様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００００４４

研究課題名（和文）高次元データの理解のための最適なスケーリングと可視化技法

研究課題名（英文）Optimal scaling and visualization methods for understanding high-dimensional data

研究代表者
清　智也（Sei, Tomonari）

東京大学・大学院情報理工学系研究科・准教授

研究者番号：２０４０１２４２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：統計学で扱うデータは高次元であることが多い。本研究では高次元データの統計的推測において各変量のスケーリングが与える影響を調べ，可視化技法への応用可能性を考察した。特に，非線形スケーリングを許容した客観的総合指数の構成法，コピュラモデルの情報幾何学的考察，Textile Plot から定まる多様体の特徴付けに関する結果を得た。また関連する成果としてスケール不変性を持つベイズ事前分布の構成法，客観的総合指数の高次元一致性が示された。

研究成果の学術的意義や社会的意義
本研究は，現代社会のあらゆる場面に現れる高次元データについて，そのスケール変換に主眼を置くことにより，変数間の従属関係を抽出するためのツールを整備したという点で意義がある。特に非線形スケーリングを用いた客観的総合指数により，多様な指標をなるべく公平に重み付ける手法が確立された。また高次元データを可視化するTextile Plotや，高次元データを確率分布として表現するコピュラモデルについて，幾何学の観点から新しい性質が導き出された。

研究成果の概要（英文）：In statistics, we often deal with high-dimensional data. We investigated the effect of scaling for each variable in high-dimensional statistical inference, together with its applicability to data visualization. In particular, we constructed an objective general index under nonlinear scaling, studied information-geometric properties of copula models, and characterize a geometric object determined by Textile Plot. As other achievements, we demonstrated a construction method of Bayesian predictive distributions having scale invariance, and high-dimensional consistency of objective general index.

研究分野：統計科学

キーワード：確率分布のスケーリング　多変量データ　コピュラ　エントロピー　可視化　Textile plot　Textile set　コンパクト可微分多様体

１．研究開始当初の背景

定量的なデータを分析する際，前処理としてしばしばスケーリングが用いられる。典型的なスケーリングは標準化であり，平均 0，分散 1 となるようにデータを線形変換する。また高次元データの可視化を意図して提案された Textile Plot では，平行座標プロットにおいて各個体に対応する折れ線グラフができるだけ水平に近づくようなスケーリングを採用している。一方，多次元データの総合評価を行うための方法として客観的総合指数がある。これは各変数と総合指数が正の共分散を持つようなスケーリングである。

以上の方法は線形変換に限ったものである。しかし非対称な分布や歪んだ分布を持つデータに対しては線形変換以外のスケーリング，すなわち非線形スケーリングを考えることも有益である。従属構造を変えない非線形スケーリングとして，コピュラモデリングにおける変数変換が知られている。そこでは各変数の周辺分布が一様分布となるような非線形変換が施され，結果として従属構造のみが抽出される。しかしコピュラで用いられる非線形スケーリングは周辺分布のみを参照するため，変換した後の分布を可視化するという点においては情報をうまく使えていない可能性があった。

２．研究の目的

本研究の目的は、高次元データを理解するための非線形スケーリング技法の確立である。特に非線形スケーリングを用いた総合指数の構成を第一の目標とした。この研究過程においてコピュラモデリングの理解が不可欠であるという認識に至り，コピュラの幾何構造を調べることも目的とした。さらにスケーリングに基づく高次元データの可視化として重要なツールである Textile Plot について，その背後に潜む微分幾何構造を明らかにすることを 3 つ目の目標とした。

３．研究の方法

非線形スケーリングを用いた総合指標を定義し，その一意存在性を数学的に厳密な形で証明する。特に最適輸送理論やコピュラ理論の知見を活かして，多次元確率分布の非線形スケーリング問題の定式化を試みる。コピュラモデルについては周辺分布が未知のセミパラメトリックの設定における幾何学的性質をダイバージェンスの観点から検討する。また Textile Plot のなす空間である Textile Set について，その微分幾何学的な性質を調べる。

４．研究成果

(1) 非線形スケーリングに基づく客観的総合指数

多次元確率分布の非線形スケーリングに基づく客観的総合指数の拡張を行った。行列の対角スケーリング問題の類推として，非線形のスケーリング問題では新たに確率分布の狭義共正値性という概念を導入した。これによって自然な拡張を得ることができた。証明は最適輸送理論で用いられる Wasserstein 空間上の自由エネルギー汎関数の最小化問題に帰着するというアイデアに基づいている。その概念図を示したものが図 1 である。確率分布全体ではエネルギーがいくらでも小さくなってしまうが，座標ごとの単調増加なスケーリングに制限すると狭義共正値性のもとで最小点が存在する。これが最適なスケーリングを与える。

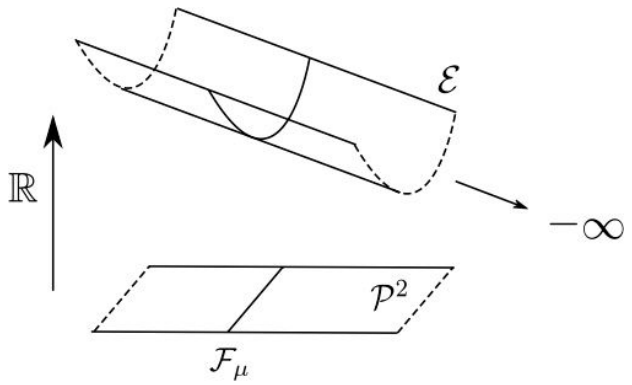

図 1：Wasserstein 空間におけるエネルギー最小化を表す概念図。

(2) コピュラモデルのダイバージェンス

本研究課題を進めていく過程で，コピュラモデルの推定問題も深く関わることが分かり，コピュ

ラモデルの情報幾何学的な性質を調べた。特に，周辺分布が未知であるようなセミパラメトリック・コピュラモデルに対するダイバージェンスを定義し，その性質を明らかにした。図 2 は，チェス盤コピュラと呼ばれる比較的扱いやすいコピュラについて，通常の Kullback-Leibler ダイバージェンスと，周辺分布が未知であることを勘案した 2 つのダイバージェンス（プロファイルダイバージェンス，順位ダイバージェンス）をプロットしたものである。図の右側では通常のダイバージェンスからの乖離が顕著であることが分かる。

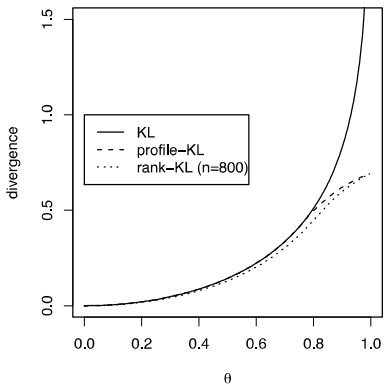

図 2：独立コピュラからチェス盤コピュラへのダイバージェンスをプロットした図。

(3) Strict Textile Set の特徴付け

先行研究で定義された Textile Set は，Textile Plot を描画するために計算される行列として可能なものを集めてできる集合である。ただしその定義において，最大固有値の部分を任意の固有値という条件に緩めていたため，実際に出力されない行列も含まれてしまっていた。これを最大固有値に限定したものを新たに定義し，Strict Textile Set と呼ぶことにした。図 3 は Strict Textile Set がある写像の逆像として表されることを示した図である。この特徴付けによって，Strict Textile Set の描像がより明確になった。

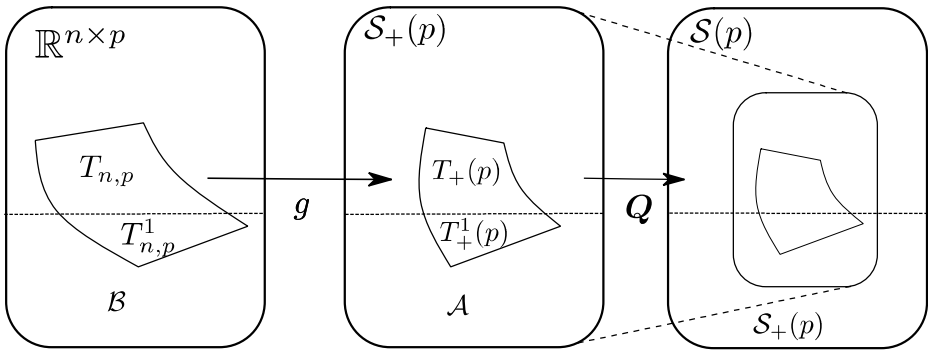

図 3：Strict Textile Set の特徴付け。

(4) その他の成果

上記に挙げた研究成果の他，当初は予期していなかった成果として以下の 3 つが挙げられる。

・客観的総合指数の高次元一致性
・Wishart 分布のベイズ予測におけるスケール不変な事前分布の構成
・スパース推定におけるホロノミック勾配法の応用

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　2件）

１．著者名 Harkonen Marc, Sei Tomonari, Hirose Yoshihiro	４．巻 3
２．論文標題 Holonomic extended least angle regression	５．発行年 2020年
３．雑誌名 Information Geometry	６．最初と最後の頁 149～181
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s41884-020-00035-1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 清 智也	４．巻 50
２．論文標題 客観的総合指標とその性質	５．発行年 2020年
３．雑誌名 品質	６．最初と最後の頁 57～60
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 清 智也、松本 和也	４．巻 68
２．論文標題 セミパラメトリックコピュラモデルにおけるダイバージェンスの性質	５．発行年 2020年
３．雑誌名 統計数理	６．最初と最後の頁 25～44
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Mitsunori Ogawa, Kazuki Nakamoto and Tomonari Sei	４．巻 －
２．論文標題 On the fractional moments of a truncated centered multivariate normal distribution	５．発行年 2020年
３．雑誌名 Communications in Statistics - Simulation and Computation	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610918.2020.1725821	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Ushio Tanaka, Masami Saga and Junji Nakano	4．巻 －
2．論文標題 NScluster: R Package for Maximum Palm Likelihood Estimation for Cluster Point Process Models using OpenMP	5．発行年 2019年
3．雑誌名 Journal of Statistical Software	6．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 清 智也	4．巻 29
2．論文標題 総合指数の数理	5．発行年 2019年
3．雑誌名 応用数理	6．最初と最後の頁 20 -- 26
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計19件（うち招待講演　1件／うち国際学会　11件）

1．発表者名 Tomonari Sei
2．発表標題 An objective general index and its generalizations
3．学会等名 Joint Statistical Meetings (JSM2020)（国際学会）
4．発表年 2020年

1．発表者名 清 智也
2．発表標題 セミパラメトリック・コピュラモデルの情報幾何学
3．学会等名 統計関連学会連合大会
4．発表年 2020年

１．発表者名
Tomonari Sei and Ushio Tanaka

２．発表標題
On geometric properties of the textile set and strict textile set

３．学会等名
Geometric Science of Information (GSI) 2019（国際学会）

４．発表年
2019年

１．発表者名
清 智也、松本 和也

２．発表標題
セミパラメトリックコピュラモデルにおけるダイバージェンス

３．学会等名
2019年度統計関連学会連合大会

４．発表年
2019年

１．発表者名
清 智也

２．発表標題
コピュラに対応する指数型分布の存在性と非一意性

３．学会等名
日本数学会2019年度秋季総合分科会

４．発表年
2019年

１．発表者名
Takuma Bando, Tomonari Sei and Kazuyoshi Yata

２．発表標題
Consistency of the objective general index in high dimensional settings

３．学会等名
International Symposium on Theories and Methodologies for Large Complex Data（国際学会）

４．発表年
2019年

１．発表者名
Tomonari Sei

２．発表標題
Holonomic gradient method for computing rank likelihood functions

３．学会等名
Differential Systems: from theory to computer mathematics（国際学会）

４．発表年
2019年

１．発表者名
Tomonari Sei

２．発表標題
Admission decision using the maximum entropy principle

３．学会等名
Entropy 2018（国際学会）

４．発表年
2018年

１．発表者名
Tomonari Sei

２．発表標題
An objective general index for separation

３．学会等名
IMS-APRM 2018（国際学会）

４．発表年
2018年

１．発表者名
清 智也

２．発表標題
高次元大標本における客観的総合指数の不一致性

３．学会等名
大規模統計モデリングと計算統計V

４．発表年
2018年

１．発表者名
清 智也

２．発表標題
A general index for admission decisions

３．学会等名
統計関連学会連合大会

４．発表年
2018年

１．発表者名
Tomonari Sei

２．発表標題
Inconsistency of diagonal scaling under high-dimensional limit

３．学会等名
International Symposium on Statistical Theory and Methodology for Large Complex Data（国際学会）

４．発表年
2018年

１．発表者名
清 智也

２．発表標題
A possible extension of regression analysis for imbalanced binary data

３．学会等名
RIMS共同研究「最尤法とベイズ法」

４．発表年
2019年

１．発表者名
Tomonari Sei

２．発表標題
Regression analysis for imbalanced binary data: an answer to Prof. Sibuya's questions

３．学会等名
Pioneering Workshop on Extreme Value and Distribution Theories -- In Honor of Professor Masaaki Sibuya（国際学会）

４．発表年
2019年

１．発表者名
清 智也

２．発表標題
多次元確率分布のスケーリング問題

３．学会等名
日本数学会2017年度秋季総合分科会（招待講演）

４．発表年
2017年

１．発表者名
Tomonari Sei

２．発表標題
Coordinate-wise transformation and Stein-type densities

３．学会等名
3rd Conference on Geometric Science of Information (GSI2017)（国際学会）

４．発表年
2017年

１．発表者名
Ushio Tanaka and Tomonari Sei

２．発表標題
How does the textile set describe geometric structures of data?

３．学会等名
IASC-ARS/NZSA 2017（国際学会）

４．発表年
2017年

１．発表者名
清 智也

２．発表標題
客観的総合指数の漸近的性質

３．学会等名
RIMS共同研究「Statistical Inference and Modelling」

４．発表年
2018年

１．発表者名 Kazuya Matsumoto and Tomonari Sei
２．発表標題 Computation of the rank likelihood of Gaussian copula models by the holonomic gradient method
３．学会等名 Current Topics on Algebraic Statistics and Related Fields（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	田中　潮 (Tanaka Ushio) (60516897)	大阪府立大学・理学（系）研究科（研究院）・助教 (24403)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関