

博士学位論文審査要旨

2017年1月19日

論文題目：Extensions of nonnegative matrix factorization
for exploratory data analysis
「探索的なデータ分析のための非負値行列因子分解の拡張」

学位申請者： 阿部 寛康

審査委員：

主 査： 文化情報学研究科 教授 宿久 洋
副 査： 文化情報学研究科 教授 川崎 廣吉
副 査： 文化情報学研究科 教授 矢野 環
副 査： 文化情報学研究科 教授 金 明哲
副 査： 大阪大学大学院人間科学研究科 教授 足立 浩平

要 旨：

本論文は、全要素が非負の実数として与えられた非負値行列に対する行列分解法である非負値行列因子分解(NMF)について、非負値行列に現れる様々な特徴を踏まえつつ探索的に分析を行うための改良を加えた新手法を提案したものである。

第1章では、NMFの概要と、本論文で注目する探索的なNMF手法のための4つの拡張方向である因子行列の数、直交制約、データに仮定する確率分布、零過剰モデルについて、これまでに発表された研究成果に言及しつつ述べている。第2章では、本論文で用いる表記について説明している。第3章では、4つの拡張方向について、それぞれの必要性に言及しながら、その特徴について数式等を用いて詳述している。第4章から第7章では、4つの拡張方向の組み合わせにより分類される、既存手法と提案手法合わせて16の手法について、問題の定式化、パラメータ更新式の導出、更新アルゴリズムについて詳述している。第8章では、紹介した各種NMF手法の特徴を捉えるための数値シミュレーションについて述べている。本シミュレーションにより、いくつかの新しく提案したNMF手法の優位性を示している。第9章では、実在するデータに対し提案手法を適用した結果を通じて、提案手法の既存手法に対する優位性と弱点、およびその他の特徴をまとめている。

本論文により、非負値行列因子分解における様々な問題に対応した手法が提案され、非負値行列に隠れた構造を捉えるための探索的な分析の可能性を広げた。よって本論文は、博士(文化情報学)(同志社大学)の学位論文として十分な価値を有するものと認められる。

総合試験結果の要旨

2017年1月19日

論文題目： Extensions of nonnegative matrix factorization
for exploratory data analysis
「探索的なデータ分析のための非負値行列因子分解の拡張」

学位申請者： 阿部 寛康

審査委員：

主査： 文化情報学研究科 教授 宿久 洋

副査： 文化情報学研究科 教授 川崎 廣吉

副査： 文化情報学研究科 教授 矢野 環

副査： 文化情報学研究科 教授 金 明哲

副査： 大阪大学大学院人間科学研究科 教授 足立 浩平

要 旨：

学位申請者は2014年度4月より本学大学院文化情報学研究科博士課程後期課程に在学しており、国内会議および国際会議での研究発表を通じて研究活動を積極的に行い、それらの成果を、国際会議 Proceedings に1本、計算機統計学関連の論文誌に1本の論文として公刊している。また、英語の語学試験にも合格していることから語学（英語）について十分な能力を有していると認定されている。

2017年1月18日水曜日18:00から約1時間の公聴会と30分の審査会において、種々の質疑応答の結果により博士（文化情報学）の学位を有するに十分な学力を有することを確認した。

よって、総合試験の結果は合格であると認める。

博士學位論文要旨

論文題目： Extensions of nonnegative matrix factorization for exploratory data analysis

探索的なデータ分析のための非負値行列因子分解の拡張

氏名： 阿部 寛康

要旨：

非負値行列因子分解 (Nonnegative Matrix Factorization, 以下 NMF) は、データとして与えられた非負値行列に対する行列分解法である。ここで非負値行列とは、その全要素が全て非負の実数として与えられた行列のことをいう。NMF は、画像認識、音声解析、遺伝子解析、テキストマイニング、推薦システム、マーケティング分析など、様々な分野で広く応用されている。それは、多様な分野で扱われるデータが非負値行列の形で与えられるということ、NMF がシンプルで扱いやすく非負値行列から説明のしやすい情報を取り出せるということが、理由としてあげられる。しかしながら、非負値行列に外れ値や多くの 0 が含まれているとき、NMF を実行するにあたって次のような困難が生じる。例えば、外れ値は行列分解の結果を意味のないものにしうる。また、0 要素が非負値行列内にたくさん存在するとき、つまり非負値行列が零過剰 (zero-inflated) な状態のとき、行列分解による非負値行列への近似が悪くなる。本稿ではこれらの困難に対処するために、非負値行列と行列分解による近似行列の間の誤差基準として多くの研究者に利用されている β ダイバージェンスに注目する。 β ダイバージェンスは外れ値に強いダイバージェンスのひとつであり、また、 β ダイバージェンスを NMF で利用することは非負値行列の各要素が Tweedie 分布に従うことを仮定したことに等しいことが知られている。Tweedie 分布はいくつかのよく知られた確率分布の一般化であり、そのケースのひとつとして知られているのが、ガンマ分布のポアソン混合である compound Poisson-gamma (CP) 分布である。CP 分布に従う確率変数はその値が 0 のときに mass をもつため、CP 分布は零過剰ポアソン分布のように容易に零過剰なモデルに拡張可能である。そこで本論文では NMF の誤差分布として零過剰 CP (ZICP) 分布を採用する。以上の議論からもわかるように、ZICP 分布に基づく NMF は外れ値や零過剰な状態に強い。以上で述べた ZICP 分布を応用した拡張に加えて、NMF により推定される非負値因子行列に列直交制約を課す Orthogonal NMF (ONMF) にも注目する。全ての要素が非負の実数で、かつ列直交するような行列は単純構造をもち、 k -means アルゴリズムで用いられる帰属行列のような役割を果たす。それゆえ ONMF で得られる結果は、より解釈容易なものになる。ONMF に関する既出の研究において ONMF 問題を解くためのいくつかの推定アルゴリズムが提案されているが、そのほとんどにおいて非負値因子行列の推定に関していくつかの問題を抱えている。本論文では、 k -means アルゴリズムの方法を応用することによって ONMF 問題を解くためのアルゴリズムを提案する。既に提案されているほとんど全ての ONMF は正規分布を誤差分布として仮定しており、ポアソン分布や CP 分布に基づく ONMF に関する研究はほとんどない。そこで k -means アルゴリズムを取り入れることによって、ポアソン分布や CP 分布を誤差分布として仮定した ONMF 問題を解くことができることを示す。NMF に関するさらなる議論としては、2 因子 NMF か 3 因子 NMF か、というものがある。2 因子 NMF は非負値データ行列を 2 つの非負値因子行列の積で近似する NMF であり、3 因子 NMF は 3 つのそれで近似する NMF である。3 因子 NMF では、得られたデータ行列の行対象および列対象にそれぞれ異なる因子が存在することが想定される。上記で述べた ZICP 分布や列直交制約は 2 因子 NMF だけでなく、3 因子 NMF にも応用可能である。本論文の目的は、NMF に関して包括的な議論を

行うことである。特に上記の2因子NMFか3因子NMFかということ、非負値因子行列に列直交制約を課すこと、仮定される確率分布やダイバージェンスのこと、零過剰モデルのこと、以上探索的な分析のためのNMFにとって有用な4つの特徴について議論を行う。加えて、これら特徴を考慮に入れたNMFについて、モデル設定、因子行列やその他パラメータ推定のための各更新式の導出、推定アルゴリズムなどを包括的に詳述する。さらに、提案する各NMFの特徴をつかむための数値シミュレーションや実データへの応用について述べる。