様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月２７日現在

機関番号：３２６８９
研究種目：若手研究（B）
研究期間：2010 ～ 2012
課題番号：22700292
研究課題名（和文）　Bregman 情報量に基づく統計モデルの拡張とその応用

研究課題名（英文）　Extension and Application of Statistical Models Based on Bregman Divergence

研究代表者
藤本　悠　（FUJIMOTO　YU）
早稲田大学・ナノ理工学研究機構・研究員
研究者番号：40434302

研究成果の概要（和文）：
本研究では統計量の算出の際にデータの偏り等に対してロバストな推定を実現する Bregman 情報量と対応付けて導出される統計モデルのクラスに着目し，実際のデータ解析への応用を視野に入れたモデルの性質の解析や推定方法の提案を行った．特に情報量との対応付けの過程で導出される正値間の乗算則の拡張を用いることで統計的独立性の一般化を行なった．これにより条件付き独立性を仮定するような既存のデータ解析の枠組みを拡張でき，判別や回帰の精度の改善が可能となることを示した．

研究成果の概要（英文）：
In this study, we focused on a class of statistical models associated with the Bregman divergence which achieves robust estimation in calculating statistics. We have proposed some methods for estimation of these models, and analyzed some properties from the viewpoint of application for data analysis. Particularly, we generalized the definition of statistical independence by using an extension of the multiplication rule between positive values; the extension is derived from the Bregman divergence. We have shown that the proposed statistical models based on generalized independence can be useful tools in practical data analysis by alleviating the conditional independence assumed in the conventional methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,000,000	300,000	1,300,000
2011 年度	900,000	270,000	1,170,000
2012 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	2,800,000	840,000	3,640,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：統計的学習理論，情報量，統計モデル，独立性の一般化，情報基礎，機械学習

１．研究開始当初の背景

Kullback-Leibler 情報量の一種の一般化である Bregman 情報量は統計モデルの推定指

標として用いることで，サンプル数の少なさや外れ値に起因したデータの偏りに対するロバストな推定の実現を可能にすることが知られている．このような情報量の一般化の考え方と対応する形で，考え得る統計モデルのクラスも一般化できることが示されているが，このような拡張された統計モデルの性質や推定方法の枠組み，及び実際のデータ解析現場への応用性などは研究開始当初まであまり論じられていなかった．特にこのような情報量の一般化の考え方と対応して得られる統計モデルそのものの拡張の枠組みに着目すると，１つの視点として統計的な独立性の定義の一般化を議論することが可能となる．金融工学において発展してきた copula や統計物理学において発展してきた Tsallis 統計など，各分野においてこのような一般化の考え方は存在しているが，これらとの関連性を論じながら統計的推論の枠組みの中で拡張した統計モデルの性質，有効性を整理し，実際のデータ解析の現場への応用可能性を検証することが必要であった．

２．研究の目的

本研究では Kullback-Leibler 情報量の一般化である Bregman 情報量の考え方と深い関係のある凸関数とその導関数によって特徴づけられる乗除算などの算術演算，及び統計的独立性の拡張を提案し，また統計的推論の枠組みの中でこの拡張を利用した統計モデルの性質，有用性を整理し，実際のデータ解析などの現場への応用の土台を構築することを目的として掲げていた．中でも特に次の点に関して整理することをこの研究の中核目的としていた．

- Bregman 情報量で用いられる凸関数の導関数で特徴づけられる乗除算の一般化に基づく確率演算の性質の解析．
- 乗除算の一般化によって導出される統計的独立性の拡張に基づく変数間の弱い特殊な依存関係表現の性質の解析．
- 弱い特殊な独立性表現をデータ解析に用いる際の方法論の提案やその有用性の整理．

３．研究の方法

上記で挙げた目的に対する本研究の取り組みとして，

(1)．提案モデルと関連研究との相違点の整理
(2)．提案モデルに関する情報量･統計的推論の観点からの整理
(3)．提案モデルの実際のデータ解析の場での応用可能性に関する議論
(4)．提案モデルを効率的に推定するためのアルゴリズムの導出

の４点を重点的に行った．

まず提案モデルのような考え方は様々な分野で個別に議論・発展がなされてきているが，これに関して体系立てられた議論が統計的機械学習の分野において成されていないという背景があった．そこで(1)では他分野での関連研究等を参考にする中で，特に関連性の深い話題として，統計物理学の分野における Tsallis 統計とその周辺で出てくる統計モデル，及び金融工学でしばしば用いられる copula などに着目をした．本研究期間を通して当該分野における文献の調査，及び各分野の研究者との議論などを通じて，関連研究との比較を行いながら理論的な背景の整理を行なってきた．また，一種の一般化線形モデルとしての観点からの提案モデルの解釈も行なった．

(2)については本研究期間で Bregman 情報量の意味でのエントロピーの観点からの特徴付けや，条件付き確率の扱いを議論することを行なってきた．これにより本研究で提案する拡張・一般化を統計的な独立性の定義に対して応用することで変数間の一種の弱い特殊な依存関係，相関の表現を可能にすることができるようになり，情報量の一般化に際して導入される凸関数の形状と実際に表される弱い依存表現の間の関係性等の整理を行った．また，複数の確率変数が存在する時に同時分布，条件付き分布，周辺分布のそれぞれの対応を本研究で提案する枠組みで一般化することで Bayes の定理と同等の関連性が成立することなどを示した．

本研究で提案する独立性の定義の一般化の枠組みは，独立性，条件付き独立性等の仮定に基づいて構築される多くの統計モデルに対する拡張として導入することができる．そのため上記(3)に対する取り組みとして，厳密な独立性が成り立たないデータの生　成機構を表現し得るような柔軟な表現力を持つ統計モデルの構築を本研究の枠組みによって行うことを提案してきた．特に一般化線形モデルの観点からの回帰問題やナイーブベイズモデルや非負値行列分解などを利用した判別問題への応用を中心に議論し，データ解析の場での有用性について実験を通して議論を行なってきた．

本研究で提案する一般化独立モデルは複数の変数間の弱い特殊な依存関係の表現を可能にする一方で，周辺分布を記述するパラメタに関する最尤推定を行うためには非線

形最適化を数値的に行う必要がある．これを避けるため，(4)では一種の近似として経験周辺分布を用いた簡易な推定法を導入し，その枠組でモデルの精度などを議論してきた．また，乗算の一般化から導出される非負値行列分解を実現するための最適化に際してもいくつか議論を行ない，変数の次元が大きいような問題に関するスケーラビリティを確保した確率的勾配降下法に基づく推定方法などの実装，提案を行なってきた．

４．研究成果

(1)．提案モデルと関連研究との相違点の整理

前述の通り，金融工学や統計物理学などの分野において本研究と関連する考え方が提案されている．例えば金融工学でしばしば用いられる copula（特に Archimedean copula と呼ばれるクラス）は本研究で対象としている一般化独立モデルと非常に良く似た考え方から構成される同時確率分布表現手法である．周辺累積密度関数を用いて同時分布を記述する copula と，周辺密度関数そのものを用いて同時分布を記述する一般化独立モデルの差異について整理を行い，例えば一様な周辺分布を用いた時に表現可能な同時分布表現の両者の違いなどを明らかにした．

また，統計物理学などの文脈で議論されている Tsallis 統計などで出てくる冪乗に基づく統計モデルの構築の手続きは，同様に冪乗で構成される凸関数を用いて構築された提案モデルと深い関係があることに言及し，弱い特殊な依存関係を表現するような統計モデルのクラスとして利用できることなどを整理してきた．

さらに，一種の一般化線形モデルとしての観点から提案モデルのクラスを議論することも行なった．これにより情報量や提案モデルの性質を定める凸関数と一般化線形モデルにおけるリンク関数の関係が対応付けられ，例えば回帰の文脈では歪んだ誤差分布を表現することで精度を改善し得るという性質があることが明らかになった．

(2)．提案モデルに関する情報量･統計的推論の観点からの整理

本研究で対象としている統計モデルは，凸関数によって定義される Bregman 情報量に基づく統計モデルの推定の文脈で自然に出てくるものとなっている．幾何学的な観点から提案モデルのイメージを描くと，図１で示したような周辺分布の組合せによって表される一般的な独立モデルを図２に示すような同時分布の空間の中でどのように曲げて配置するかが提案モデルの表現力の柔軟性の鍵となることが分かった．

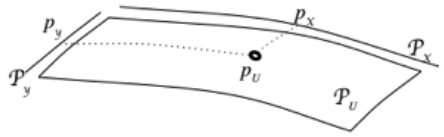

図 1．独立モデルの幾何学的な解釈

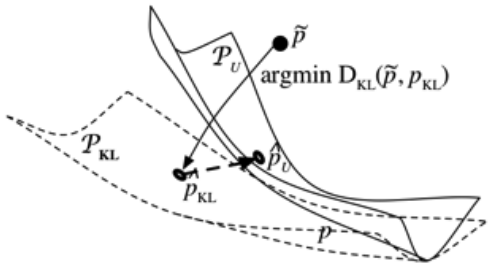

図 2．提案モデルの推定方法の幾何学的な解釈

合わせて，統計的な推論の観点から本研究で提案する枠組みの整理を行なった．確率変数が複数存在する時にはいわゆる Bayes の定理によってそれぞれの変数に関する同時分布，条件付き分布，周辺分布の関係を議論することができる．本研究で提案する一般化の枠組みを用いると Bayes の定理に関する拡張が可能となり，一般の条件付き分布と一意に対応付けられる条件付き関数によって整合性のある各分布間の関係が成立することなどを示した．

(3)．提案モデルの実際のデータ解析の場での応用可能性に関する議論

本研究で中心的に論じてきた一般化した独立性は，見方を変えると弱い特殊な依存関係を表していることになる．このことを念頭におき提案モデルのデータ解析の文脈での有用性を検証するために，

① 一般化線形モデルの観点からの拡張
② ナイーブベイズモデルの拡張
③ 非負値行列分解(NMF)の拡張

の３点を具体的に試みてきた．

まず①では対数線形モデルの一般化を本研究で用いる弱い独立性を表すモデルによって行い，これによって推薦システムなどの文脈で重要となる協調フィルタリング問題への応用が可能なことを示した．行列分解手法による同問題へのアプローチと比較すると，乗算の規則を支配する１パラメタのチューニングを行うことで推定精度が比較的大きく改善されることが確認でき，実問題への応用に際した本モデルの柔軟性，有用性を示す一例となっている．

また，②では確率変数間の弱い依存関係を提案手法により表現するという非常に単純な拡張によって，例えばナイーブベイズモデ ルを判別器として利用する際の判別性能の改善が可能になるといった応用上興味深い知見を得ることができた．

また，③では本研究で用いる乗算則の一般化によって非負値行列分解の拡張を行うことが可能となることを示し，行列分解の結果として得られる基底で張られる空間が乗算の一般化に対応して曲がり方が変化するような超曲面となることを示した(図 3 参照)．これによって，一種非線形なデータの分布を柔軟に表現するための１つの方法として本提案モデルが利用できることに言及し，また，ベンチマークデータに対する判別問題の観点から本モデルの有用性を検証した．結果として，データの散布状況を提案手法で得られる基底で張られる空間が柔軟に表現し得るという様子を確認し，上記の非線形性が有効に働くことで判別精度が上がるという仕組みが確認できた．

(4)．提案モデルを効率的に推定するためのアルゴリズムの導出

本研究で提案する考え方は比較的広範な手法への応用が見込める一方で，実際のデータ解析の場で利用する際には既存の手法で用いられている最適化の方法論がそのままでは使えない場合が出てくる．そのため，(3)で提案したそれぞれのモデルに対してどのように推定を実現するかを具体的に導出してきた．

例えば，与えられたデータに対して最尤理論の枠組みで提案拡張独立モデルを推定する際には周辺分布に相当する一変量分布の最適化と変量間の弱い依存関係を表すパラメタに関する非線形同時最適化が必要となる．特に厳密解を数値計算的に導く場合と段階的推定法を応用して近似解を獲得する場合とで，どのような状況でどの程度精度が異なるのかを汎化性能の観点から議論を試み，その結果推定に用いるサンプル数に応じて両推定法の優位性が変わってくることなどを確認した．また，二値変量間の同時確率表を対象として，従来の独立モデルと拡張独立モデルの幾何学的な構造の比較を行い，段階的推定法を用いる際に一様分布への収縮が起きることなどを確認した．加えて，一般化した独立性を応用したナイーブベイズモデルの拡張などに対してもこの段階的推定法が利用できることを確認した．また，例えば変量の次元数が非常に大きくなり得るような問題に対して拡張した非負値行列分解を行うことを考え，確率的勾配降下法を利用したスケーラビリティが見込める推定手法の提案を行い，有用性の確認を行った．

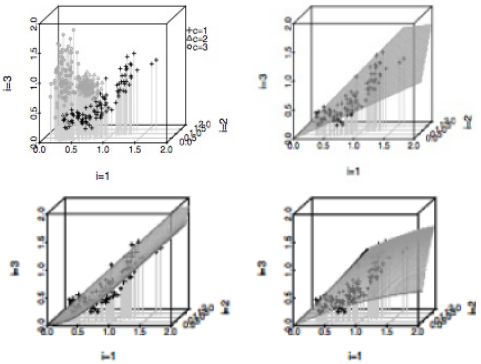

図 3. 提案手法によるデータの散布状況の柔軟な表現の様子（左上:サンプルの散布図．右上:通常の行列分解で得られた基底で表現される平面．下:提案手法によって得られた基底で表現される曲面．）

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Yu Fujimoto & Noboru Murata, “A Generalization of Independence in Statistical Models for Categorical Distribution”, International Journal of Data Mining, Modelling and Management, 査読有り, Vol.4, No.2, 2012, pp.172-187, DOI:10.1504/IJDMMM.2012.046809

〔学会発表〕（計７件）

① Yu Fujimoto & Noboru Murata, “Extended Independent Model Based on Modified Product Rule from the Copula Viewpoint”, Copulae in Mathematical and Quantitative Finance, 2012/07/10-11, Cracow, Poland.

② Yu Fujimoto & Noboru Murata, “Nonnegative Matrix Factorization via Generalized Product Rule and its Application for Classification”, The 10th International Conference on Latent Variable Analysis and Signal Separation, pp.263-271, 2012/03/14, Tel-Aviv, Israel.

③ Yu Fujimoto, “Extended Independent Model Based on Modified Product Rule: Comparison of FIML and TSML Estimation Approaches”, International Workshop on Anomalous Statistics, Generalized Entropies, and Information Geometry, 2012/03/08,

Nara, Japan.

④ 藤本悠, “Bregman divergence に基づく統計モデルの推定と拡張”, 統計的機械学習セミナー（招待講演）, 2011/11/24, 東京都

⑤ 藤本悠 & 村田昇, “一般化乗算に基づく NMF の拡張と判別問題への応用”, 第 14 回情報論的学習理論ワークショップ, 2011/11/09, 奈良県

⑥ 藤本悠, “推薦システムにおける一般化線形モデルの適用について -主効果モデルによる評価得点推定”, 第 13 回情報論的学習理論ワークショップ, 2010/11/04, 東京都

⑦ Yu Fujimoto & Noboru Murata, “A Generalization of Independence in Naïve Bayes Model”, The 11th International Conference on Intelligent Data Engineering and Automated Learning, pp.153-161, 2010/09/02, Paisley, Scotland, UK

６．研究組織

(1)研究代表者

藤本 悠 (FUJIMOTO YU)

早稲田大学・ナノ理工学研究機構・研究員

研究者番号：40434302

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：