【研究ノート】

3群以上にも適用可能なデータ分割の指標について

田　中　研太郎

概要

本論文では，各群がなるべく似たものになるようにデータをいくつかの群に分割する，という問題を扱う。各群の等質性を測る指標として，最適輸送における考え方と線形判別の手法を組み合わせたものを提案する。提案する指標は，群が3つ以上の場合にも適用可能である。簡単な例を通して指標の有効性を検証する。

1. データの分割について

本論文では，データが与えられたときに，なるべく偏りが生じないようにいくつかの群に適切にデータを分割するにはどのようにすればよいのか，という問題を扱う。統計分析や機械学習においては，判別分析などのように，分割後の各群が互いに異なる性質を持つようにデータを分割・分類することを考えることが多いが，いま扱いたい問題はそれとは逆で，各群が互いになるべく等質になるような分割を考える。そのような問題は，例えば，学生たちでグループ作業を実施するときのグループ分けのときなどに生じる。もし，学生の属性（学年，所属学部，過去の成績など）がグループ間で著しく異なると，作業の進度や完成度が必要以上にバラバラになってしまう可能性がある。また，グループ間の偏りが大きくなると，グループ内の偏りは小さくなるため，グループ内での多様性が損なわれてしまい，相互理解の体験が損なわれてしまう。他にも，チームスポーツにおける練習試合でのチーム分けにおいても，等質なグループ分けが有効な場合があると考えられる。

このようなデータ分割をしたいときによく行われる方法としては，乱数などを使って無作為にデータ分割を行う，といったものが考えられる。無作為なデータ分割を行えば，確かに平均的には未観測な変数も含めて偏りは生じないが，一方で，毎回のデータ分割が適切になるという保証はない。結果として，偏りが少なそうな等質なデータ分割が見つかるまで延々と無作為なデータ分割を行う，という作業になってしまう場合も多い。ここで問題になるのは，偏りが少なそう，というのがどういう状態を指すのかが曖昧である，ということである。それが曖昧なままだと，作業をいつ終えていいのかも判然とせず困ってしまう。そこで，田中（2022）において，最適輸送を使った指標を提案したが，そこでは2群へのデータ分割しか扱っていなかった。本論文では，最適輸送を使った指標が，3群以上のデータ分割の場合にも

拡張可能であることを紹介する。さらに，最適輸送の考え方だけでは各群の等質性を表し切れていない部分があるので，線形判別の手法を用いてその問題点を修正した指標を提案する。

2. 最適輸送を用いた等質性の指標

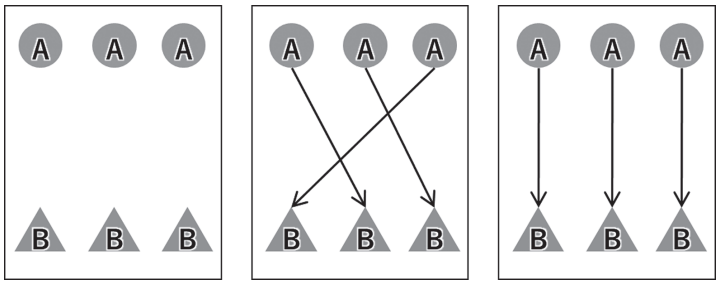

図１：(左) グループA, Bの位置関係, (中央) 最適でない輸送, (右) 最適輸送

まず，2つの群の間の最適輸送について考える。この場合の最適輸送とは，一方からもう一方へ何かを輸送するときの総コスト（総距離）が最小になるような輸送方法のことである。例えば, 図1（左）のように, 3個の点からなるグループAの各点に1単位のものがあるとして，それらを3個の点からなるグループBに1単位ずつ分配・輸送することを考える。このとき，輸送方法にはいろいろなパターンが考えられる。図1（中央）の矢印の対応関係で輸送をする方法も考えられるが，輸送コストを最小にする輸送方法は図1（右）で与えられる。次に，グループA, Bの位置関係が図2（左）の場合を考える。このとき，最適輸送は図2（右）のようになるが，その輸送の総コストは図1（右）よりも小さい。いま，図1，2のA，Bの位置関係を比べると，図2の方が混ざり合いが大きく，より等質なデータ分割になっていることが分かる。つまり，最適輸送の総コストが小さい方が，データ分割としてより等質なものになっているため，その総コストの値を等質性の指標として用いることができると考えられる。以上の話を，2つの群における最適輸送の問題として一般的に考える。2つの群における最適輸送の問題は，次のような線形計画問題として定式化されることが知られている。

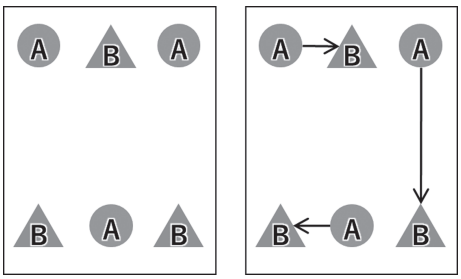

図２：(左) A, Bの位置関係, (右) 最適輸送

$$\min_{\gamma \in R_+^{m \times n}} \sum_{i,j} \gamma_{ij} M_{ij}$$
$$s.t. \quad \boldsymbol{\gamma 1} = \boldsymbol{a}, \quad \boldsymbol{\gamma}^T \mathbf{1} = \boldsymbol{b}, \quad \boldsymbol{\gamma} \geq \mathbf{0} \tag{1}$$

記号の説明を行う。まず，グループA，Bにはそれぞれm,n個の点があり，各グループの各点の設定量をベクトル$\boldsymbol{a} = (a_1, \cdots, a_m)$と$\boldsymbol{b} = (b_1, \cdots, b_n)$で表すことにする。$\boldsymbol{a}$と$\boldsymbol{b}$の要素はすべて正で,それぞれの合計は1になるように正規化されているとする。そして，グループAのi番目からグループBのj番目に1単位のものを輸送するときのコストをM_{ij}で表すことにする。また，1は要素がすべて1のベクトルを表す。この線形計画問題を解いたときの目的関数の値が最適輸送の総コストになり，その総コストの値が小さいほどグループAとBがより等質であることを表す。コストM_{ij}の決め方は色々と考えられるが，ここでは，2点間のユークリッド距離を考えることにする。式（1）の解法については，Cuturi（2013）において，Sinkhorn-Knoppアルゴリズムによる高速な近似解法が知られている。

3. データを3群以上に分割する場合の指標

ここまでの話では，データを2分割する場合の指標しか構成できていないが，式（1）を素直に拡張することで，3群以上に分割する場合の指標も構成できる。本論文では3群にデータを分割する場合だけについての記述に留めるが，より一般の個数への分割も同様の方法で可能である。

グループA，B，Cにはそれぞれl, m, n個の点があり，各グループの各点の設定量をベクトル$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$で表すとする。$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$の要素はすべて正で, それぞれの合計は1になるように正規化されているとする。そして，グループAのi番目とグループBのj番目とグループCのk番目の間で1単位のものを融通するときのコストをM_{ijk}とする。この，融通する，ということが具体的にはどのような状態を表すのかは想像しづらく，また，輸送問題という呼称を使用し続けることが適切なのかは疑問が残るが，線形計画問題としては，以下のように，式（1）を素直に拡張することができる。

$$\min_{\gamma \in R_+^{l \times m \times n}} \sum_{i,j,k} \gamma_{ijk} M_{ijk}$$
$$s.t. \quad \sum_{j,k} \gamma_{ijk} = a_i, \quad \sum_{i,k} \gamma_{ijk} = b_j, \quad \sum_{i,j} \gamma_{ijk} = c_k, \quad \boldsymbol{\gamma} \geq \mathbf{0} \tag{2}$$

コストM_{ijk}をどう決めるのかは悩ましいところだが，計算と拡張のしやすさという点から，3点の重心からの距離の和を用いることにした（図3）。ただし，解が一意に決まるという保証はない。4群以上への分割の場合も，式（2）と同様の定式化が可能である。

なお，3群以上への最適輸送問題の拡張は，multi-marginal optimal transportと呼ばれている。また，multi-marginal optimal transport においても，Sinkhorn-Knoppアルゴリズムを素直に拡張することで近似解を求めることができる。ただし，群の数が大きくなると，計算量が冪で大きくなるため，あまりにも多くの群に分けると現実的な時間内に解くことは難しくなってしまう。

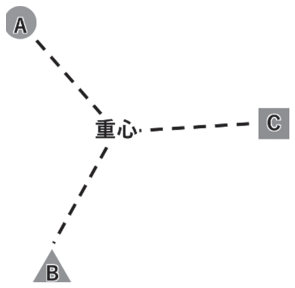

図３：コストM_{ijk}最適輸送

4. 問題点とその修正

最適輸送の総コストをデータ分割（グループ分け）の指標にすることで，3群以上に分割する場合にも容易に拡張できることは前節までに説明したが，実は，最適輸送の概念だけでは，図4のような2つのグループ分けを区別することができない。図4の左のグループ分けよりも右のグループ分けのほうが，グループAとBがより混ざり合っていて等質であると考えられる。しかし，最適輸送の総コストは，図4のどちらのグループ分けでも同じ値になってしまう（点線の長さの和はどちらのグループ分けでも同じ）。このような現象は，コスト（M_{ij}やM_{ijk}など）の取り方を変えるだけで修正することは難しい。そこで，新たな指標を提案することにする。

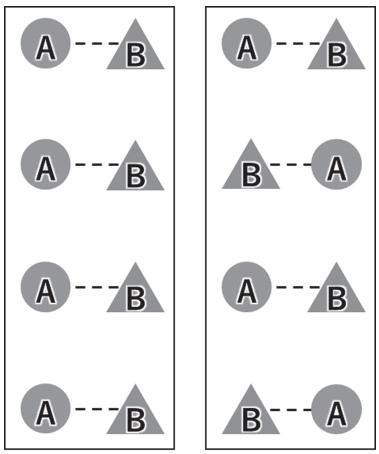

図４：最適輸送で区別できない例

まず，グループ内コストの和，平均コスト，バラつきコストを以下で定める。

- グループ内コストの和：各グループ内において重心からの各データの距離の総和（グループ内コスト）を計算し，それらの和をとったもの
- 平均コスト：各グループの重心の差異からなるコスト
- バラつきコスト：各グループ内における重心からの各データの距離の総和（グループ内コスト）の差異からなるコスト

そのうえで，以下の指標を提案する。

$$\text{提案する指標}=\frac{\text{最適輸送の総コスト}+\tau\cdot(\text{平均コスト})+\kappa\cdot(\text{バラつきコスト})}{(\text{グループ内コストの和})^{\nu}} \qquad (3)$$

ここで，τ,κ,νは，$\tau\geq 0$，$\kappa\geq 0$，$\nu\geq 0$の範囲で適当に設定するパラメーターである。この

指標を用いると，最適輸送の概念を用いつつ，図4の左と右のグループ分けのそれぞれが違う値を持つことになるので区別することができる。この指標の導入について簡単に説明する。まず，最適輸送の総コストは，線形判別における群間分散に相当し，グループ間のコストを表す量である。つまり，最適輸送だけでは線形判別における群内分散（グループ内のコスト）を考慮することができていない。そこで，提案する手法では，線形判別の手法に倣って，グループ間のコストをグループ内のコストの和で割ることにした。平均コストとバラつきコストは，その修正を補助する役割で付加している。各グループの平均やバラつきが全てお互いに等しければ，平均コストもバラつきコストもゼロになる。これらのコストも輸送コストと同様に，基準となる量を重心として設定して，各グループの量の重心からの距離の和をとることで構成している。そして，どちらもグループ間の差異を表す量なので分子に加えている。なお，グループ内コストの和，平均コスト，バラつきコストという3つの量だけでは区別することができず，最適輸送の総コストでは区別できる，という例（図5）もあるため，計算量はかかってしまうが，最適輸送の総コストはやはり必要である。

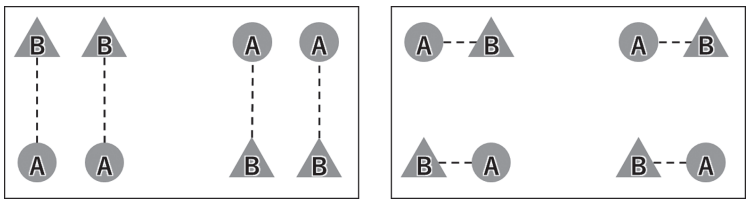

図５：最適輸送の総コストのみに依存する例

5. 数値例

式（3）の指標の使用例として，図6の左と右の3群A,B,Cへのグループ分けにおいての指標の計算結果を報告する。ここで，式（3）のパラメーターτ,κ,νは，τ=0.2, κ=0.8, ν=1という値に設定した（この設定にとくに根拠はない）。まず，最適輸送の総コストは，左のグループ分けは2.4250…，右のグループ分けは2.0である。次に，平均コストは，左は0.4807…，右は0である。そして，バラつきコストは，左は0.3023…，右は0.1199…となった。

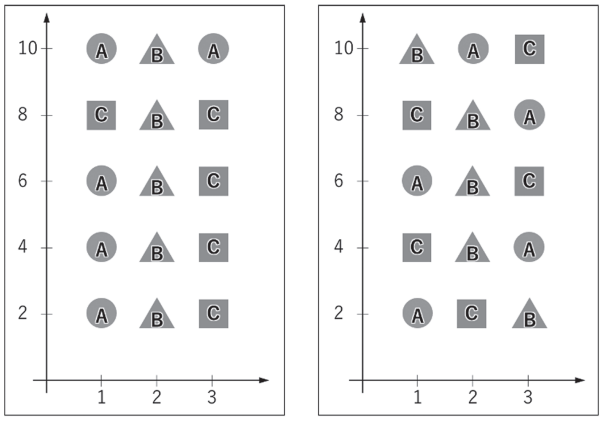

図６：指標による比較の例

また，グループ内コストの和は，左は2.5422…，右は2.6291…となった。以上より，式（3）の指標は，左が1.0869…，右が0.7972…となった。式（3）の指標（τ =0.2, κ =0.8, ν =1の場合）で比較すると，右の方が小さいので，より等質な分割であるといえる。左は横軸の値の混ざり合いが右に比べて少ないため，この結果は妥当であると考えられる。

6. まとめ

本論文では，データを分割したときの各群の等質性を表す指標を導入した。最適輸送の総コストは等質性を表す指標として有用であり，さらに，3群以上の場合への拡張も容易ではあるが，最適輸送の総コストだけでは表しきれない性質がある。それらの性質を反映するような修正を最適輸送の総コストに付け加えることで，適切な指標を構築することができたと考えられる。また，数値例によって指標の妥当性を示すことができた。ただし，コスト（M_{ij} や M_{ijk} など）の定義を，計算のしやすさと拡張性を優先して決めてしまったため，指標の構成がやや技巧的になってしまっているかもしれない。今後の課題として，より等質性の情報を反映しやすく，かつ計算も容易なバランスの取れたコストの定義について考えたい。そして，それらの指標を利用して，データを分割する効率的なアルゴリズムを構築したいと考えている。

（成蹊大学経営学部教授）

謝辞　本研究は，成蹊大学教員研修制度の長期研修における成果の一部です。ここに感謝を申し上げます。

参考文献

Cuturi, M.（2013）: “Sinkhorn distances: Lightspeed computation of optimal transport”, *Advances in Neural Information Processing Systems*, 26, pp. 2292–2300.

田中研太郎.（2022）:“データ分割と最適輸送”, *成蹊大学経済経営論集*, 53(2), pp. 117–120,（http://hdl.handle.net/10928/1534）.