多変量分布間の回帰モデル

東京大学経済学研究科　岡野遼

東京大学総合文化研究科　今泉允聡

1　はじめに

ユーグリッド空間に値を取らない複雑なデータに対して, それらを用いて統計分析をする方法を研究することは統計学において長らく関心がもたれてきた. 例えば, 関数データや多様体値データはそのような複雑なデータの代表例である. 前者は無限次元空間に値をとるデータであり, 後者は非線形制約を満たすデータであるという点で, それぞれユーグリッド空間に値をとる通常のデータと異なっている. このような複雑なデータを扱って統計分析をする際には, それぞれのデータの特徴を考慮してモデルや手法の設計を行う必要がある.

近年, このような複雑データの一種として, 確率分布に値をとるデータの解析が注目を集めている. 例えば, ある年のある国における年齢別死亡率は, 横軸を年齢, 縦軸を死亡率とする密度関数として表示することで, 一次元分布に値をとるデータとみなすことができる. このような確率分布値のデータは関数データのように無限次元空間に値をとりつつも, 密度関数であるための非線形制約を満たす必要がある点で従来の複雑データにはない特徴をもっており, それを扱うための新たな統計的手法の研究が近年盛んに行われている. 例えば, [2], [3] は, 説明変数と結果変数がともに分布の形で与えられるような回帰（分布間回帰）モデルをそれぞれ提案し, 世界 37 カ国の年齢別死亡率分布を用いた実証研究を行っている. また, [1] では確率分布値データに対する主成分分析手法, [5] では確率分布値データに対する相関指標, [9] では確率分布値時系列データに対する自己回帰モデルの提案がそれぞれ行われている.

本研究では, 確率分布値データ間の回帰モデルとして, 新たなものを提案する. 分布間の回帰モデルは [2], [3] で提案されているが, それらは分布の次元を一次元の場合に限定している. これらのモデルは, 一次元分布間の最適輸送問題が陽に解けるという事実に強く依存しており, 一般に最適輸送問題が陽に解けない多変量分布の場合にそのまま拡張することは困難である. 本研究では, 同一の location-scale 分布族に属する分布間の最適輸送問題が, 多変量分布であっても例外的に陽な解を持つことに注目し, 多変量分布の場合でも適用可能な, 新たな分布間回帰モデルを提案する. 以下では, 確率分布値のデータをモデリングする際に必要となる最適輸送問題と Wasserstein 空間についてまず第二節で紹介し, その後第三節で提案するモデルの概要を述べる.

2　最適輸送問題と Wasserstein 空間

確率分布値のデータを扱う際に必要となる最適輸送問題と Wasserstein 空間の基本事項について述べる. 最適輸送問題の数学的側面に関しては [8] が詳しく, 計算アルゴリズム的側面に関しては [6] が詳しい. また, Wasserstein 空間に関しては, [4] にまとめられている.

2.1 最適輸送問題

$\mathcal{W}$ を $\mathbb{R}^d$ 上の確率分布で有限な二次モーメントを持つもの全体とする. 二つの確率分布 $\mu, \nu \in \mathcal{W}$ に対して, それらの間の最適輸送問題を,

$$\inf_{T:T\#\mu=\nu} \int_{\mathbb{R}^d} \{T(x) - x\}^2 d\mu(x) \tag{1}$$

により定義する. ただし, # は $T\#\mu(A) = \mu(T^{-1}(A))$ で定まる測度の押し出しを表している. $X \sim \mu, Y \sim \nu$ なる確率ベクトル X, Y を用いると, 問題 (1) は,

$$\inf_{T:T(X)\sim Y} \mathbb{E}[\{T(X) - X\}^2] \tag{2}$$

と言い換えることもできる. 最適輸送問題は, 分布 μ の質量を分布 ν に輸送するのに最もコストの少ない方法を求める問題と解釈できる. 今, 分布 μ, ν は絶対連続（つまり, 密度関数を持つ）ことを仮定すると, 最適輸送問題 (1) 及び (2) は一意な解を持つことが保証される. その解を最適輸送写像と呼び, t_μ^ν と表すことにする. 分布が一次元（つまり, $d = 1$）の場合, 最適輸送写像 t_μ^ν は μ の分布関数 F_μ と ν の分位点関数 F_ν^{-1} を用いて, $t_\mu^\nu = F_\nu^{-1} \circ F_\mu$ と陽に表せることが知られている. 一方, 分布が多変量（つまり, $d \geq 2$）の場合には, 一般に最適輸送写像を陽に表すことはできない.

最適輸送写像 t_μ^ν を用いて, 確率分布 $\mu, \nu \in \mathcal{W}$ 間の Wasserstein 距離を

$$d_W(\mu, \nu) := \sqrt{\int_{\mathbb{R}^d} \{T_\mu^\nu(x) - x\}^2 d\mu(x)}$$

により定める. Wasserstein 距離 d_W は数学的な意味で $\mathcal{W}$ 上の距離になっており, 分布のサポートの距離構造を捉えられるという特徴を持つ. 分布が一次元の場合には, 最適輸送写像の陽な表現から, $\mu, \nu \in \mathcal{W}$ 間の Wasserstein 距離は, その分位点関数分位点関数 F_μ^{-1}, F_ν^{-1} を用いて,

$$d_W(\mu, \nu) = \sqrt{\int_0^1 \{F_\mu^{-1}(u) - F_\nu^{-1}(u)\}^2 du}$$

と表すことが可能である.

2.2 Wasserstein 空間

上で定義した確率分布の集合 $\mathcal{W}$ に, Wasserstein 距離 d_W を与えてできる距離空間 $(\mathcal{W}, d_W)$ を Wasserstein 空間と呼ぶ. Wasserstein 空間は無限次元でかつ非線形な空間であるという特徴を持つ. 確率分布値のデータを扱う際には, それらを Wasserstein 空間における点とみなしてモデリングを行うことになる.

確率分布値のデータを扱う際に, 幾何的な概念を Wasserstein 空間に導入することがしばしば有用である. 絶対連続な確率分布 $\mu_* \in \mathcal{W}$ に対し, μ_* における $\mathcal{W}$ の接空間を,

$$T_{\mu_*} := \overline{\{t(t_{\mu_*}^\mu - \mathrm{id}) : \mu \in \mathcal{W}, t > 0\}}^{\mathcal{L}^2_{\mu_*}}$$

により定義する. ここで, id は恒等写像を表し, $\mathcal{L}^2_{\mu_*}$ は $\mathbb{R}^d$ 上の μ_*-二乗可積分関数全体に内積 $\langle , \rangle_{\mu_*}$ を与えてできるヒルベルト空間を表す. この接空間 T_{μ_*} は $\mathcal{L}^2_{\mu_*}$ の部分空間であることが知られている. さらに, μ_* に

おける exponential map $\exp_{\mu_*} : T_{\mu_*} \to \mathcal{W}$ を

$$\exp_{\mu_*} g = (g + \mathrm{id})\#\mu$$

によって定め, μ_* における logarithmic map $\log_{\mu_*} : \mathcal{W} \to T_{\mu_*}$ を

$$\log_{\mu_*} \mu = t^{\mu}_{\mu_*} - \mathrm{id}$$

によって定める. これらの map によって, Wasserstein 空間 $\mathcal{W}$ の点とその接空間 T_{μ_*} の点との間の対応が与えられる.

3 研究内容

$\mathcal{W}$ を $\mathbb{R}^d$ 上の確率分布で有限な 2 次モーメントを持つもの全体とし, $\mathcal{W}$ には Wasserstein 距離 d_W が与えられているとする. $\mathcal{F}$ を $\mathcal{W} \times \mathcal{W}$ 上の同時分布とし, $(\nu_1, \nu_2) \sim \mathcal{F}$ とする. 分布間回帰は ν_1 を説明変数, ν_2 を結果変数とするような Wasserstein 空間の間の回帰問題として定式化される.

分布間回帰のモデルは近年主に [2] と [3] で一次元分布（$d = 1$）の場合に提案されている. 具体的には, [2] は Wasserstein 空間の幾何を用いて, 確率分布を制約のない関数に変換し, 分布値データ間の回帰を関数値データ間の回帰に帰着させるというアプローチをとっている. また, [3] では, 一次元分布間の最適輸送写像はそれが単調増加であることで特徴づけられるという事実に基づき, 回帰関数が $\mu \mapsto T\#\mu$（T は未知の単調増加関数）という形であると仮定する分布間回帰モデルを提案している. しかし, これらのモデルは, 一次元分布間の最適輸送問題が陽に解けることとその解の形に強く依存しており, 一般に最適輸送問題が陽に解けない多変量分布（$d \geq 2$）の場合にそのまま拡張することは困難である. 本研究では, 多変量分布に対しても利用可能な新たな分布間回帰モデルを作ることを目標とする.

3.1 キーアイデア：同一の location-scale 族への制限

本研究では, 同一の location-scale 分布族に属する分布間の最適輸送問題が, 多変量分布であっても例外的に陽な解を持つことに注目する. P を $\mathbb{R}^d$ 上の絶対連続な分布で, $X \sim P$ とする. あるベクトル $a \in \mathbb{R}^d$ と行列 $B \in \mathbb{R}^{d \times d}$ を用いて, $a + BX$ と表されるような確率ベクトルが従う分布全体を, P により生成される location-scale 族といい, $\mathcal{G}_P$ と書くことにする. 例えば, P が標準ガウス分布であれば, $\mathcal{G}_P$ はガウス分布全体となる. また, 分布 $\mu \in \mathcal{G}_P$ が平均 m, 共分散行列 Σ を持つ場合, $\mu = P(m, \Sigma)$ と表すことにする. すると, $\mathcal{G}_P$ に属する非退化な共分散行列を持つ二つの分布 $\mu_1 = P(m_1, \Sigma_1), \mu_2 = P(m_2, \Sigma_2)$ の間の最適輸送写像 $t^{\mu_2}_{\mu_1}$ 及び Wasserstein 距離 $d_W(\mu_1, \mu_2)$ はそれぞれ以下のように具体的に表されることが知られている：

$$t^{\mu_2}_{\mu_1}(x) = m_2 + \Sigma_1^{-1/2}[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}]^{1/2}\Sigma_1^{-1/2}(x - m_1),$$
$$d_W(\mu_1, \mu_2) = \sqrt{\|m_1 - m_2\|^2 + \mathrm{tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}]}.$$

また, 最適輸送写像が陽に表されることから, Wasserstein 空間を location-scale 族に制限して得られる空間 $(\mathcal{G}_P, d_W)$ について, $(\mathcal{G}_P, d_W)$ とその接空間との間の対応を陽に与えることが可能である. 以下で述べる空間 $(\mathcal{G}_P, d_W)$ の幾何的な性質については, [7] に基づいている. 簡単のため, P を $\mathbb{R}^d$ 上の平均 0 の絶対連続な分布とし, P によって生成される平均 0 の scale 族を $\mathcal{G}_P$ とする. また, $\mathrm{Sym}(d)$ をサイズ $d \times d$ の対称行列全体とし, $\mathrm{Sym}^+(d)$ をサイズ $d \times d$ の半正定値対称行列全体とする. すると, まず $(\mathcal{G}_P, d_W)$ の $P(0, \Sigma^*)$ における接

空間は, 内積空間 $T_{\Sigma_*} = (\mathrm{Sym}(d), G_{\Sigma_*})$ となる. ここで, 内積 G_{Σ_*} は

$$G_{\Sigma_*}(V, W) = \mathrm{tr}(V\Sigma_* W), \quad V, W \in \mathrm{Sym}(d)$$

により与えられる. また, expenential map $\exp_{\Sigma_*} : T_{\Sigma_*} \to \mathcal{G}_P$ は

$$\exp_{\Sigma_*} V = P(0, (V + I)\Sigma_*(V + I)), \quad V \in T_{\Sigma_*}, \tag{3}$$

で与えられ, logarithmic map $\log_{\Sigma_*} : \mathcal{G}_P \to T_{\Sigma_*}$ は

$$\log_{\Sigma_*} P(0, \Sigma) = \Sigma_*^{-1/2}[\Sigma_*^{1/2}\Sigma\Sigma_*^{1/2}]^{1/2}\Sigma_*^{-1/2} - I, \quad P(0, \Sigma) \in \mathcal{G}_P. \tag{4}$$

で与えられる.

3.2 提案するモデル

P_1 及び P_2 を $\mathbb{R}^d$ 上の絶対連続な確率分布とし, P_1, P_2 により生成される location-scale 族をそれぞれ $\mathcal{G}_{P_1}, \mathcal{G}_{P_2}$ とする. $\mathcal{F}$ を $\mathcal{G}_{P_1} \times \mathcal{G}_{P_2}$ 上の同時分布とし, $(\nu_1, \nu_2) \sim \mathcal{F}$ とする. 以下では, ν_1 を説明変数, ν_2 を結果変数とするような, $\mathcal{G}_{P_1}$ から $\mathcal{G}_{P_2}$ への回帰を考える. [2] と同様に空間 $(\mathcal{G}_{P_1}, d_W)$ 及び $(\mathcal{G}_{P_2}, d_W)$ の幾何を用いて, 確率分布 ν_1 及び ν_2 をそれぞれ非線形制約のない行列に変換し, 分布値データ間の回帰を行列データ間の回帰に帰着させるというアプローチを取る.

$j = 1, 2$ に対し, ν_j の $\mathcal{G}_{P_j}$ における Fréchet 平均 $\nu_{j\oplus} = P(m_{j\oplus}, \Sigma_{j\oplus})$ を以下で定義する：

$$\nu_{j\oplus} := \underset{\mu \in \mathcal{G}_{P_j}}{\arg\min}\, \mathbb{E} d_W^2(\mu, \nu_j).$$

次に, $j = 1, 2$ に対し, 空間 $(\mathcal{G}_{P_j}, d_W)$ の幾何を用いて確率分布 $\nu_j = P_j(m_j, \Sigma_j)$ を非線形制約のない行列に変換する. 具体的には, Fréchet 平均 $\nu_{j\oplus} = P(m_{j\oplus}, \Sigma_{j\oplus})$ において $(\mathcal{G}_{P_j}, d_W)$ の接空間をはり, $P_j(m_j, \Sigma_j) \mapsto (m_j, \log_{\Sigma_{j\oplus}} P(0, \Sigma_j))$ という変換を行う. ここで, $\log_{\Sigma_{j\oplus}} P(0, \Sigma_j)$ は (4) により与えられる. 行列の集合 S_d を $S_d = \{X = (a, B), a \in \mathbb{R}^d, B \in \mathrm{Sym}(d)\}$ で定め, 変換後の行列を $X = (m_1, \log_{\Sigma_{1\oplus}} P(0, \Sigma_1)), Y = (m_2, \log_{\Sigma_{2\oplus}} P(0, \Sigma_2))$ とおくことにする. このような変換を行うことで, $\nu_1 \in \mathcal{G}_{P_1}$ から $\nu_2 \in \mathcal{G}_{P_2}$ という分布間の回帰は, $X \in S_d$ から $Y \in S_d$ という非線形制約のない行列間の回帰に帰着できる. 我々は, 行列のペア (X, Y) の間に以下のような線形モデルを仮定する：

$$Y = \langle X, \mathbb{B} \rangle + E, \quad \mathbb{E}[E|X] = 0. \tag{5}$$

ここで, $E \in S_d$ は誤差を表す行列であり, $\mathbb{B} \in \mathbb{R}^{d\times(d+1)\times d\times(d+1)}$ は回帰係数に相当する 4 次テンソルである. また, $\langle X, \mathbb{B} \rangle \in \mathbb{R}^{d\times(d+1)}$ は, (q_1, q_2) 成分が以下で与えられるような, X と $\mathbb{B}$ の間の contracted tensor product である

$$\langle X, \mathbb{B} \rangle_2[q_1, q_2] = \sum_{p_1=1}^{d} \sum_{p_2=1}^{d+1} X[p_1, p_2]\mathbb{B}[p_1, p_2, q_1, q_2].$$

テンソル $\mathbb{B}$ に何も制約を課さない場合, $\mathbb{B}$ の要素数 (すなわち, モデルのパラメータ数) は分布の次元 d に応じて非常に大きくなってしまい, over-parametrize の問題が発生する. これを回避するため, $\mathbb{B}$ に低ランク性を仮

定する. すなわち, R を小さな自然数とし, A_1, A_2, A_3, A_4 をそれぞれサイズ $d\times R, (d+1)\times R, d\times R, (d+1)\times R$ の行列として, $\mathbb{B}$ がこれらの行列を用いて

$$\mathbb{B} = [\![A_1, A_2, A_3, A_4]\!] := \sum_{r=1}^{R} a_{1r} \circ a_{2r} \circ a_{3r} \circ a_{4r} \tag{6}$$

と分解されることを仮定する. ここで, a_{kr} は行列 A_k の第 r 列であり, $\circ$ は外積を表す. このように $\mathbb{B}$ が低ランク R を持つことを仮定することで, パラメータ数を $d^2(d+1)^2$ から $(4d+2)R$ へと抑えることができている. (6) の分解が一意となるための仮定を A_1, A_2, A_3, A_4 に課すと, このモデルにおいて関心のあるパラメータは

$$\theta_0 = (\mathrm{vec}(A_1)^\top, \mathrm{vec}(A_2)^\top, \mathrm{vec}(A_3)^\top, \mathrm{vec}(A_4)^\top)^\top$$

となる. 最後に, 確率 1 で $\langle X, \mathbb{B}\rangle \in \mathbb{R}^d \times \log_{\Sigma_{2\oplus}}(\mathcal{G}_0)$ となることをモデルに仮定する.

3.3 推定量の構成法とその理論的性質

これまでは各確率分布が完全に観測されることを暗に仮定してきたが, 実際にはそのような状況は稀で, 我々は各分布からの離散観測のみが得られる場合がほとんどである. 従って, 以下では次のような二段階のデータ生成メカニズムを仮定して, パラメータの推定を行う：

- まず確率分布のペア $(\nu_{1i}, \nu_{2i}) \sim (\nu_1, \nu_2), i = 1, ..., n$ が潜在的に独立に生成される
- 次に各分布 ν_{ji} から d 次元確率ベクトル $W_{jir}, r = 1, ..., N$ が観測される

従って我々は, 観測値 $W_{jir} \sim \nu_{ji}, j = 1, 2, i = 1, ..., n, r = 1, ..., N$ に基づいて, パラメータ

$$\theta_0 = (\mathrm{vec}(A_1)^\top, \mathrm{vec}(A_2)^\top, \mathrm{vec}(A_3)^\top, \mathrm{vec}(A_4)^\top)^\top$$

の推定を行うことになる.

推定量は以下の手順で構成する.

1. 離散観測 $W_{jir}, r = 1, ..., N$ を用いて, 各分布 $\nu_{ji} = P_j(m_{ji}, \Sigma_{ji})$ の平均ベクトルの推定値 $\hat{m}_{ji}$ と共分散行列の推定値 $\hat{\Sigma}_{ji}$ を構成し, $\hat{\nu}_{ji} = P_j(\hat{m}_{ji}, \hat{\Sigma}_{ji})$ とおく. 平均ベクトルと共分散行列の推定においては, 最尤推定量を用いる.
2. $j = 1, 2$ に対し, 推定された分布 $\hat{\nu}_{ji}, i = 1, ..., N$ を用いて, 経験 Fréchet 平均

$$\hat{\nu}_{j\oplus} = N(\hat{m}_{j\oplus}, \hat{\Sigma}_{j\oplus}) := \operatorname*{arg\,min}_{\mu \in \mathcal{G}_{P_j}} \frac{1}{N} \sum_{i=1}^{N} d_W^2(\mu, \hat{\nu}_{ji})$$

   を計算する. この最適化問題は, [4] の Section 5.4.1 で紹介されている steepest descent アルゴリズムを用いて容易に解くことができる.
3. $j = 1, 2$ に対し, 経験 Fréchet 平均 $\hat{\nu}_{j\oplus}$ において $(\mathcal{G}_{P_j}, d_W)$ の接空間をはり, 推定された分布 $\hat{\nu}_{ji}$ を行列に変換する：

$$\hat{X}_i = (\hat{m}_{1i}, \log_{\hat{\Sigma}_{1\oplus}} P(0, \hat{\Sigma}_{1i})), \quad \hat{Y}_i = (\hat{m}_{2i}, \log_{\hat{\Sigma}_{2\oplus}} P(0, \hat{\Sigma}_{2i})).$$

4. 最小二乗法により, パラメータの推定量を計算する：

$$\hat{\theta}_{n,N} = \operatorname*{arg\,min}_{\theta \in \Theta} \hat{M}_{n,N}(\theta), \quad \hat{M}_{n,N}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|\hat{Y}_i - \langle \hat{X}_i, \mathbb{B}(\theta)\rangle\|_F^2.$$

ここで, Θ はモデルの条件を満たすようなパラメータからなるある集合である. また, $\|\cdot\|_F$ はフロベニウスノルムを表す.

この推定量の理論的性質に関しては, 真の Fréchet 平均 $\nu_{1\oplus}, \nu_{2\oplus}$ が既知の状況のもとで, 以下の収束レートと漸近正規性の結果が得られている.

Theorem 1　真の Fréchet 平均 $\nu_{1\oplus}, \nu_{2\oplus}$ が既知であると仮定する. この時, いくつかの正則条件の下で, $\|\hat{\theta}_{n,N} - \theta_0\|_F = O_p(n^{-1/2} + N^{-1/4})$ が成立. さらに, N は n の数列で, $N(n) = n^q (q > 2)$ であれば, ある共分散行列 V が存在して, $n \to \infty$ のとき $\sqrt{n}(\check{\theta}_{n,N} - \theta_0) \xrightarrow{d} N(0, V)$ となる.

当日では, 推定に関するさらなる理論的結果と, 数値実験・実データ解析の例なども報告する予定である.

参考文献

[1] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic pca in the wasserstein space. *arXiv preprint arXiv:1307.7721*, 2013.

[2] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

[3] Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*.

[4] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[5] Alexander Petersen and Hans-Georg Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.

[6] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[7] Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[8] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[9] Chao Zhang, Piotr Kokoszka, and Alexander Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52, 2022.