異なるダイバージェンスに基づく分布的ロバスト最適化モデルの比較
A Comparison of Distributionally Robust Optimization Models Defined by Different Divergence Measures

経営システム工学専攻
前田　彩

1. 序論

不確実性のあるデータから生じうる範囲を想定し, その中で最悪な状況を基に最適解を導く「ロバスト最適化（Distributionally Robust Optimization, DRO）」の研究が注目を集めており, 金融や設計, 文字認識など様々な分野で使われはじめている. しかし DRO は扱いにくい問題であるとされており，不確実な範囲（不確実性集合）をどう定義し, どうしたら解きやすく変形できるかが研究の目的になっている. 不確実性集合を定義する方法は「代替分布の台を経験分布の台に限った部分集合」と「代替分布の台を経験分布の台に限らない部分集合」の 2 つ存在し, 従来使われていた ϕ-ダイバージェンスを用いる DRO は前者であり，最近 [5] によって提案されたワッサースタイン距離を用いた DRO は後者に対応する. 一方で, ワッサースタイン距離を経験分布の台に限定した Earth Mover's Distance（EMD）を用いることで前者の DRO を定義することもできる. 本研究では, ワッサースタイン距離を用いた 2 つの DRO の比較を行う. 具体的にロジスティック回帰を取り上げ, 2 つの DRO に適用しそれらの事後パフォーマンスを比較した.

数値実験では標本の大きさに対する説明変数が多いデータでは前者に対応する一般的なワッサースタイン距離を用いた DRO の方が, 標本の大きさに対して説明変数が少ないデータでは 後者に対応する EMD を用いた DRO の方が良いパフォーマンスをしたことを示す. また, これ以降は 2 つのワッサースタイン距離を分けて考えるため, 後者の場合を「広義のワッサースタイン距離を用いた DRO」, 前者の場合を「狭義のワッサースタイン距離を用いた DRO」と呼ぶ.

2. 定式化

経験的最適化と DRO についてまとめ, 広義のワッサースタイン距離を用いた DRO と狭義のワッサースタイン距離を用いた DRO について示す.

2.1. 経験的最適化と分布的ロバスト最適化

本論文では以下のような最適化問題を考える.

$$\max_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{Y})] \tag{2.1}$$

ここで $f(\boldsymbol{x}, \boldsymbol{Y})$ は $\boldsymbol{x}$ と $\boldsymbol{Y}$ による報酬関数, $\boldsymbol{x} \in \mathbb{R}^b$ は決定変数, $\boldsymbol{Y}$ は確率変数を表す. $\mathbb{P}$ は $\boldsymbol{Y}$ が従う真の分布を表す. 現実における多くの場合, 分布 $\mathbb{P}$ は未知であるため, この問題 (2.1) は実行ができない. しかし, $\mathbb{P}$ に基づいて生成された $\boldsymbol{Y}$ の経験データ $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n, \boldsymbol{Y}_i \in \mathbb{R}^m$ を得ることができ, (2.1) の代わりに

$$\max_{\boldsymbol{x}} \left\{ \mathbb{E}_{\hat{\mathbb{P}}_n}[f(\boldsymbol{x}, \boldsymbol{Y})] = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}, \boldsymbol{Y}_i) \right\} \tag{2.2}$$

を扱うことが多い. このとき $\hat{\mathbb{P}}_n$ は $\boldsymbol{Y}$ の経験分布である. (2.2) を経験的最適化問題（Empirical Optimization, EO）と呼ぶ. EO は標本数が大きければ最適値が (2.1) に近づくと言われているが, 多くの場合, 事後データのパフォーマンスを劣化させる. これに対し DRO は経験分布から想定した範囲内における最悪な分布に対して最適な解を導きだす方法論である. 本研究では以下のロバスト最適化問題について考える :

$$\max_{\boldsymbol{x}} \min_{\mathbb{Q} \in \mathcal{M}(\hat{\mathbb{P}}_n)} \left\{ \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{Y})] \right\} \tag{2.3}$$

このとき $\mathbb{Q}$ は代替分布，$\mathcal{M}(\hat{\mathbb{P}}_n)$ は不確実性集合である. 不確実性集合 $\mathcal{M}(\hat{\mathbb{P}}_n)$ から目的関数を最大にする最悪な分布を決定し, その最悪な分布を基に最適化する問題である. 前述したとおり, 不確実性集合を定義する方法は一般的に 2 つ存在し, ワッサースタイン距離を用いることで 2 つの場合について DRO の定式化を記し, その比較をする.

2.2. 広義のワッサースタイン距離の DRO

一般的なワッサースタイン距離を定義し, 広義のワッサースタイン距離の DRO とロジスティック回帰に当てはめた場合の定式化をする.

定義 1 (ワッサースタイン距離) 確率分布 $\mathbb{P}$ に対する確率分布 $\mathbb{Q}$ のワッサースタイン距離 $W(\mathbb{Q}|\mathbb{P})$ を以下で定義する.

$$W(\mathbb{Q}|\mathbb{P}) := \left| \begin{array}{ll} \inf\limits_{\Pi\in M(\Xi^2)} & \int_{\Xi^2} d(Y,Y')\Pi(\mathrm{dY},\mathrm{dY}') \\ \text{s.t.} & \Pi(\mathrm{dY},\Xi) = \mathbb{Q}(\mathrm{dY}) \\ & \Pi(\Xi,\mathrm{dY}') = \mathbb{P}(\mathrm{dY}') \end{array} \right.$$

Ξ は分布の台が含む空間, $\Xi\times\Xi$ 上の確率分布の集合を $M(\Xi^2)$ とする. Y は分布 $\mathbb{Q}$ についての確率変数, Y' は分布 $\mathbb{P}$ についての確率変数であり, Π は分布 $\mathbb{P}$ と分布 $\mathbb{Q}$ の同時確率分布である. 任意のノルム $\|\cdot\|$ を用いて $d(Y,Y') = \|Y-Y'\|$ と与えるとする. これは分布 $\mathbb{Q}$ と分布 $\mathbb{P}$ の 2 つの分布間の距離を示していると解釈でき, 定義 1 を用いて不確実性集合を $\mathcal{M}(\hat{\mathbb{P}}_n) := \{\mathbb{Q} : W(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon\}$ と表すことができる. $\varepsilon \ge 0$ はワッサースタイン球の半径である. 半径 ε が分布的ロバスト最適化問題におけるパラメータとなり, 不確実性集合の大きさを調節することができる. 定義 1 と (2.3) から, DRO の定式化を以下のように表せる:

$$\sup_{\boldsymbol{x}} \inf_{\mathbb{Q}\in\mathcal{M}(\hat{\mathbb{P}}_n)} \left\{\mathbb{E}_{\hat{\mathbb{P}}_n}[f(\boldsymbol{x},\boldsymbol{Y})] | W(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon\right\}.$$

この DRO の解を $x(\varepsilon)$ とする. 双対性を利用することで簡単な凸計画問題に帰着することができる [5]. 具体的にロジスティック回帰を用いて, 広義のワッサースタイン距離を用いた場合の定式化とそれを変形させた等価な定式化を示す.

ロジスティック回帰への適用

ロジスティック回帰はデータ分析の手法として最もよく用いられる回帰分析の 1 つであり, 説明変数をもとに 2 値の被説明変数がどちらであるかを推測する手法である. 被説明変数 $y_i \in \{-1,1\}$, 説明変数 $\boldsymbol{z}_i \in \mathbb{R}^m (i = 1,\cdots,n)$ があるとき $\mathrm{Prob}(y_i|\boldsymbol{z}_i) = (1+\exp(-y_i(\boldsymbol{z}_i^T\boldsymbol{x}+x_0)))^{-1}$, と表せる. このとき x_0 と $\boldsymbol{x} \in \mathbb{R}^m$ は未知の回帰係数である. 回帰係数を求めるにはロジスティックの対数尤度関数を最大化すればよく, 対数尤度関数は

$$l_{\boldsymbol{x}}(\boldsymbol{z_i}, y_i) = -\log(1+\exp(-y_i(\boldsymbol{z}_i^T\boldsymbol{x}+x_0))$$

であり, これに対する定式化を示す:

$$\sup_{\boldsymbol{x}} \inf_{\mathbb{Q}\in\mathcal{M}(\hat{\mathbb{P}}_n)} \left\{\mathbb{E}_{\hat{\mathbb{P}}_n}[l_{\boldsymbol{x}}(\boldsymbol{z},\boldsymbol{y})] | W(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon\right\}.$$

これは, 双対性を利用して以下のような等価な扱いやすい凸計画問題に変形することができる [1].

$$\left| \begin{array}{ll} \max\limits_{\boldsymbol{x},\lambda,s_i} & \varepsilon\lambda + \dfrac{1}{n}\sum\limits_{i=1}^{n} s_i \\ \text{s.t.} & -l_{\boldsymbol{x}}(\hat{\boldsymbol{z}}_i,\hat{y}_i) \ge s_i \\ & -l_{\boldsymbol{x}}(\hat{\boldsymbol{z}}_i,-\hat{y}_i) - \lambda\kappa \ge s_i \\ & \|\boldsymbol{x}\| \le \lambda \end{array} \right. \tag{2.4}$$

このとき, $\lambda, s_i (i = 1,\cdots,n)$ は双対係数であり, 正の重み κ は 1 とした. $\|\cdot\|$ は任意のノルムである.

2.3. 狭義のワッサースタイン距離の DRO

ワッサースタイン距離の条件を限定した EMD を定義し, 狭義のワッサースタイン距離の DRO とロジスティック回帰に当てはめた定式化を示す.

定義 2 (EMD) 確率分布 $\mathbb{P}$ に対する確率分布 $\mathbb{Q}$ の EMD である $W'(\mathbb{Q}|\mathbb{P})$ を以下で定義する.

$$W'(\mathbb{Q}|\mathbb{P}) := \left| \begin{array}{ll} \min\limits_{\pi_{ij}} & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} d_{ij}\pi_{ij} \\ \text{s.t.} & \sum\limits_{i=1}^{n} \pi_{ij} = p_j (j = 1,\cdots,n) \\ & \sum\limits_{j=1}^{n} \pi_{ij} = q_i (i = 1,\cdots,n) \\ & \pi_{ij} \ge 0 \end{array} \right.$$

このとき, 確率分布 $\mathbb{P} = (p_1,\cdots,p_n)$, 確率分布 $\mathbb{Q} = (q_1,\cdots,q_n)$ であり, π_{ij} は p_j と q_i の同時確率分布である. 分布 $\mathbb{Q}$ は分布 $\mathbb{P}$ にとって絶対連続な分布であり, これにより, 不確実性集合において「代替分布の台が経験分布の台に限る部分集合」を定義することができる. EMD を用いた不確実性集合は $\mathcal{M}'(\hat{\mathbb{P}}_n) := \{\mathbb{Q} : W'(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon'\}$ であり, $\varepsilon' \ge 0$ はワッサースタイン球の半径である.

定義 2 用いて, 狭義のワッサースタインの DRO の定式化を次に示す [3].

$$\max_{\boldsymbol{x}} \min_{\mathbb{Q}\in\mathcal{M}(\hat{\mathbb{P}}_n)} \left\{\mathbb{E}_{\hat{\mathbb{P}}_n}[f(\boldsymbol{x},\boldsymbol{Y})] | W'(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon'\right\} \tag{2.5}$$

これも双対性を利用して簡単な凸計画問題に帰着することができ, 具体的にロジスティック回帰を用いて, 広義のワッサースタイン距離を用いた場合の定式化とそれを変形させた等価な定式化を示す.

ロジスティック回帰への適用

前節のロジスティック損失関数を (2.5) で表す:

$$\max_{\boldsymbol{x}} \min_{\mathbb{Q}\in\mathcal{M}(\hat{\mathbb{P}}_n)} \left\{\mathbb{E}_{\hat{\mathbb{P}}_n}[l_{\boldsymbol{x}}(\boldsymbol{z},\boldsymbol{y})] | W'(\mathbb{Q}|\hat{\mathbb{P}}_n) \le \varepsilon'\right\} \tag{2.6}$$

双対性を利用して (2.6) と等価な扱いやすい凸計画問題に帰着させる [3].

$$\left|\begin{array}{ll} \max\limits_{\boldsymbol{x},\mu_j,\theta} & \varepsilon'\theta + \sum\limits_{j=1}^{n} p_j\mu_j \\ \text{s.t.} & \theta\sum\limits_{j=1}^{n} d_{ij} + \mu_j \le l_{\boldsymbol{x}}(\boldsymbol{z}_i, y_i) \\ & \theta \ge 0 \end{array}\right. \tag{2.7}$$

このとき $\theta, \mu_j (j = 1, \cdots, n)$ は双対係数であり, $\|\cdot\|$ は任意のノルムである.

3. 平均分散フロンティア

ワッサースタイン距離を用いた場合に限らず多くの DRO は, 不確実なデータから起こりうる範囲は何かしらのパラメータに依存している. パラメータを動かすことで, 範囲の不確実なデータから起こりうる調節をすることができる. [3] では, パラメータを動したときの最適値のパフォーマンスの振る舞いについて書かれており, わずかなパラメータを与えることで, 最適値を保障しつつ, 最適値の変動（分散）を大幅に抑えられることが指摘されている. パラメータによるロバスト最適化問題の最適値と変動（分散）の関係を見ることができ, それを平均分散フロンティアと呼んでいる. 最適値と変動（分散）については次のように示す.

$$\begin{aligned} \mu(\varepsilon) &:= \mathbb{E}_{\hat{\mathbb{P}}_n}[f(x^*(\varepsilon), \boldsymbol{Y})] = \frac{1}{n}\sum_{i=1}^{n} f(x^*(\varepsilon), \boldsymbol{Y}_i) \\ \sigma^2(\varepsilon) &:= \mathbb{V}_{\hat{\mathbb{P}}_n}[f(x^*(\varepsilon), \boldsymbol{Y})] \\ &= \frac{1}{n}\sum_{i=1}^{n}\{f(x^*(\varepsilon), \boldsymbol{Y}_i) - \mu(\varepsilon)\}^2 \end{aligned}$$

このとき, $x^*(\varepsilon)$ はパラメータ ε のときの最適解である. パラメータに対する最適値と分散の関係を観察できるので, 最適値のパフォーマンスをどれくらい下げてしまうが, 分散をどれ程下げられたかという点で評価することができ, 本研究の 2 つの最適化問題の比較に有効であると考えた.

4. 数値実験

数値実験を行い, 平均分散フロンティアを用いて 2 つの DRO を比較する.

4.1. 使用するデータおよび検証方法

ロジスティック回帰で使用する実データ WCBD, ILPD[4][1] を使用した. WCBD の標本数は 569, 説明変数は 3 であり, ILPD の標本数は 583, 説明変数は 10 である.

[1]Breast Cancer Wisconsin は略称として WDBC, Indian Liver Patient Dataset の略称を ILPD と呼ぶ. また, WDBC は,30 個ある説明変数から 3 つに絞ったデータを利用する.

4.2. 検証結果

事前データの比較

図 4.1 は WDBC,ILPD の事前データの B-WRO と N-WRO[2] の平均分散フロンティアである. パラメータが 0 のときは, EO であり平均も分散も同じ結果が得られる. どちらもパラメータを与えることで, 大幅に分散が減っていることがわかる. WDBC はパラメータ $\varepsilon = 0.005, \varepsilon' = 2$ まで平均はほとんど変わらず分散を大きく下げているが, さらに大きいパラメータを与えると, B-WRO と N-WRO どちらも似たような振る舞いをしながら平均が小さくなっている. ILPD は, N-WRO の方が緩やかに平均が落ちている. 標本数に対する説明変数が多い ILPD にとって, 代替分布を経験分布の台に限定していない B-WRO は不確実性集合が大きくなり, より最悪な分布を想定したのではないかと考えた. 一方で, N-WRO は経験分布の台に限定しており, B-WRO より不確実性集合が小さく設定され, 比較的にパフォーマンスが良くなったのではないかと考えられる.

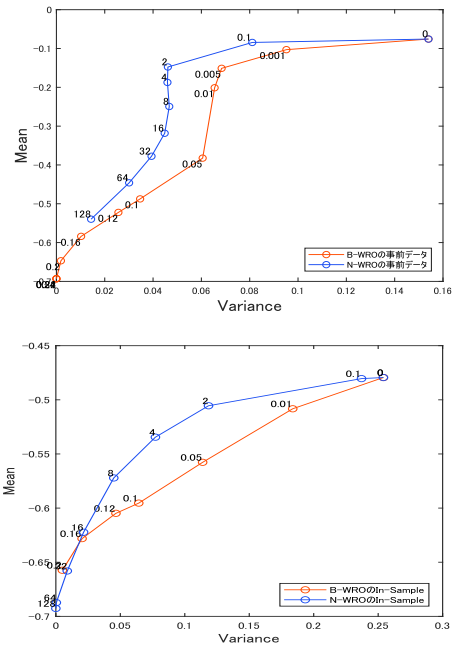

図 4.1: WDBC(上), ILPD(下) の事前データの平均分散フロンティア

[2]広義なワッサースタイン距離の分布的ロバスト最適化問題を B-WDR（Broad Sense Wasserstein Distributionally Robust Optimization）とし, 狭義なワッサースタイン距離の分布的ロバスト最適化問題を N-WRO（Narrow Sense Wasserstein Distributionally Robust Optimization）とする.

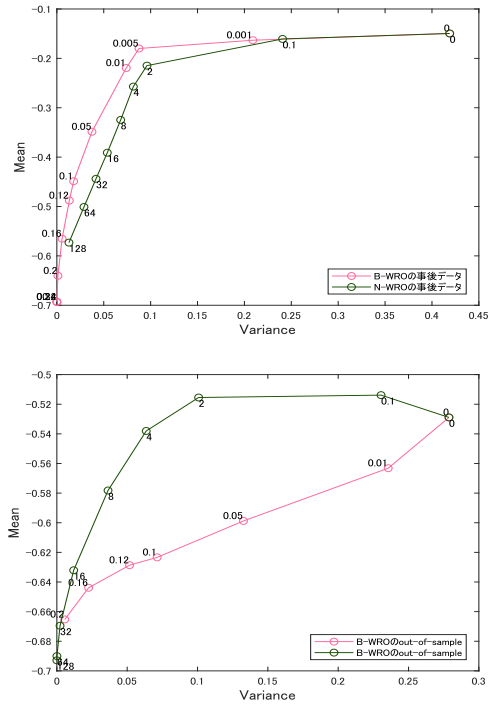

図 4.2: WDBC(上), ILPD(下) の事後データの平均分散フロンティア

事後データの比較

WDBC, ILPD に対する B-WRO と N-WRO の事後データから平均分散フロンティアを示し, 比較する. 図 4.2 から, 事前データと同様にどの平均分散フロンティアもパラメータを与えることで, 大幅に分散が小さくなった. 事前データではどちらも N-WRO の方が良いパフォーマンスをしていたが, 事後データでは異なる結果がでた. WDBC の場合は B-WRO のほうが, ILPD の場合は N-WRO のほうがパフォーマンスが良くなった. 説明変数に対する変数のデータ数ではないかと考えた. 2 つは標本数がほとんど変わらないが, WDBC の方が説明変数 3, ILPD は 11 である. 説明変数に対する変数のデータ数が少ない場合は, 最悪な分布を経験分布上のみから想定する-WRO の方が良いパフォーマンスをしするのではないかと考えた.

5. 結論

本論文では DRO において「代替分布の台が経験分布の台に限らない場合」と「代替分布の台が経験分布の台に限る場合」の 2 つ場合について比較を行った. 前者の場合である広義のワッサースタイン距離を用いた DRO と後者の場合である狭義のワッサースタイン距離を用いた DRO の定式化を示し, 具体的にロジスティック回帰を取り上げて, 2 つの DRO について数値実験を行った. 数値実験の結果として, ロジスティック回帰については標本の大きさに対して説明変数が少ない実データでは前者の場合である広義のワッサースタイン距離を用いた DRO の方が, 標本の大きさに対して説明変数が多い実データでは行為者の場合である狭義義のワッサースタイン距離を用いた DRO の方が良いパフォーマンスをした. また, 本論ではロジスティック回帰だけでなく効用関数を用いたポートフォリオ選択問題に関しても定式化し, 数値実験を行っている. 事前データの比較ではロジスティックと同じように狭義のワッサースタイン距離を用いた DRO のパフォーマンスが良くなる傾向があったが, 事後データの比較においてはロジスティック回帰のように 2 つの DRO について差別化できる特徴を見ることはできなかった. 詳細については, 本論を見ていただきたい. 今後の課題としては, 狭義のワッサースタイン距離を用いた分布的ロバスト最適化問題については計算時間がかかる傾向があったので, より早い計算ができる方法を考えるべきである. また, ロジスティック回帰については大きな違いを見つけることはできたが他のデータを利用して検証する必要がある.

参考文献

[1] D. Kuhn, S. S. Abadeh and P. M. Esfahani, “Distributionally Logistic Regression,” arXiv:1509.09259v3, 2015.

[2] J. Gotoh, M. J. Kim and A. E. B. Lim, “Calibration of Distributionally Robust Empirical Optimization Models,” arXiv:1711.06565v1, 2017.

[3] J. Gotoh, ‘’Robust Empirical Optimization with Wasserstein Uncertainty, *unpublished*,2017.

[4] M. Lichman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. URL: `http://archive.ics.uci.edu/ml`.

[5] P. M. Esfahani, D. Kuhn, ”Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations,” *Mathematical Programming ISSN 1436-4646*, 2017.