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　６ 年　 ５ 月 １８ 日現在

機関番号：１５４０１
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１９７５７

研究課題名（和文）最適輸送理論に基づく補助変数を用いた統計的推測

研究課題名（英文）Statistical inference using auxiliary variables based on optimal transportation

研究代表者
伊森　晋平（Imori, Shinpei）

広島大学・先進理工系科学研究科（理）・准教授

研究者番号：８０７４７３４５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：最適輸送理論（Wasserstein距離）に基づいた補助変数の活用に関する研究を行った．特に，完全データにおける混合分布に対するWasserstein距離を上から評価する式を導出した．また，フレシェ距離に基づくデータ（分布）の分類問題を考え，補助変数を追加した際のフレシェ距離の性質やフレシェ距離の推定量の収束レートを求めた．さらに，共変量シフト下での貪欲法やガンマダイバージェンスに基づく貪欲法の収束レートを導出した．

研究成果の学術的意義や社会的意義
有用な補助変数を活用することで主要変数の推定精度の向上が見込まれるため，その理論的な性質の解明は重要である．本研究で導出したWasserstein距離の評価式やフレシェ距離の推定量の収束レートは，これまでの補助変数の活用では考えられていなかったため，今後の補助変数の活用，選択手法の開発の基盤になりうると期待される．また，貪欲法に関する研究内容は高次元データにおける効率的な解析に役立つと考えられる．

研究成果の概要（英文）：Use of auxiliary variables based on the optimal transportation (Wasserstein distance) is studied. An upper evaluation of Wasserstein distance between two mixture distributions in complete data is developed. A property of the Frechet distance when using auxiliary variables and the convergence rate of estimators of the Frechet distance are investigated. Moreover, convergence rates of greedy algorithms under the covariate shift or based on the Gamma divergence are derived.

研究分野：数理統計学

キーワード：補助変数　最適輸送理論　フレシェ距離　Greedy algorithm　共変量シフト　ガンマダイバージェンス

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

補助変数を用いた際の主要変数に対する統計的推測を考える．有用な補助変数を利用することで主要変数の母集団パラメータの推定精度は向上しうるが，そうではない補助変数の利用は悪化を招くことが知られている．先行研究の Imori and Shimodaira (2019, Entropy)では，不完全データにおいて補助変数および主要変数の枠組みを定めており，補助変数は訓練データでは観測されるが，テストデータでは観測されない（あるいは解析に用いない）変数として捉えることができる．さらに，Imori and Shimodaira (2019)では，完全データの Kullback-Leibler ダイバージェンスに基づくリスク関数によってテストデータの予測分布の良さを測り，その漸近不偏推定量として情報量規準を導出し，有用な補助変数の選択を試みている．

しかしながら，有用な補助変数の持つ性質は十分に明らかになっておらず，また，情報量規準を用いた補助変数の選択は，大規模データ，特に，変数の数が多い高次元データにおいて，計算量の観点から適していないという問題点がある．

２．研究の目的

有用な補助変数の活用および選択を効率的に実施するためには，補助変数と主要変数の関連が解析結果に与える影響の仕組みを明らかにすることが肝要である．そこで，近年，機械学習分野で活発に研究が行われている最適輸送理論に基づき，補助変数を用いた主要変数の統計的推測に関する研究を行い，大規模データに適用可能で理論的にも妥当である，有用な補助変数の活用方法や選択手法を構築することが本研究の目的である．

３．研究の方法

(1) 混合分布における Wasserstein 距離に関する研究：

混合分布はラベルを潜在変数とみなすことで不完全データと捉えることができる．ラベルを含む完全データの解析を行うことに興味がある状況を考える．Imori and Shimodaira (2019)におけるシミュレーションで示されているように，このような状況下では，有用な補助変数を活用することで，主要変数に関する未知パラメータの推定精度を向上させることができる．

Imori and Shimodaira (2019)では予測分布の良さを測る尺度として Kullback-Leibler ダイバージェンスが用いられている．一方で，最適輸送理論において，分布間の距離を測る尺度として用いられる Wasserstein 距離を利用し予測分布の良さを測ることができれば，良い補助変数とはどのようなものであるかを明らかにするための手助けになることが期待できる．

(2) Fréchet（フレシェ）距離に基づく分類問題に関する研究：

補助変数の良さは扱う問題や解析手段に依存すると考えられる．そこで，先行研究の枠組みとは異なる，Fréchet 距離に基づくデータ（分布）の分類問題を考える．Fréchet 距離は 2 つの正規分布間の距離を表す尺度であり，Wasserstein 距離の特別な場合であると考えることができる．ターゲットとなるデータの従う分布が 2 つの正規分布のどちらに近いかを Fréchet 距離で測ることによって，分類を行うことができる．このような状況下における補助変数の有用性および選択方法について研究を行う．Fréchet 距離を用いて分類問題を行う場合に，補助変数を活用することで Fréchet 距離がどのように変化するのかを明らかにすることが重要である．

p 次元正規分布 $N_p(\mu_1,\Sigma_1)$ と $N_p(\mu_2,\Sigma_2)$ 間の Fréchet 距離（の平方）は次のように明示的に表すことができる(Dowson and Landou, 1982; Journal of Multivariate Analysis):

$$F(\mu_1,\Sigma_1,\mu_2,\Sigma_2)^2 = \|\mu_1-\mu_2\|^2 + \mathrm{tr}\left(\Sigma_1+\Sigma_2-2(\Sigma_1\Sigma_2)^{1/2}\right).$$

平均ベクトル μ_1,μ_2 と分散共分散行列 Σ_1,Σ_2 は母集団分布の未知パラメータであるため，実際のデータ解析では，これらを推定する必要がある．そこで，未知パラメータ $(\mu_1,\mu_2,\Sigma_1,\Sigma_2)$ を推定量 $(\hat{\mu}_1,\hat{\mu}_2,\hat{\Sigma}_1,\hat{\Sigma}_2)$ で置き換えた際の Fréchet 距離の推定量がどのようなスピードで真値に収束するのかを明らかにする．

(3) 高次元データに適用可能な貪欲法に関する研究：

高次元データに適用するためには情報量規準の総当たり法などでは理論的，計算量的な観

点から実用に不向きであると考えられる．高次元データに対する有効なアプローチの一つとして，適当な基準の下でステップごとにモデルを更新する貪欲法(greedy algorithm)が知られている．そのうちの一つとして，orthogonal greedy algorithm (OGA)が知られており，高次元統計解析における変数選択手法および予測手法として盛んに研究されている．

先行研究の理論結果を，補助変数を用いた解析の枠組みに直接適用することは困難であるため，関連研究として，訓練データとテストデータで説明変数が従う分布の違いに着目した共変量シフト(Shimodaira, 2000; Journal of Statistical Planning and Inference)下での OGA や，外れ値に対してロバストな推定結果を与えることができるガンマダイバージェンス(Fujisawa and Eguchi, 2008; Journal of Multivariate Analysis) に基づく貪欲法に関する研究を行う．

４．研究成果

(1) 混合分布に対し，完全データの Wasserstein 距離の明示的な表現を計算することは困難であったため，上から評価した式の導出を試みた．

2 群の混合分布に従う確率変数を $Y \in \mathbb{R}$,混合分布のラベルを表す確率変数を $Z \in \{0,1\}$ とし，完全データの確率変数を $X = (Y,Z)$ とする．ここで，Y は観測され，Z は潜在変数であり観測されないことに注意する．X の従う確率分布として，p_X と q_X の二つを考えるとき，p_X と q_X 間の Wasserstein 距離 $W(p_X, q_X)$ を評価する．

本研究では，完全データの Wasserstein 距離の上からの評価式を導出した．X の従う確率分布が p_X のとき，$Z = z \in \{0,1\}$ を条件づけた下で，Y の従う分布は $p_Y^{(z)}$ に従うものとする．同様に，X の従う確率分布が q_X のとき，$Z = z \in \{0,1\}$ を条件づけた下で，Y の従う分布を $q_Y^{(z)}$ とする．本研究で導出した評価式では $p_Y^{(a)}$ と $q_Y^{(b)}$ 間の Wasserstein 距離 $W\left(p_Y^{(a)}, q_Y^{(b)}\right)$ $(a,b \in \{0,1\})$ および混合分布の構成比率（つまり，$Z = 1$ となる確率）が用いられる．したがって，$p_Y^{(a)}$ と $q_Y^{(b)}$ を推定することができれば，この評価式を推定することが可能となる．例えば，$p_Y^{(a)}$ と $q_Y^{(b)}$ がそれぞれ正規分布であり，$W\left(p_Y^{(a)}, q_Y^{(b)}\right)$ が Fréchet 距離であれば，それぞれの平均ベクトルや分散共分散行列の推定量を用いて $W\left(p_Y^{(a)}, q_Y^{(b)}\right)$ を計算することができるようになる．

しかしながら，この評価式の近似精度などの妥当性については，今後の課題として残されている．

(2) Fréchet 距離を用いた分類問題において，補助変数の活用がどのように影響するのかについて研究を行った．分布間の距離が大きくなるほど分類しやすくなると考えられるが，正規分布の次元が大きくなるほど，母集団分布間の Fréchet 距離が大きくなる単調性が見受けられた．すなわち，補助変数を活用することで，データの分類精度が向上することが示唆されている．

実際の利用では，Fréchet 距離を構成する平均ベクトルと分散共分散行列を推定する必要があるため，その推定精度とのトレードオフで有用な補助変数が定まるものと考えられる．そこで，それぞれ標本平均ベクトルと標本分散共分散行列に置き換えた Fréchet 距離の推定量を考え，その収束レートを導出した．その結果，正規分布の次元がサンプルサイズに比べて十分に小さい場合は収束することが示された．しかし，この収束レートは，正規分布の次元がサンプルサイズと比べてある程度大きいような高次元データにおいては有用ではない．

この原因の一つは高次元データにおける標本平均ベクトルや標本分散共分散行列の推定精度にある．そこで，近年盛んに研究が行われている，高次元データにおいても妥当な収束レートを持つ推定方法に着目し，適当な収束レートを持つ平均ベクトルと分散共分散行列の推定量の使用を考えることで，Fréchet 距離の推定量の収束レートを改善することが可能となった．

(3) 真の構造を誤特定している高次元線形回帰モデルにおいて，共変量シフトを考えた場合の OGA に関する研究を行った．Shimodaira (2000)などの先行研究では，訓練データとテストデータでの説明変数の従う分布の違いに着目してサンプルに重みをつけることで推定量の補正を行っており，そのような加重推定に基づく OGA の修正版や通常の OGA の適用を考える．これらの手法に関して，共変量シフト下での収束レートを導出した．結果として，適当な条件下においては，共変量シフト下であったとしても，OGA の修正版が良い収束レートを達成することがわかった．

また，分布間の違いを測る尺度の一つであるガンマダイバージェンスに基づく推定量の高次元データにおける漸近挙動を導出した．そしてこの結果を用いて，ガンマダイバージェンスに基づく貪欲型アルゴリズムの収束レートを導出した．

Wasserstein 距離や Fréchet 距離に基づく貪欲法に関する研究には至っておらず，この点は今後の課題である．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Imori Shinpei	４．巻 -
２．論文標題 Asymptotic Optimality of Cp-Type Criteria in High-Dimensional Multivariate Linear Regression Models	５．発行年 2023年
３．雑誌名 Statistica Sinica	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） 10.5705/ss.202020.0425	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 －

〔学会発表〕　計9件（うち招待講演　3件／うち国際学会　4件）

１．発表者名 Shinpei Imori
２．発表標題 Weighted orthogonal greedy algorithm for prediction under covariate shift
３．学会等名 2023 International Conference for Statistics and Data Science (2023 ICSDS)（招待講演）（国際学会）
４．発表年 2023年

１．発表者名 Shinpei Imori
２．発表標題 Importance weighted orthogonal greedy algorithm with estimated weight function
３．学会等名 The 6th International Conference on Econometrics and Statistics (EcoSta 2023)（招待講演）（国際学会）
４．発表年 2023年

１．発表者名 Shinpei Imori
２．発表標題 On classification problem based on Frechet distance with auxiliary variables
３．学会等名 The Institute for Mathematical Statistics - Asia-Pacific Rim Meeting (IMS-APRM 2024（国際学会）
４．発表年 2024年

１．発表者名
伊森晋平

２．発表標題
高次元多変量線形回帰モデルにおける変数選択について

３．学会等名
日本数学会2022年度秋季総合分科会（招待講演）

４．発表年
2022年

１．発表者名
伊森晋平，若木宏文

２．発表標題
Frechet距離を用いた分類問題について

３．学会等名
科研費シンポジウム「多様な分野における統計科学の理論とその応用」

４．発表年
2022年

１．発表者名
伊森晋平

２．発表標題
Frechet距離に基づく分類と補助変数について

３．学会等名
研究集会『多変量統計学・統計的モデル選択の新展開』

４．発表年
2023年

１．発表者名
伊森 晋平，橋本 真太郎，Ching-Kang Ing

２．発表標題
外れ値に対して頑健な貪欲型変数選択手法について

３．学会等名
2021 年度統計関連学会連合大会

４．発表年
2021年

１．発表者名 Shinpei Imori
２．発表標題 Variable selection in high-dimensional multivariate linear regression models with group structure
３．学会等名 The 4th International Conference on Econometrics and Statistics (EcoSta 2021)（国際学会）
４．発表年 2021年

１．発表者名 伊森 晋平, 橋本 真太郎
２．発表標題 Gamma-divergence に基づく変数選択について
３．学会等名 科研費シンポジウム（機械学習・統計学・最適化の数理と AI 技術への展開）
４．発表年 2020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

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			
台湾	National Tsing Hua University			