追求するよりも, 推薦と同時に何らかの付加情報を効率的に提示する手法の研究が重要であると考えられる. 本論文で提案したユーザとアイテムの同時可視化結果を推薦と同時に提示することで, 推薦の根拠をユーザにわかりやすく伝えることが可能となる.

第6章　まとめ

本論文では, 情報理論の観点からデータ分布空間における教師付き距離構造学習問題を議論した. 一般的な連続変数ベクトルデータに対する教師付き距離構造学習の枠組みとして, 条件付きエントロピー最小化基準による方法を提案した. また, 特殊な離散観測データに対する距離構造学習のために, データの新しい生成モデルを提案し, モデルに基づきデータ間距離の解析をする方法を議論した.

まず, 条件付きエントロピー最小化基準による学習の枠組みの理論的な正当化を行い, その具体例として線型変換による次元削減手法と, 特徴空間における距離構造を学習する手法である Multiple Kernel Learning(MKL) の手法を新たに提案した. 距離構造の学習問題は多くの研究者の関心を集めているが, 高度な最適化理論を援用した手法が数多く開発される一方で, データの分布まで考慮した情報論的な手法はそれほど多くない. また, 情報論的な観点からの距離構造の学習手法の多くは, Shannon の微分エントロピーや相互情報量を推定することが困難であることから, 比較的推定が容易な Renyi エントロピーで代用したものが多かった. 本論文ではエントロピーの推定に k 近傍法に基づく効率的な手法を用い, さらに同時エントロピーを周辺エントロピーの和で近似することで, 従来研究では避けられてきた Shannon の微分エントロピーを用いて理論的に妥当な枠組みを構築した. エントロピーの推定に k 近傍法を用いたもう一つの効果として, データの局所性がエントロピー推定を介して反映されることが挙げられる. これにより, 本論文で提案した条件付きエントロピー最小化に基づく次元削減手法は, 判別問題の前処理として用いたときにデータの局所性を陽に用いた既存の高精度な手法と同等の性能を示す. さらに, 提案する枠組みから, 情報論的な MKL 手法を提案した. これまで提案されきた MKL 手法の多くは, SVM のマージン最大化基準に基づく最適化問題として定式化されたものであった. 本論文で提案した情報論的な MKL 手法は, データの確率分布を無視した既存の MKL 手法と比較して情報理論的に明確な背景を持つ手法である. 実データに対する実験の結果から,

既存の MKL 手法に比べ幾つかの問題に対して優れた性能を有することも示された.

次に, 特殊な, しかし近年その重要性を増している離散観測データの空間における距離構造の学習問題を議論した. 種々のデータに対して, そのデータが分布する空間の距離構造を学習することは判別や可視化などの性能向上のために重要である. しかし, アイテムの購買履歴や, ユーザがアイテムに対して離散値で評価を与えたデータのような非標準的なデータに対しては, データ間の距離や類似度としてとりあえずユークリッド距離を用いるといったアプローチすら困難である. 一方, データの生成モデルが既知であれば, 例えば Fisher カーネルのような方法でデータ間に類似度を定義することが可能である. そこで, 複数のユーザによる複数のアイテムへの評価データを本論文ではグループ化ランキング観測データと定義し, その生成モデルを構築することを通して, アイテム間の関係やユーザ間の関係などが直観的に理解出来るようなデータ解析手法を提案した. そして, 正規化 Fisher カーネルを用いてデータ間に類似度を定義し, 協調フィルタリングシステムとしての応用を検討した. 実データを用いた評価実験の結果, 提案するモデルに基づいてユーザ間に定義した類似度を用いることで, 従来の協調フィルタリングと同等の推薦精度が達成できた. 本論文で提案する協調フィルタリング手法はアイテムとユーザの可視化と併用することで, 推薦の根拠を明示することが可能であるという付加価値を有する.

今後考える必要がある課題を，以下にいくつか挙げる.

条件付きエントロピー最小化基準に基づく MKL アルゴリズムにおいて, 本論文では, 要素カーネル関数の凸結合によって得られるカーネル関数の最適化問題のみを考察したが, 第 3 章において示したように, 積やテンソル積など他の幾つかの演算に対してもカーネル関数は閉じている. 例えば, [56] はカーネル関数の結合係数が $\boldsymbol{x}$ に依存する形の結合の最適化問題を考えた. また, [86] ではカーネル関数値の多項式の形の組み合わせ最適化を考えている. 本論文で示した MCEM アルゴリズムは, 凸結合以外の方法によるカーネル関数の組み合わせも扱うことが可能であり, 考えるカーネル関数結合のクラスを広げることで, さらに高精度な判別が期待できる.

なお, MCEM アルゴリズムの収束性はまだ明らかになっていない. 実験の結果, MCEM アルゴリズムの繰り返しで必ずしも単調に条件付きエントロピーが減少しないという状況も観測された. 収束するための条件や,

収束の保証されたアルゴリズムへの改良も重要な課題である. また, [9] で行われているように, 最適化問題 (4.11) において判別関数のパラメタ $\boldsymbol{\alpha}$ とカーネル結合の係数 $\boldsymbol{\beta}$ を同時に最適化する手法の開発も興味深い.

生成モデルに基づく距離構造の学習について, 文献 [87] では Plackett-Luce モデルを Bayes 推定の枠組みで扱う試みがなされている. この研究では, Gumbel 分布と呼ばれる分布を Plackett-Luce モデルにおけるアイテム選好度パラメタの事前分布としている. 本論文で提案したグループ化ランキングモデルの Bayes 統計からの考察は興味深い今後の課題の一つである. また, 文献 [88] においても Bayes 推定の枠組みで評価値を予測する方法が提案されており, ユーザとアイテムの可視化方法が提案されている. 推薦の根拠を適切な可視化手法とともに効果的に提示する方法は, 今後の推薦システム研究において重要な位置を占めると考えられ, 今後の重要な研究課題の一つである.

本論文に関わる業績

論文

1. Hideitsu Hino, Yu Fujimoto, Noboru Murata, "A Grouped Ranking Model for Item Preference Parameter," *Neural Computation*, Vol.22, Issue 9, 2010

2. Hideitsu Hino, Noboru Murata, "Conditional Entropy Minimization Criterion for Dimensionality Reduction and Multiple Kernel Learning," *Neural Computation*, (to appear)

国際会議

1. Yu Fujimoto, Hideitsu Hino, Noboru Murata, "An Estimation Method for Bradley-Terry and its Related Models based on the Bregman Divergence," *Learning Workshop 2010 (Computational and Biological Learning Society)*, Utah, United States, April, 2010

2. Yu Fujimoto, Hideitsu Hino, Noboru Murata, "ITEM-USER PREFERENCE MAPPING WITH MIXTURE MODELS -Data Visualization for Item Preference-," *International Conference on Knowledge Discovery and Information Retrieval (KDIR2009)*, Madeira, Portugal, October, 2009.

3. Hideitsu Hino, Noboru Murata, "An Information Theoretic Perspective of the Sparse Coding,", *6-th International Symposium on Neural Networks(ISNN2009),* Wuhan, China, May, 2009.

4. Hideitsu Hino, Yu Fujimoto, Noboru Murata, "Item Preference Parameters from Grouped Ranking Observations," *13-th Pacific-Asia*

Conference on Knowledge Discovery and Data Mining(PAKDD2009), Bangkok, Thailand, April, 2009.

国内会議/研究会

1. 日野英逸, 村田昇, “条件付きエントロピー最小化に基づく教師付き次元削減手法,” 第12回情報論的学習理論ワークショップ (IBIS2009), 福岡, 2009 年 10 月
2. 日野英逸, 藤本悠, 村田昇, “Grouped ranking モデル: Plackett-Luce モデルの一般化とその応用,” 第 11 回情報論的学習理論ワークショップ (IBIS2008), 仙台, 2008 年 10 月

謝辞

統計科学, 機械学習理論の知識をほとんど持たない状態で博士課程に進学した私を快く受け入れて頂き, 3 年間に渡り必要にして十分な指導・示唆をして頂いた村田昇教授に心から感謝致します．また，本論文の審査をして頂いた松本隆教授, 内田健康教授，井上真郷准教授, そして産業技術総合研究所の赤穂昭太郎博士に深く感謝致します.

京都大学在学中には, 大学院情報学研究科 数理工学専攻 数理物理学講座力学系理論分野の先生方及び大学院生諸氏に多大な影響を受けました. 岩井敏洋教授, 上野嘉夫助教授 (現 公立はこだて未来大学教授), 山口義幸助手 (現 助教) には, 卒業論文, 修士論文の指導を通して研究の面白さを教えていただきました. 当時の研究室の先輩, 同輩, 後輩の皆様にも, 研究に関する議論を通して様々なことを学ばせて頂きました. ここに感謝いたします.

日立製作所システム開発研究所在籍時には, 高橋健太研究員, 村上隆夫研究員との議論が非常に有用でした. 特に, 高橋氏には, 企業における要素技術研究の進め方などをご教示頂きました.

早稲田大学の情報学習システム研究室の諸氏には, 機械学習理論の分野の初心者であった私の勉強・研究に付き合って頂きました. 研究を進めていくにあたって，様々な議論に付き合って頂いた本研究室の学生の方々に感謝します．研究環境の整備の面でも多大なお世話になりました.

また, 共同研究を通して私の知見を広げて下さった青山学院大学助教の藤本悠博士, 早稲田大学高等研究所助教の小川哲司博士, ヘルシンキ工科大学の Nima Reyhani 氏, 産業技術総合研究所研究員の藤木淳博士に感謝いたします.

最後に，博士課程への進学を決めた時に, 一度は諦めかけていた私を応援し, 後押しをしてくれた父, 母, 義父, 義母, 兄, 姉には，常に感謝の念を抱いております. そして, いつも変わらず応援し, 支えてくれている妻康恵と, 娘の瑛真に何よりも感謝します.

付録

付録 1: 条件付きエントロピー最小化基準の正当性

第 4 章において Fisher の判別分析との関係から条件付きエントロピー最小化基準による学習を提案したが, ここではまた別な情報論的観点から, 条件付きエントロピー最小化基準の妥当性を示す.

次元削減問題では, 変換されたデータはコンパクトに纏まっていることが望ましい. このコンパクトなデータ表現という考えを情報理論の文脈で考察すると, 次元削減における望ましい変換とは相互情報量が小さい変換ということになる:

$$I(\boldsymbol{X};\boldsymbol{Z}) = H(\boldsymbol{Z}) - H(\boldsymbol{Z}|\boldsymbol{X}). \tag{6.1}$$

これは, $A : \boldsymbol{x} \mapsto A^T\boldsymbol{x} = \boldsymbol{z}$ をデータ圧縮過程とみなしたとき, 相互情報量 $I(\boldsymbol{X};\boldsymbol{Z})$ が小さいということが, データが高い圧縮率で変換されたことを意味するためである [18]. 相互情報量 (6.1) はもとのデータ $\boldsymbol{X}$ の分布と変換されたデータ $\boldsymbol{Z}$ の分布のみによって決定されるため, 式 (6.1) は教師無し次元削減のための基準とみなすことができる. 教師付き次元削減の枠組みにおいては, データは各クラスにおいてコンパクトに分布していることが望ましい. この場合, 次元削減変換の良さを, クラス条件付きの相互情報量

$$I(\boldsymbol{X};\boldsymbol{Z}|Y) = H(\boldsymbol{Z}|Y) - H(\boldsymbol{Z}|\boldsymbol{X},Y)$$

によって評価することは自然である. また, 多くの場合変換 $\boldsymbol{x} \mapsto \boldsymbol{z}$ は非確率的であり, その場合には $H(\boldsymbol{Z}|\boldsymbol{X},Y)$ は 0 であり, 変換の良さは本質的にはクラス条件付きエントロピー $H(\boldsymbol{Z}|Y)$ で評価される. この議論より, 提案した次元削減の基準はデータ圧縮理論の枠組みにおいても妥当であると主張できる.

次に, 確率変数 $\boldsymbol{X}$ のネゲントロピーとクラス条件付きネゲントロピー

を考える [20]. これらは

$$J(\boldsymbol{Z}) = H_G(\boldsymbol{Z}) - H(\boldsymbol{Z}), \tag{6.2}$$
$$J(\boldsymbol{Z}|Y) = H_G(\boldsymbol{Z}|Y) - H(\boldsymbol{Z}|Y) \tag{6.3}$$

で定義される量であり, このネゲントロピーの表現から

$$\begin{aligned} H(\boldsymbol{Z}) - H(\boldsymbol{Z}|Y) &= \{H_G(A^T\boldsymbol{X}) - H_G(A^T\boldsymbol{X}|Y)\} - \{J(A^T\boldsymbol{X}) - J(A^T\boldsymbol{X}|Y)\} \\ &= \frac{1}{2}\log\frac{|A^T\Sigma A|}{\prod_{y=1}^{C}|A^T\Sigma_y A|^{p(y)}} - \{J(A^T\boldsymbol{X}) - J(A^T\boldsymbol{X}|Y)\} \end{aligned}$$

を得る. ここで Σ と Σ_y はそれぞれ全データ D とクラス y に属するデータ D_y の共分散行列であり, $p(y)$ はクラス事前確率である. ここで, 条件付きエントロピー $H(\boldsymbol{Z}|Y)$ は以下のように 3 つの項に分解できる:

$$\begin{aligned} H(\boldsymbol{Z}|Y) &= H(A^T\boldsymbol{X}|Y) \\ &= H(A^T\boldsymbol{X}) + \{J(A^T\boldsymbol{X}) - J(A^T\boldsymbol{X}|Y)\} - \frac{1}{2}\log\frac{|A^T\Sigma A|}{\prod_{y=1}^{C}|A^T\Sigma_y A|^{p(y)}} \\ &= H_G(A^T\boldsymbol{X}) - J(A^T\boldsymbol{X}|Y) - \frac{1}{2}\log\frac{|A^T\Sigma A|}{\prod_{y=1}^{C}|A^T\Sigma_y A|^{p(y)}}. \end{aligned} \tag{6.4}$$

以下, これら 3 項の意味を考察する.

正規分布のエントロピー項

第 1 項 $H_G(A^T\boldsymbol{X})$ は, 全てのデータが正規分布に従うとした場合のエントロピーを表す. この項の値は共分散行列 Σ によって完全に定まり, この項を最小化することは全てのデータが小さく纏まって分布するように変換することに相当する. しかし, この項はデータのスカラー倍によっていくらでも小さくなるため, 判別性には影響しない項である. 変換 A に適当な正則化が行われるという仮定の下で, 第 1 項 $H_G(A^T\boldsymbol{X})$ の最小化は判別性能には寄与しないといえる.

条件付きネゲントロピー項

式 (6.2) で定義されるように, ネゲントロピーは一種の正規化されたエントロピーであり, 独立成分分析の分野で非正規性の尺度としてよく利用される [89; 90; 20].

式 (6.3) より, $H_G(\boldsymbol{Z}|Y)$ の影響もあるため厳密な議論ではないが, 条件付きエントロピー $H(A^T\boldsymbol{X}|Y)$ の最小化により条件付きネゲントロピー $J(A^T\boldsymbol{X}|Y)$

は最大化されることがわかる. つまり, 条件付きエントロピー最小化により, 変換後のデータの非正規性が最大化され, データは正規分布とかけ離れた特徴的な分布をするように変換されると考えられる.

不均一分布を仮定した判別分析項

式 (6.4) の第 3 項は, 不均一分布を仮定した判別分析 (Heteroscedastic Discriminant Analysis:HDA [91]) における目的関数と同一である. この項の最適化により, クラス判別性を高めることになる. FDA では各クラスに属するデータは同一の共分散構造を持つ正規分布に従うという仮定をおいていた. HDA はこの仮定を取り除くものであり, 多くの研究がなされている [91; 92; 93; 94].

付録 2: 線型次元削減に関するより詳細な実験

第 4 章において行った実験に加え, 提案する線型次元削減手法 LCEM についてより詳しい実験結果を示す. まず, 第 4 章で示した実験と同じデータ, 同じ手法について, 次元を 1 次元に削減したときの判別精度を表 6.1 に示す. この表 6.1 から, 多くのデータに対して LCEM アルゴリズムは他の手法と同等か少し良い判別結果を得られることがわかる.

次に, 削減する次元を変えたときの判別精度を, 削減された次元の関数として図 6.1 に示す. この結果から, 全体的に LCEM アルゴリズムは良好な判別結果を与えるが, 全てのデータセットについて一貫して他の手法よりも良いような手法な存在しないことがわかる. 図 6.1 から, 次元が上がるにつれた誤り率はおおむね低減することがわかる. しかし, “ringnorm” データでは 7 次元付近で最良の判別結果が得られている. これは, 最適な次元を求めるために何らかのモデル選択手法が必要であることを示唆している.

付録 3: MCEM.Q アルゴリズムにおけるカーネル結合最適化手法の導出

MCEM.Q アルゴリズムにおける $\boldsymbol{\beta}$ の最適化の方法を導出する. 式 (4.12), (4.13) のように, 条件付きエントロピーとその上界に着目する. 式 (4.13) の最右辺は, KFDA の目的関数と同じであった. KFDA ではこの条件付

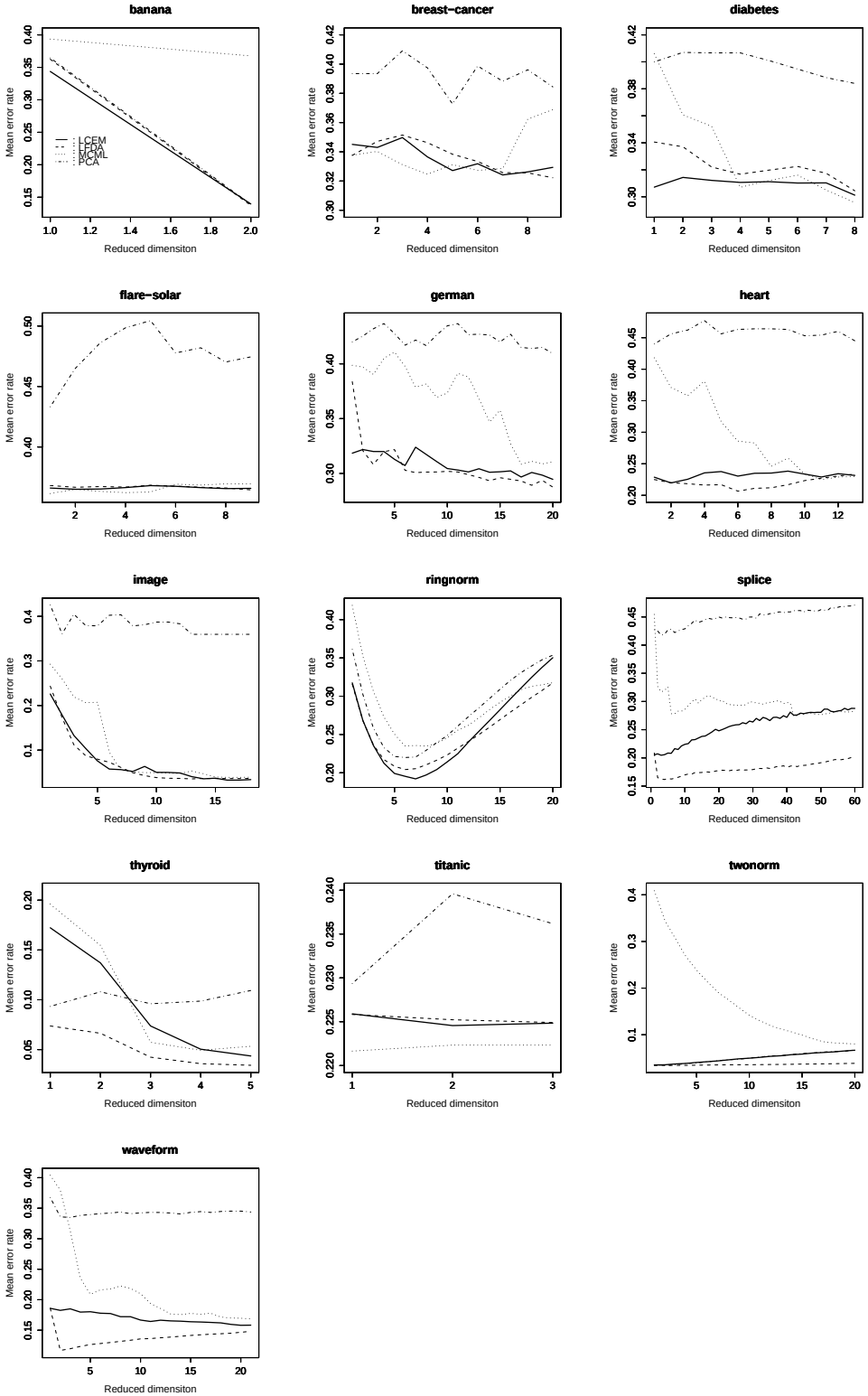

図 6.1: 削減された次元の関数としての, 平均誤り判別率. 4 種類の次元削減手法を用いて, データを低次元空間に射影された上で最近傍法で判別した.

表 6.1: 線型次元削減手法により 1 次元のデータ次元を削減した上で判別した結果. 判別誤り率の平均 (パーセント表記) と分散. 最も精度の良い結果と, 5%の有意水準の t テストによって同等と検定された結果は太字で表示してある.

Data name	PCA	FDA	MCML	LFDA	LCEM
banana	36.5(0.6)	38.3(4.0)	39.4(1.3)	36.2(1.2)	**34.4**(1.6)
breast-cancer	38.9(5.5)	34.9(5.1)	**33.8**(5.4)	**33.9**(4.7)	34.5(4.8)
diabetes	40.0(4.2)	31.3(2.8)	40.6(2.2)	34.1(2.4)	**30.7**(2.5)
flare-solar	43.8(5.7)	**36.4**(1.9)	**36.2**(2.6)	36.8(1.9)	36.6(2.0)
german	42.0(2.3)	**32.0**(2.6)	39.9(3.3)	38.4(3.3)	**31.8**(2.8)
heart	44.0(25.9)	**22.9**(4.1)	41.8(5.6)	**22.5**(3.2)	**22.9**(3.2)
image	44.0(9.0)	**22.1**(0.9)	29.3(1.5)	31.2(1.6)	22.6(1.4)
ringnorm	36.1(8.4)	**31.7**(1.0)	41.9(0.9)	**31.6**(1.6)	**31.9**(1.1)
splice	42.7(4.3)	**20.4**(0.8)	45.4(2.0)	**20.9**(0.9)	**20.6**(0.6)
thyroid	9.3(3.8)	17.9(4.9)	19.6(3.3)	**7.4**(3.4)	17.2(4.2)
titanic	23.0(1.6)	22.5(1.1)	**22.2**(1.0)	22.6(1.5)	22.5(1.0)
twonorm	3.6(0.3)	**3.5**(0.5)	40.9(1.2)	**3.4**(0.4)	**3.5**(0.5)
waveform	36.8(19.0)	**18.6**(1.2)	40.4(1.2)	**18.6**(1.1)	**18.7**(1.1)

きエントロピーの上界を $\boldsymbol{\alpha}$ に関して最小化した. カーネルの凸結合の係数 $\boldsymbol{\beta}$ についても同様に, 条件付きエントロピーの上界を $\boldsymbol{\beta}$ の二次形式の形で表現することができる.

まず, $V_y(\boldsymbol{\beta})$ を要素カーネルを陽に用いて表す. $K(s)$ で s 番目の要素カーネル関数 $k(\cdot\,,\cdot\,;\lambda_s)$ のグラム行列を表し, $\boldsymbol{k}_i(s)$ を $K(s)$ の第 i 行ベクトルとする. このとき, $V_y(\boldsymbol{\beta})$ は

$$V_y(\boldsymbol{\beta}) = \frac{1}{N_y}\sum_{i\in D_y}(\boldsymbol{k}_i-\bar{\boldsymbol{k}}^y)(\boldsymbol{k}_i-\bar{\boldsymbol{k}}^y)^T, \tag{6.5}$$

$$\boldsymbol{k}_i = \sum_{s=1}^{S}\beta_s\boldsymbol{k}_i(s),\ \ \bar{\boldsymbol{k}}^y=\sum_{s=1}^{S}\beta_s\bar{\boldsymbol{k}}^y(s) \tag{6.6}$$

のように書ける. 今, 定数項と定数倍要素を無視すると, 条件付きエント

ロピーの上界は次式で得られる:

$$
\begin{aligned}
&\sum_{y=1}^{C} \frac{N_y}{N} \log |\boldsymbol{\alpha}^T V_y(\boldsymbol{\beta}) \boldsymbol{\alpha}| \\
&= \sum_{y=1}^{C} \frac{N_y}{N} \log \left| \boldsymbol{\alpha}^T \left\{ \frac{1}{N_y} \sum_{i \in D_y} (\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)(\boldsymbol{k}_i - \bar{\boldsymbol{k}}^y)^T \right\} \boldsymbol{\alpha} \right| \\
&= \sum_{y=1}^{C} \frac{N_y}{N} \log \left| \boldsymbol{\alpha}^T \left\{ \frac{1}{N_y} \sum_{i \in D_y} (\tilde{K}_i - \tilde{K}^y) \boldsymbol{\beta} \boldsymbol{\beta}^T (\tilde{K}_i - \tilde{K}^y)^T \right\} \boldsymbol{\alpha} \right| \\
&= \sum_{y=1}^{C} \frac{N_y}{N} \log \left| \frac{1}{N_y} \sum_{i \in D_y} (\boldsymbol{\gamma}_{iy}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \boldsymbol{\gamma}_{iy}) \right| \\
&= \sum_{y=1}^{C} \frac{N_y}{N} \log \left| \boldsymbol{\beta}^T \left\{ \frac{1}{N_y} \sum_{i \in D_y} (\boldsymbol{\gamma}_{iy} \boldsymbol{\gamma}_{iy}^T) \right\} \boldsymbol{\beta} \right| \\
&= \sum_{y=1}^{C} \frac{N_y}{N} \log \left| \boldsymbol{\beta}^T \Gamma_y \boldsymbol{\beta} \right| \leq \log \left| \boldsymbol{\beta}^T \sum_{y=1}^{C} \frac{N_y}{N} \Gamma_y \boldsymbol{\beta} \right| = \log |\boldsymbol{\beta}^T \Gamma_w \boldsymbol{\beta}|.
\end{aligned}
$$

ただし, S 個のグラム行列の第 i 行ベクトルを並べたものを $\tilde{K}_i = (\boldsymbol{k}_i(1), \ldots, \boldsymbol{k}_i(S)) \in \mathbb{R}^{N \times S}$ として, $\boldsymbol{k}_i = \sum_{s=1}^{S} \beta_s \boldsymbol{k}_i(s) = \tilde{K}_i \boldsymbol{\beta}$ とした. 同様に, クラス y に属するデータの S 個のグラム行列の平均を行に並べたものを $\tilde{K}^y = (\bar{\boldsymbol{k}}^y(1), \ldots, \bar{\boldsymbol{k}}^y(S)) \in \mathbb{R}^{N \times S}$ として, $\bar{\boldsymbol{k}}^y = \sum_{s=1}^{S} \beta_s \bar{\boldsymbol{k}}^y(s) = \tilde{K}^y \boldsymbol{\beta}$ とした. また, $\boldsymbol{\gamma}_{iy} = (\tilde{K}_i - \tilde{K}^y)^T \boldsymbol{\alpha} \in \mathbb{R}^S$, $\Gamma_y = \frac{1}{N_y} \sum_{i \in D_y} \boldsymbol{\gamma}_{iy} \boldsymbol{\gamma}_{iy}^T$ として, $\Gamma_w = \sum_{y=1}^{C} \frac{N_y}{N} \Gamma_y$ と定めた. 上式の最後の不等式は, Jensen の不等式から導いた.

同様にして, 正則化項 $-\eta H(f(X; \boldsymbol{\alpha}, \boldsymbol{\beta}))$ も $\boldsymbol{\beta}$ に関する二次形式で $\boldsymbol{\beta}^T \Gamma_b \boldsymbol{\beta}$ のように近似出来る. ここで, $\Gamma_b = \sum_{y \in \{\pm 1\}} (N_y/N) \Gamma_b^y$ であり, $\Gamma_b^y = (\tilde{K}^y - \tilde{K})^T \boldsymbol{\alpha} \boldsymbol{\alpha}^T (\tilde{K}^y - \tilde{K})$ とした. なお, ここでは簡単のために, 条件付きエントロピーと同様に上界を用いて近似したが, 本来エントロピー項 $H(f(X; \boldsymbol{\alpha}, \boldsymbol{\beta}))$ は下界を評価すべきであることに注意する. 計算が容易な形でエントロピーの下界を $\boldsymbol{\beta}$ の二次形式で評価することは今後の課題とする.

以上より, 条件付きエントロピーの上界の $\boldsymbol{\beta}$ に関する最小化問題は次

の問題として定式化される:

$$\min_{\boldsymbol{\beta}} \quad \boldsymbol{\beta}^T(\Gamma_w - \eta\Gamma_b)\boldsymbol{\beta} \tag{6.7}$$

$$\text{subject to} \quad \sum_{s=1}^{S} \beta_s = 1, \;\; \beta_s \geq 0, \;\; s = 1, \ldots, S. \tag{6.8}$$

この問題は二次計画問題であり, 一意な解が, 例えば内点法などにより効率的に求められる. $\boldsymbol{\beta}$ をこの二次計画問題の解として定める MCEM アルゴリズムを, MCEM.Q アルゴリズムと呼ぶ.

一方, 扱わなければいけない問題のデータ数が非常に多く, 計算効率が要求される場合には, $\boldsymbol{\beta}$ に関する制約を緩和してより簡単な問題を効率的に解くというアプローチも考えられる. この場合は, $\Gamma_w - \eta\Gamma_b$ の最小固有値に対応する固有ベクトルを最小化問題 (4.14) の解に用いる. 固有ベクトルの負の成分は 0 に置き換えることにする. こうして固有値問題によって $\boldsymbol{\beta}$ を定める MCEM アルゴリズムを, MCEM.E アルゴリズムと呼ぶ.

付録 4: グループ化ランキングモデルの詳細

第 5 章で提案したグループ化ランキングモデルの背景にあるデータ生成プロセスを詳述し, 確率モデルとして正しいものであることを示す. まず, ランキングデータの研究においては, 少なくとも 2 種類の" 不完全データ" が存在することに注意する. 一つは, 未評価のアイテムがデータに含まれている状況である. ランキングモデルは, こうした未観測データの問題を解決しなければならない. グループ化ランキングモデルにおけるこうした未観測データの扱いについては付録 7 で述べる. もう一つの不完全データは, 本論文で扱うような, グループ化ランキングの形でしかデータが与えられないという問題である. この問題は, タイ (同点) の問題とも呼ばれている.

グループ化ランキング観測データ D^u はアイテムのグループ $\{G_m^u\}_{m=1}^M$ から構成されている. 各グループ G_m^u に含まれるアイテムの数は, そのデータを生成したユーザの評価の傾向を反映している. 例えば, 厳しい評価をつける傾向があるユーザは, ごく少数のアイテムにしか最高の評価を与えないであろうし, 評価が甘いユーザは多くのアイテムに良い評価を与えるであろう. こうしたユーザの傾向という概念を表現するために, 文献 [63] にならって **composition** の概念を導入する.

定義 6.1 N 個の要素からなる集合を M 個の部分集合に分類するときの *composition* とは, その和が N になるような整数列 $\gamma^u = (\gamma_1^u, \ldots, \gamma_M^u)$ である.

Composition $\gamma^u = (\gamma_1^u, \ldots, \gamma_M^u)$ は, γ_1^u 個のアイテムが初めのグループ G_1^u に含まれ, γ_2^u 個のアイテムが第二のグループ G_2^u に含まれるようなグループ化ランキング観測データに対応する.

ここで, ユーザは次の 2 つのステップに従ってグループ化ランキングデータを生成すると仮定する:

1. N 個のアイテムに完全なランキングを与える:
   ユーザ u によって与えられた完全なランキングデータを $O^u = (I_{u(1)} \succ I_{u(2)} \succ \cdots \succ I_{u(N)})$ と表す. ここで $u(i)$ は i 番目にランクされたデータの添字を表す.

2. N 個のアイテムを M 個のグループに, ランキングの順序は変えずに分割する[1]:
   ユーザ u の composition γ^u に従って分割されたランキングデータを,

   $$(O^u, \gamma^u) = (\underbrace{(I_{u(1)} \succ \cdots \succ I_{u(\gamma_1^u)})}_{\gamma_1^u}, \ldots, \underbrace{(I_{u(\sum_{m=1}^{M-1}\gamma_m^u+1)} \succ \cdots \succ I_{u(N)})}_{\gamma_M^u})$$

   で表す. そして, 組 (O^u, γ^u) をグループ化ランキング観測データ $D^u = \{G_1^u, \ldots, G_M^u\}$ と同一視する.

つまり, 各ユーザは全てのアイテムに対して完全なランキングを与えると仮定するが, 何らかの理由によりランキングはグループ化ランキング観測データ $D^u = \{G_1, \ldots, G_M\}$ として部分的にしか観測されないと考えるのである. 例えばユーザは大雑把な評価をつけるように指示されていたのかもしれないし, 詳細な順序付けを報告する余裕が無かったという場合もありうる. なお, データ O^u は Plackett-Luce モデルで扱う完全ランキングデータそのものである. グループ化ランキングモデルでは, 与えられるデータはグループ化ランキング観測データ $\{D^u = \{G_m^u\}_{m=1}^M\}_{u=1}^U$ であり, 各グループ内の順序は観測出来ない. $N = M$ で各グループが一つのアイテムしか含まないならば, このモデルは Plackett-Luce model と同じである. この意味で, グループ化ランキングモデルは Plackett-Luce モデルの一般化である.

[1] この分割ステップは, 初めのランキングのステップとは独立であると仮定する.

このグループ化ランキングモデルは確かに確率モデルであることを示す. N 個のアイテムをグループ $\{G_m^u\}_{m=1}^M$ に分割するとして, composition γ^u を変えながらその全ての考えうるパターンを列挙したものは, N 個のアイテムの全ての並び順を列挙したものと同じである. 従ってその確率の和は 1 になる. つまり,

$$\begin{aligned}\sum_{\gamma^u}\sum_{\{G_m^u|\gamma^u\}_{m=1}^M} P(\{G_m^u\}_{m=1}^M) &= \sum_{\gamma^u}\sum_{\{G_m^u|\gamma^u\}_{m=1}^M} P((O^u,\gamma^u)) \\ &= \sum_u P\big(I_{u(1)} \succ I_{u(2)} \succ \cdots \succ I_{u(N)}\big) = 1\end{aligned}$$

である. ここで, γ^u に関する和は N アイテムに関する全ての取りうる composition について考える. $\{G_m^u|\gamma^u\}_{m=1}^M$ に関する和は, composition γ^u を固定した上でのグループ内のアイテムの全ての順列について考える. また, Plackett-Luce モデルにおける u に関する和は N 個のアイテムの考えられる $N!$ 通りの順列全てについて考える.

付録 5: 近似尤度の理論的評価

グループ化ランキングモデルの尤度 $l(\theta;m,u)$ とその近似 $\tilde{l}(\theta;m,u)$ の差の上界を導出する. 尤度 $l(\theta;m,u)$ の表式に含まれる $\sum_{\pi_m^u\in\mathcal{S}(G_m^u)}$ を取り除くため, $l(\theta;m,u)$ を最大化する一つの π_m^{u*} を固定して, この置換で計算される $l(\theta;m,u)$ を $l^*(\theta;m,u)$ と記す. 当然, $l^*(\theta;m,u) \geq l(\theta;m,u)$ であ

る. 今, $l(\theta; m, u)$ と $\tilde{l}(\theta; m, u)$ との差は次式で上から抑えられる:

$$
\begin{aligned}
& l(\theta; m, u) - \tilde{l}(\theta; m, u) \leq l^*(\theta; m, u) - \tilde{l}(\theta; m, u) \\
= \;& \log \gamma_m^u! + \sum_{i \in G_m^u} \log \theta_i - \sum_{i=1}^{\gamma_m^u} \log \left(\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^{u*}(j)} \right) \\
& - \left[\gamma_m^u \left\{ \log \Theta_m^u - \log \left(\sum_{n=m}^{M} \Theta_n^u \right) \right\} + \log \gamma_m^u! - \gamma_m^u \log \gamma_m^u \right] \\
= \;& \gamma_m^u \log \gamma_m^u + \log \frac{\prod_{i \in G_m^u} \theta_i}{(\Theta_m^u)^{\gamma_m^u}} + \log \left(\frac{\left(\sum_{n=m}^{M} \Theta_n^u \right)^{\gamma_m^u}}{\prod_{i=1}^{\gamma_m^u} (\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^{u*}(j)})} \right) \\
\leq \;& \gamma_m^u \log \gamma_m^u + \frac{\prod_{i \in G_m^u} \theta_i}{(\Theta_m^u)^{\gamma_m^u}} + \frac{\left(\sum_{n=m}^{M} \Theta_n^u \right)^{\gamma_m^u}}{\prod_{i=1}^{\gamma_m^u} (\sum_{n=m}^{M} \Theta_n^u - \sum_{j<i} \theta_{\pi_m^{u*}(j)})} - 2 \\
\leq \;& \gamma_m^u \log \gamma_m^u + \left(\frac{\sum_{n=m}^{M} \Theta_n^u}{\sum_{n=m+1}^{M} \Theta_n^u} \right)^{\gamma_m^u} - 1.
\end{aligned}
$$

ここで, 不等式 $\log x \leq x - 1, \; (x > 0)$ を用いた. 尤度の差 $l(\theta; m, u) - \tilde{l}(\theta; m, u)$ は, γ_m^u の増加関数によって上から抑えられたことになる.

尤度の近似式 $\tilde{l}(\theta; m, u)$ は本来最尤推定のために尤度関数を計算が容易な形で評価するために導かれたものであり, $\tilde{l}(\theta; m, u)$ は尤度の近似としては厳密なものではない. 実用上は, ユーザは全アイテムのうちのごく一部にしか評価を与えず, 差 $l(\theta; m, u) - \tilde{l}(\theta; m, u)$ はそれほど大きくはならないのが普通である. 従って, em アルゴリズムによる近似尤度の最大化により十分良いパラメタ推定値を得ることが期待できる. 尤度のより厳密な近似は, 今後の重要な課題の一つである.

付録 6: グループ化ランキングモデルのための em アルゴリズムの導出

本論文で対象としている状況では, em アルゴリズムにおける e 射影は各観測部分多様体 $\mathcal{D}_u = \{ \theta \mid \sum_{i \in G_m^u} \theta_i = \hat{\Theta}_m^u \}$ 上の点 $\hat{\theta}^u(t)$ を

$$
\hat{\theta}^u(t) = \arg \min_{\theta \in \mathcal{D}_u} KL(\theta, \theta(t)) \; , \quad u = 1, \ldots, U
$$

として定める手続きである．ここで, $KL(\theta, \theta(t)) = \sum_{i=1}^{N} \theta_i \log \frac{\theta_i}{\theta_i(t)} = \sum_{m=1}^{M} \sum_{i \in G_m^u} \theta_i \log \frac{\theta_i}{\theta_i(t)}$ である．グループ G_m^u の θ_i の和は $\hat{\Theta}_m^u$ になるように制約されているので, あるユーザ u が与えたデータの中の一つのグループ G_m^u について, 最小化問題 $\sum_{i \in G_m^u} \theta_i \log \frac{\theta_i}{\theta_i(t)}$ を考えれば十分である．この最小化問題は, 次式のように定式化出来る:

$$\min_{\theta} \sum_{i \in G_m^u} \theta_i \log \frac{\theta_i}{\theta_i(t)} , \qquad \text{subject to} \qquad \sum_{i \in G_m^u} \theta_i = \hat{\Theta}_m^u, \ \theta_i > 0 \ . \quad (6.9)$$

この問題 (6.9) を解くために, ラグランジュの未定乗数 λ を導入し, ラグランジアン $F(\theta, \lambda)$ を

$$F(\theta, \lambda) = \sum_{i \in G_m^u} \theta_i \log \frac{\theta_i}{\theta_i(t)} + \lambda \left(\hat{\Theta}_m^u - \sum_{i \in G_m^u} \theta_i \right)$$

とする．ラグランジアンを $\theta_i, \ i \in G_m^u$ で微分したものをゼロとおくことで,

$$\frac{\partial F}{\partial \theta_i} = \log \frac{\theta_i}{\theta_i(t)} + 1 - \lambda = 0.$$

ここで, 式 (6.9) の制約から,

$$\lambda = \log \left(\frac{\hat{\Theta}_m^u}{\sum_{i \in G_m^u} \theta_i(t)} \right) + 1,$$

を得る．従って, e 射影による部分多様体上のパラメタ $\hat{\theta}_i^u(t)$ は

$$\hat{\theta}_i^u(t) = \frac{\theta_i(t)}{\sum_{j \in G_{m|i}^u} \theta_j(t)} \hat{\Theta}_{m|i}^u, \quad i = 1, \ldots, N, \ u = 1, \ldots, U$$

で得られる．ここで $G_{m|i}^u$ はアイテム I_i が属するグループであり, $\hat{\Theta}_{m|i}^u$ で対応するグループパラメタを表す.

m 射影では, 確率単体 Δ_{N-1} 上の点 $\theta(t+1)$ で, 各部分多様体 $\{\mathcal{D}_u\}_{u=1}^{U}$ 上の点 $\hat{\theta}^u(t)$ からの KL ダイバージェンスの和を最小化するような点を求

める. つまり,

$$\begin{aligned}
\theta(t+1) &= \arg\min_{\theta} \frac{1}{U}\sum_{u=1}^{U} KL(\hat{\theta}^u(t),\theta) \\
&= \arg\min_{\theta} \frac{1}{U}\sum_{u=1}^{U}\sum_{i=1}^{N}\left(\hat{\theta}_i^u(t)\log\hat{\theta}_i^u(t) - \hat{\theta}_i^u(t)\log\theta_i\right) \\
&= \arg\max_{\theta} \frac{1}{U}\sum_{u=1}^{U}\sum_{i=1}^{N}\hat{\theta}_i^u(t)\log\theta_i \qquad (6.10) \\
&= \arg\max_{\theta} \sum_{i=1}^{N} p_i \log\theta_i \qquad (6.11)
\end{aligned}$$

で定まる点を求める. ここで $p_i = \frac{1}{U}\sum_{u=1}^{U}\hat{\theta}_i^u(t)$ である. e 射影と同様に, ラグランジュの未定乗数 μ を導入してラグランジアン $H(\theta,\mu)$ を

$$H(\theta,\mu) = \sum_{i=1}^{N} p_i\log\theta_i + \mu\left(1 - \sum_{i=1}^{N}\theta_i\right)$$

で定義する. このラグランジアンを $\theta_i, i = 1,\cdots,N$ に関して微分して 0 とおくことにより,

$$\frac{\partial H}{\partial\theta_i} = \frac{p_i}{\theta_i} - \mu = 0$$

を得る. 制約 $\sum_{i=1}^{N}\theta_i = 1$ と, $\sum_{i=1}^{N} p_i = \frac{1}{U}\sum_{u=1}^{U}\sum_{i=1}^{N}\hat{\theta}_i^u = 1$ なる条件を用いると, $\mu = 1$ を得る. 従って, m 射影によるパラメタの更新は次式で計算できる

$$\theta_i(t+1) = p_i = \frac{1}{U}\sum_{u=1}^{U}\hat{\theta}_i^u(t).$$

なお, em アルゴリズムは観測データに過適合することが多い. これを防ぐために, (5.11) に対して正則化項を加えて,

$$L_{em}^{\mathrm{reg}}(\theta) = \sum_{u=1}^{U}\min_{\theta^u\in\mathcal{D}_u} KL(\theta^u,\theta) + \epsilon KL(\theta_{\mathrm{unif}},\theta),$$

なる最適化問題を考える. ここで $\theta_{\mathrm{unif}} = (\frac{1}{N},\ldots,\frac{1}{N})$ である. この場合, em アルゴリズムの m ステップ (5.14) は

$$\theta_i(t+1) = \frac{1}{U+\epsilon}\left(\sum_{u=1}^{U}\hat{\theta}_i^u(t) + \epsilon\frac{1}{N}\right), \quad i = 1,\ldots,N \qquad (6.12)$$

のように修正される. 本論文では, 人工データを用いた実験においては正則化した (6.12) を用いた. このときの最適な ϵ は, データセットによって一般には異なり, クロスバリデーションなどで推定することができる. しかし, 本論文における実験では簡単のためデータ数 U を用いて $\epsilon = U/2$ と定めた. また, データ数が多い実データを用いた実験においては, 正則化項を加えて過適合を防ぐ代わりに, アルゴリズムの繰り返しを収束する前に早い段階で停止するという方法をとった.

付録 7: 未評価アイテムの取り扱い

ここでは, グループ化ランキング観測データにおいてユーザが評価していないアイテムが存在する場合の推定アルゴリズムを導く. 実用上, 全てのユーザが全てのアイテムを評価するという状況は考えにくく, アイテム選好度パラメタの推定アルゴリズムは未評価アイテムを扱える必要がある.

観測データに未評価のアイテムがある場合でも, 図 5.2 のアルゴリズムにおける初期化ステップはそのままでよい. つまり, M 変数に関する U 個の最適化問題 (5.10) を解けば良い. しかし, em アルゴリズムにおける e-ステップは, 射影 (5.12), (5.13) を未評価アイテムを考慮したものに修正する必要がある. 記述の簡単のため, ユーザのインデックス u を省略すると, 図 5.2 のアルゴリズムにおける e 射影は次のように修正される. このユーザが評価したアイテムに関しては

$$\hat{\theta}_i(t) = \frac{\theta_i(t)}{\sum_{j \in G_{m|i}} \theta_i(t)} \hat{\Theta}_{m|i} \times \frac{e^{-KL(\hat{\Theta}, \Theta(t))}}{e^{-KL(\hat{\Theta}, \Theta(t))} + \Theta_*(t)},\ i \in G_m,\ m = 1, \ldots, M \tag{6.13}$$

となり, 未評価アイテムに対しては

$$\hat{\theta}_i(t) = \theta_i(t) \times \frac{1}{e^{-KL(\hat{\Theta}, \Theta(t))} + \Theta_*(t)},\ i \notin {}^{\forall}G_m, \tag{6.14}$$

となる. ここで, 未評価のアイテムに対応するパラメタ全ての和を

$$\Theta_*(t) = \sum_{i \notin {}^{\forall}G_m} \theta_i(t),$$

で表し, $KL(\hat{\Theta}, \Theta(t)) = \sum_{l=1}^{M} \hat{\Theta}_l \log \frac{\hat{\Theta}_l}{\Theta_l(t)}$ とした. もし未評価アイテムがない場合には, $\Theta_*(t) = 0$ であり, 上式はアルゴリズム 5.2 における e 射影に一

致する. これらの修正した更新則は次のように導出できる. 図 5.2 のアルゴリズムにおける e 射影を導出するために, 最適化問題 (5.10) の解 $\{\hat{\Theta}_m\}_{m=1}^M$ を $\sum_{i\in G_m}\theta_i$ に関する制約条件として利用した. グループ化ランキング観測データに未評価アイテムがある場合には, 等式制約 $\sum_{i\in G_m}\theta_i=\hat{\Theta}_m$ を, 比に関する制約 $\sum_{i\in G_m}\theta_i\propto\hat{\Theta}_m$ に置き換える. そして, これらの値の比を $\sum_{i\in G_m}\theta_i$ の制約として用いるのである. つまり, $\hat{\Theta}_m/\hat{\Theta}_M=c_m,\ m=1,\dots,M-1$ として, グループ化ランキング観測データによって定義される部分多様体は

$$\mathcal{D}=\left\{\theta\in\Delta_{N-1}\left|\frac{\sum_{i\in G_m}\theta_i}{\sum_{i\in G_M}\theta_i}=c_m,\ m=1,\dots,M-1\right.\right\}$$

のように修正される. その上で, アルゴリズムの前回の繰り返しによって得られている $\theta(t)$ から部分多様体 $\mathcal{D}$ への e 射影は, 次の最適化問題の解として得られる:

$$\min_{\theta}\quad \sum_{i=1}^N\theta_i\log\frac{\theta_i}{\theta_i(t)}\ , \tag{6.15}$$

$$\text{subject to}\quad \frac{\sum_{j\in G_m}\theta_j}{\sum_{j\in G_M}\theta_j}=c_m,\ m=1,\dots,M-1, \tag{6.16}$$

$$\sum_{i=1}^N\theta_i=1. \tag{6.17}$$

この問題のラグランジアンは

$$F(\theta,\{\lambda_m\}_{m=1}^{M-1},\mu)=\sum_{i=1}^N\theta_i\log\frac{\theta_i}{\theta_i(t)}+\sum_{m=1}^{M-1}\lambda_m\left(\frac{\sum_{j\in G_m}\theta_j}{\sum_{j\in G_M}\theta_j}-c_m\right)+\mu\left(\sum_{i=1}^N\theta_i-1\right)$$

で, 未定乗数は $(\{\lambda_m\}_{m=1}^{M-1},\mu)$ である. ラグランジアンを $\theta_i, i=1,\dots,N$ に関して微分してゼロとおくことで, 次の方程式系を得る:

$$\theta_i=\begin{cases}\theta_i(t)\bar{\mu}^{-1}e^{-\lambda_m/\Theta_M}, & i\in G_m,\ m=1,\dots,M-1,\\ \theta_i(t)\bar{\mu}^{-1}\exp\left(\sum_{m=1}^{M-1}\frac{\lambda_m}{\Theta_M}\frac{\Theta_m}{\Theta_M}\right), & i\in G_M,\\ \theta_i(t)\bar{\mu}^{-1}, & i\notin{}^{\forall}G_m,\end{cases}\ .$$

ただし, $\bar{\mu} = e^{1+\mu}$ とした. 各グループに属する θ_i を加え合わせるとで,

$$\begin{aligned}
\Theta_m &= \Theta_m(t)\bar{\mu}^{-1}e^{-\lambda_m/\Theta_M}, \ \ m = 1, \ldots, M-1, && (6.18)\\
\Theta_M &= \Theta_M(t)\bar{\mu}^{-1}\exp\left(\sum_{m=1}^{M-1}\frac{\lambda_m}{\Theta_M}\frac{\Theta_m}{\Theta_M}\right), && (6.19)\\
\Theta_* &= \Theta_*(t)\bar{\mu}^{-1} && (6.20)
\end{aligned}$$

となり, これらの連立方程式を $\lambda_m/\Theta_M,\ m=1,\ldots,M-1$ に関して解けば,

$$\begin{aligned}
\frac{\lambda_m}{\Theta_M} &= \sum_{l=1}^{M}\left(\log\frac{\Theta_m(t)}{\Theta_l(t)} - \log\frac{c_m}{c_l}\right)c_l\hat{\Theta}_M && (6.21)\\
&= \sum_{l=1}^{M}\hat{\Theta}_l\log\frac{\hat{\Theta}_l}{\Theta_l(t)} + \log\frac{\Theta_m(t)}{\hat{\Theta}_m} && (6.22)\\
&= KL(\hat{\Theta}, \Theta(t)) + \log\frac{\Theta_m(t)}{\hat{\Theta}_m} && (6.23)
\end{aligned}$$

を得る. ただし, $c_M = \frac{\hat{\Theta}_M}{\hat{\Theta}_M} = 1$ である.

次に, $\bar{\mu}$ を考える. $\sum_{m=1}^{M-1}\Theta_m + \Theta_M + \Theta_* = 1$ であることと, 式 (6.18), (6.19), (6.20) から,

$$\begin{aligned}
\bar{\mu} &= \sum_{m=1}^{M-1}\Theta_m(t)e^{-\frac{\lambda_m}{\Theta_M}} + \Theta_M(t)\exp\left(\sum_{l=1}^{M-1}\frac{\lambda_l}{\Theta_M}c_l\right) + \Theta_*(t) && (6.24)\\
&= \sum_{m=1}^{M}\Theta_m(t)e^{-\frac{\lambda_m}{\Theta_M}} + \Theta_*(t) && (6.25)
\end{aligned}$$

である. ただし, $\lambda_M = -\sum_{m=1}^{M-1}c_m\lambda_m$ とした. なお, 式 (6.23) は $m=M$ についても成立する. 式 (6.23) を式 (6.25) に代入して,

$$\bar{\mu} = \sum_{m=1}^{M}\hat{\Theta}_m e^{-KL(\hat{\Theta},\Theta(t))} + \Theta_*(t) = e^{-KL(\hat{\Theta},\Theta(t))} + \Theta_*(t)$$

を得る. 最後に, 式 (6.23) と式 (6.25) を用いると, 未評価アイテムに対する e 射影の公式が式 (6.13) と式 (6.14) で得られる.

アルゴリズムの m-ステップは e-射影に依存し, 観測データに直接は依存しないため, m-ステップはアルゴリズム 5.2 と同様に行ってよい.

参考文献

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Fan. R. K. Chung. *Spectral graph theory*, Vol. 92 of *CBMS Regional Conference Series*. American Mathematical Society, Providence, 1997.

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, Vol. 15, No. 6, pp. 1373–1396, 2003.

[4] Xiaofei He and Partha Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[5] R.A.Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, Vol. 7, pp. 179–188, 1936.

[6] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighborhood component analysis. In *NIPS*, 2004.

[7] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pp. 451–458. MIT Press, Cambridge, MA, 2006.

[8] Thore Graepel. Kernel matrix completion by semidefinite programming. In *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks*, pp. 694–699, London, UK, 2002. Springer-Verlag.

[9] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, Vol. 5, pp. 27–72, 2004.

[10] 日野英逸, 村田昇. 条件付きエントロピー最小化に基づく教師付き次元削減手法. IBIS2009: 第 12 回情報論的学習理論ワークショップ, 2009.

[11] Hideitsu Hino and Noboru Murata. Conditional entropy miminization criterion for dimensionality reduction and multiple kernel learning. *Neural Computation*, Vol. 22, No. 9, 2010.

[12] 日野英逸, 藤本悠, 村田昇. Grouped ranking モデル: Plackett-luce モデルの一般化とその応用. IBIS2008: 第 11 回情報論的学習理論ワークショップ, 2008.

[13] H. Hino, Y. Fujimoto, and N. Murata. Item preference parameters from grouped ranking observations. In *13-th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2009)*, 2009.

[14] Yu Fujimoto, Hideitsu Hino, and Noboru Murata. Item-user preference mapping with mixture models -data visualization for item preference-. In *International Conference on Knowledge Discovery and Information Retrieval (KDIR2009)*, 2009.

[15] Yu Fujimoto, Hideitsu Hino, and Noboru Murata. An estimation method for bradley-terry and its related models based on the bregman divergence. In *Learning Workshop: Computational and Biological Learning Society*, 2010.

[16] Hideitsu Hino, Yu Fujimoto, and Noboru Murata. A grouped ranking model for item preference parameter. *Neural Computation*, to appear.

[17] Claude Elwood Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, Vol. 27, pp. 379–423,623–656, 1948.

[18] T.M.Cover and J.A.Thomas. *Elements of information theory*. John Wiley and Sons, Inc., 1991.

[19] Alfred Renyi. On measures of information and entropy. In *4th Berkeley Symposium on Mathematics, Statistics and Probability*, pp. 547–561, 1960.

[20] A.Hyvärinen, J.Karhunen, and E.Oja. *Independent Component Analysis*. J. Wiley, New York, 2001.

[21] Mark L.G. Althouse and Chein-I Chang. Image segmentation by local entropy methods. *Image Processing, International Conference on*, Vol. 3, p. 3061, 1995.

[22] Jose A. Costa, Alfred O. Hero, and III. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE TRANS. ON SIGNAL PROCESSING*, Vol. 52, pp. 2210–2221, 2004.

[23] L. F. Kozachenko and n. N. Leonenko. Sample estimate of entropy of a random vector. *Problems of Information Transmission*, Vol. 23, pp. 95–101, 1987.

[24] Lev Faivishevsky and Jacob Goldberger. ICA based on a smooth estimation of the differential entropy. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pp. 433–440. 2009.

[25] Matthew P. Wand and M. Jones. *Kernel Smoothing*. Chapman & Hall/CRC, December 1994.

[26] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for SVMs. In *NIPS*, pp. 668–674, 2000.

[27] Q. Tao, D. Chu, and J. Wang. Recursive support vector machines for dimensionality reduction. *IEEE Transactions on Neural Networks*, Vol. 19, No. 1, pp. 189–193, 2008.

[28] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pp. 1473–1480. MIT Press, Cambridge, MA, 2006.

[29] M.Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Mach. Learn. Res.*, Vol. 8, pp. 1027–1061, 2007.

[30] Jose M. Leiva-Murillo and Antonio Artes-Rodriguez. A gaussian mixture based maximization of mutual information for supervised feature extraction. In Carlos Garcia Puntonet and Alberto Prieto, editors, *ICA*, Vol. 3195 of *Lecture Notes in Computer Science*, pp. 271–278. Springer, 2004.

[31] Samuel Kaski and Jaakko Peltonen. Informative discriminant analysis. In *In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). AAAI Press, Menlo Park, CA*, pp. 329–336. AAAI Press, 2003.

[32] Sajama and Alon Orlitsky. Supervised dimensionality reduction using mixture models. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pp. 768–775, New York, NY, USA, 2005. ACM.

[33] Goldberger J. Peltonen J. and Kaski S. Fast semi-supervised discriminative component analysis. In *MLSP 2007: Machine Lerning for Signal Processing*, pp. 312–317, 2007.

[34] Ran He, Bao-Gang Hu, and Xiaotong Yuan. Robust discriminant analysis based on nonparametric maximum entropy. In Zhi-Hua Zhou and Takashi Washio, editors, *ACML*, Vol. 5828 of *Lecture Notes in Computer Science*, pp. 120–134. Springer, 2009.

[35] J.W. Fisher III and J. Principe. Entropy manipulation of arbitrary nonlinear mappings. In *Proc. IEEE Workshop Neural Nets for Signal Proc., 14-23, Amelia Island*, pp. 14–23. IEEE Press, 1997.

[36] J.C.Principe and Xu Dongxin. An introduction to information theoretic learning. In *Neural Networks, 1999. IJCNN '99. International Joint Conference on*, pp. 1783–1787, 1999.

[37] Kari Torkkola and William M. Campbell. Mutual information in learning feature transformations. In *In Proceedings of the 17th Inter-*

national Conference on Machine Learning, pp. 1015–1022. Morgan Kaufmann, 2000.

[38] Kari Torkkola. Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.*, Vol. 3, pp. 1415–1438, 2003.

[39] Kenneth E. Hild, Deniz Erdogmus, Kari Torkkola, and Jose C. Principe. Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 9, pp. 1385–1392, 2006.

[40] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[41] 赤穂昭太郎. カーネル多変量解析. 岩波書店, 2008.

[42] 斉藤三郎. 再生核の理論入門. 牧野書店, 2002.

[43] Huyen Do, Alexandros Kalousis, Adam Woznica, and Melanie Hilario. Margin and radius based multiple kernel learning. In *ECML/PKDD (1)*, pp. 330–343, 2009.

[44] Seung-Jean Kim, Alessandro Magnani, and Stephen Boyd. Optimal kernel selection in kernel fisher discriminant analysis. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 465–472, New York, NY, USA, 2006. ACM.

[45] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müllers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48, 1999.

[46] Alain Rakotomamonjy, Francis R. Bach, Stéphane Canu, and Yves Grandvalet. SimpleMKL. *JMLR*, Vol. 9, pp. 2491–2521, 2008.

[47] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, Vol. 7, pp. 1531–1565, 2006.

[48] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pp. 487–493, Cambridge, MA, USA, 1999. MIT Press.

[49] Koji Tsuda, Shotaro Akaho, Motoaki Kawanabe, and Klaus-Robert Müller. Asymptotic properties of the fisher kernel. *Neural Comput.*, Vol. 16, No. 1, pp. 115–137, 2004.

[50] Koji Tsuda, Taishin Kin, and Kiyoshi Asai. Marginalized kernels for biological sequences. *Bioinformatics*, Vol. 18, pp. 268–275(8), July 2002.

[51] Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, Vol. 67, pp. 106–135, 2005.

[52] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[53] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, Vol. 42, No. 3, pp. 287–320, March 2001.

[54] Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 149–158. AAAI Press, 1999.

[55] Gert R. G. Lanckriet, Minghua Deng, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Pacific Symposium on Biocomputing*, pp. 300–311, 2004.

[56] Darrin P. Lewis, Tony Jebara, and William Stafford Noble. Nonstationary kernel combination. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 553–560, New York, NY, USA, 2006. ACM.

[57] C. L. Mallows. Non-null ranking models.I. *Biometrika*, Vol. 44, No. 1/2, pp. 114–130, 1957.

[58] M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society, Series B*, Vol. 48, No. 3, pp. 359–369, 1986.

[59] P. Diaconis. *Group Representations in Probability and Statistics*. IMS Lecture Notes, 1988.

[60] T. B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Computational Statistics & Data Analysis*, Vol. 41, No. 3-4, pp. 645–655, January 2003.

[61] M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *22nd Conference on Uncertainty in Artificial Intelligence (UAI07)*, Vancouver, British Columbia, July 2007.

[62] J. Huang, C. Guestrin, and L. Guibas. Efficient inference for distributions on permutations. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007.

[63] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. In *Advances in Neural Information Processing Systems 20(NIPS2007)*, 2007.

[64] L. M. Busse, P. Orbanz, and J. M. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th international conference on Machine learning(ICML2007)*, pp. 113–120, New York, NY, USA, 2007. ACM.

[65] R. A. Bradley and M. Terry. The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, Vol. 39, pp. 324–345, 1952.

[66] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, Vol. 26, No. 2, pp. 451–471, 1998.

[67] T. Huang, R. C. Weng, and C. Lin. Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, Vol. 7, pp. 85–115, 2006.

[68] T. Takenouchi and S. Ishii. Ternary bradley-terry model-based decoding for multi-class classification. In *IEEE International Workshop on Machine Learning For Signal Processing*, 2008.

[69] R. L. Plackett. The analysis of permutations. *Applied Statistics*, Vol. 24, No. 2, pp. 193–202, 1975.

[70] D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Anaals of Statistics*, Vol. 32, No. 1, pp. 384–406, 2004.

[71] J. I. Marden. *Analyzing and Modeling Rank Data*, Vol. 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1995.

[72] S. Amari and N. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[73] S. Amari. Information geometry of the EM and *em* algorithms for neural networks. *Neural Networks*, Vol. 8, No. 9, pp. 1379–1408, 1995.

[74] 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 Series. B*, Vol. 39, , 1977.

[75] H. Sahbi and N. Boujemaa. Fuzzy clustering: Consistency of entropy regularization. *Advances in Soft Computing*, Vol. 2, pp. 95–107, 2005.

[76] A. Zenebe and A. F. Norcio. Visualization of item features, customer preference and associated uncertainty using fuzzy sets. In *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society*, pp. 7–12, 2007.

[77] G. Mei and C. R. Shelton. Visualization of collaborative data. In *Proceedings of the Twenty-Second International Conference on Uncertainty in Artificial Intelligence*, pp. 341–348. AUAI Press, 2006.

[78] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engneering*, Vol. 17, No. 6, pp. 734–749, 2005.

[79] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175–186, Chapel Hill, North Carolina, 1994. ACM.

[80] T. Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems 12*, pp. 914–920, 2000.

[81] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[82] J. Riedl and J. Konstan. Movielens dataset. 2000.

[83] C. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pp. 22–32, New York, NY, USA, 2005. ACM.

[84] M. A. Croon and R. Luijkx. *Latent structure models for ranking data*. Springer, New York, 1993.

[85] Toshihiro Kamishima and Shotaro Akaho. Efficient clustering for orders. In *Proceedings of the 2nd International Workshop on Mining Complex Data*, pp. 274–278, 2006.

[86] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pp. 396–404. 2009.

[87] J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 377–384, New York, NY, USA, 2009. ACM.

[88] David H. Stern, Ralf Herbrich, and Thore Graepel. Matchbox: large scale online bayesian recommendations. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pp. 111–120, New York, NY, USA, 2009. ACM.

[89] Pierre Comon. Independent component analysis, a new concept? *Signal Process.*, Vol. 36, No. 3, pp. 287–314, 1994.

[90] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, Vol. 2, pp. 94–128, 1999.

[91] N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun.*, Vol. 26, No. 4, pp. 283–297, 1998.

[92] Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, Vol. 58, pp. 155–176, 1996.

[93] Marco Loog and Robert P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 26, No. 6, pp. 732–739, 2004.

[94] Yu Zhang and Dit-Yan Yeung. Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In Wray L. Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor, editors, *ECML/PKDD (2)*, Vol. 5782 of *Lecture Notes in Computer Science*, pp. 602–616. Springer, 2009.