BOSTON UNIVERSITY

Ditributionally Robust Learning and Applications to Predictive and Prescriptive Health Analytics

Ruidi Chen[1] and Ioannis Ch. Paschalidis[1,2]
1 Division of Systems Eng., Boston Univ.
2 Dept. of Electrical & Computer Eng., Boston Univ.

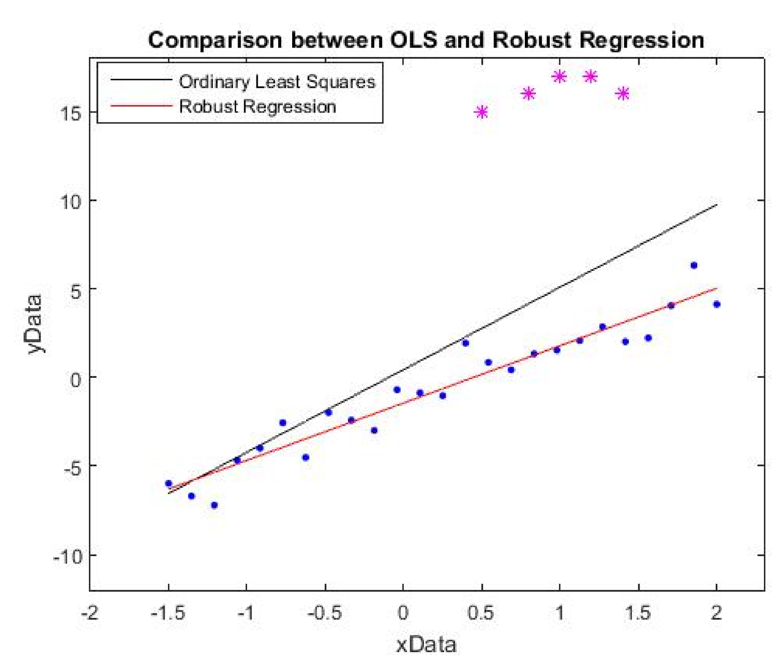
The Center for Information and Systems Engineering CISE

Abstract

We present a Wasserstein metric-based *Distributionally Robust Optimization (DRO)* approach to develop novel learning algorithms that are robust to adversarial perturbations in the input data. Our approach minimizes the worst case loss over a family of distributions on the observed data that are close to the empirical distribution in the sense of the Wasserstein metric. The min-max Wasserstein DRO problem can be relaxed to a convex regularized learning formulation that links robustness to regularization. We establish bounds on the prediction and estimation biases of the solution to our formulation under mild conditions. The proposed approach has been applied to develop robust regression, classification, and optimal decision making techniques that have been shown to outperform the classical methodologies both theoretically and empirically. Two notable applications include detecting CT exams with an abnormally high radiation exposure, and prescribing optimal treatments for patients with diabetes or hypertension. The latter implies an applicability of this framework to data-driven decision making.

Wasserstein DRO Approach

- Intuition:
  - Estimate the regression line that is not skewed by outliers, through minimizing some expected loss function under the "true" distribution of $(\mathbf{x}, y)$, where $\mathbf{x}$ is the feature vector and y is the response variable.



  - The samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, may be contaminated with outliers.
  - Solution: hedge against a family of plausible distributions.
- The Wasserstein DRO problem:

$$\inf_{\beta \in \mathscr{D}} \sup_{\mathbb{Q} \in \mathscr{B}} E^{\mathbb{Q}}[|y - \mathbf{x}'\beta|].$$

- Notation:
  - β: the regression coefficient to be estimated; $\mathbb{Q}$: the probability distribution of $(\mathbf{x}, y)$.
  - $\mathscr{B}$: the Wasserstein ball of distributions centered at the empirical distribution $\hat{\mathbb{P}}_N$: $\mathscr{B} = \{\mathbb{Q} \in \mathscr{M}(\mathscr{Z}) : d_W(\mathbb{Q}, \hat{\mathbb{P}}_N) \le \varepsilon\}$, where the Wasserstein distance is defined through,

$$d_W(\mathbb{Q}, \hat{\mathbb{P}}_N) \triangleq \min_{\Pi \in \mathscr{P}(\mathscr{Z} \times \mathscr{Z})} \left\{ \int_{\mathscr{Z} \times \mathscr{Z}} \|(\mathbf{x}_1, y_1) - (\mathbf{x}_2, y_2)\| \, \Pi(d(\mathbf{x}_1, y_1), d(\mathbf{x}_2, y_2)) \right\},$$

with Π the joint distribution of $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$, with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$.

- The DRO problem could be relaxed to

$$\inf_{\beta \in \mathscr{D}} \varepsilon \|(-\beta, 1)\|_* + \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\beta|.$$

Performance Guarantees

Let β^* and $\hat{\beta}$ be the true and estimated regression coefficients, respectively.

Theorem 1 *When $\|(\mathbf{x}, y)\| \le R$ a.s., $\|(-\beta, 1)\|_* \le \bar{B}$, for any $0 < \delta < 1$, with probability at least $1 - \delta$ with respect to the sampling,*

$$\mathbb{E}[|y - \mathbf{x}'\hat{\beta}|] \le \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\hat{\beta}| + \frac{2\bar{B}R}{\sqrt{N}} + \bar{B}R\sqrt{\frac{8\log(2/\delta)}{N}}.$$

Theorem 2 *Under sub-Gaussian assumption on $(\mathbf{x}, y)$, when $\|(-\beta, 1)\|_2 \le \bar{B}_2$, with a high probability,*

$$\|\hat{\beta} - \beta^*\|_2 \le \frac{\bar{C}R\bar{B}_2\mu}{N\lambda_{min}} w(\mathscr{B}_u)\Psi(\beta^*).$$

Groupwise Wasserstein Grouped LASSO

- Goal: recover the group-wise sparsity in the predictors, e.g.,
  - Dummies used to represent different levels of a categorical predictor;
  - Gene expression data: gene pathways define the groups.
- Notation:
  - $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, and $\beta = (\beta^1, \ldots, \beta^L)$, where $\mathbf{x}^l, \beta^l$ are the predictor and regression coefficient of group l (which has p_l predictors), respectively.
  - $\mathbf{z_w} \triangleq (\frac{1}{\sqrt{p_1}}\mathbf{x}^1, \ldots, \frac{1}{\sqrt{p_L}}\mathbf{x}^L, My)$, where M is very large.
- Define the Wasserstein metric using the following norm:

$$\|\mathbf{z_w}\|_{2,\infty} \triangleq \max\left\{ \frac{1}{\sqrt{p_1}}\|\mathbf{x}^1\|_2, \ldots, \frac{1}{\sqrt{p_L}}\|\mathbf{x}^L\|_2, M|y| \right\},$$

the DRO problem could be relaxed to the following GWGL formulation:

$$\inf_{\beta} \left(\varepsilon \sum_{l=1}^{L} \sqrt{p_l}\|\beta^l\|_2 + \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathbf{x}_i'\beta| \right).$$

Grouping Effect

Theorem 3 *Suppose the predictors are standardized and the response is centered. If $\mathbf{x}_{,i}$ is in group l_1 and $\mathbf{x}_{,j}$ is in group l_2, and $\|\hat{\beta}^{l_1}\|_2 \neq 0$, $\|\hat{\beta}^{l_2}\|_2 \neq 0$, where $\hat{\beta} = (\hat{\beta}^1, \ldots, \hat{\beta}^L)$ is the solution to GWGL, define*

$$D(i, j) = \left| \frac{\sqrt{p_{l_1}}\hat{\beta}_i}{\|\hat{\beta}^{l_1}\|_2} - \frac{\sqrt{p_{l_2}}\hat{\beta}_j}{\|\hat{\beta}^{l_2}\|_2} \right|.$$

Then,

$$D(i, j) \le \frac{\sqrt{2(1-\rho)}}{\sqrt{N}\varepsilon},$$

where $\rho = \mathbf{x}_{,i}'\mathbf{x}_{,j}$ is the sample correlation, and p_{l_1}, p_{l_2} are the number of features in groups l_1 and l_2, respectively.

Prediction-based Optimal Decision Making

- **Problem**: given a set of actions $[M] \triangleq \{1, \ldots, M\}$, choose the one that yields the best future outcome y, with the aid of auxiliary data $\mathbf{x}$ that is predictive of y.
- **Idea**: *predict* the outcome under each action using a robust nonlinear framework, and *prescribe* the actions based on their predictions.
- **Applications**: Prescribe optimal treatments for patients with diabetes or hypertension.

Robust Nonlinear Prescription

- Assumption: under each action m, $y_m = \mathbf{x}_m'\beta_m^* + h_m(\mathbf{x}_m) + \varepsilon_m$.
- Method:
  - For each $m \in [M]$, derive a robust estimate of β_m^*, denoted by $\hat{\beta}_m$, by solving the Wasserstein DRO problem.
  - Given a new sample $\mathbf{x}$, find its K_m nearest neighbors in each action group m using the metric: $\|\mathbf{x} - \mathbf{x}_{mi}\|_{\hat{\mathbf{W}}_m} = \sqrt{(\mathbf{x} - \mathbf{x}_{mi})'\hat{\mathbf{W}}_m(\mathbf{x} - \mathbf{x}_{mi})}$, where $\hat{\mathbf{W}}_m = \text{diag}((\hat{\beta}_{m1})^2, \ldots, (\hat{\beta}_{mp})^2)$, and $\mathbf{x}_{mi}$ is the sample of group m.
  - Compute a K-NN estimate of the future outcome for $\mathbf{x}$ under action m: $\hat{y}_m(\mathbf{x}) = 1/K_m \sum_{i=1}^{K_m} y_{m(i)}$, where $y_{m(i)}$ is the response of the i-th closest neighbor to $\mathbf{x}$ in group m.
  - Prescribe action m with probability $e^{-\xi \hat{y}_m(\mathbf{x})} / \sum_{j=1}^{M} e^{-\xi \hat{y}_j(\mathbf{x})}$, where $\xi > 0$ is a pre-specified constant.

Prescriptive Performance

Theorem 4 *Assume $\hat{y}_m(\mathbf{x})$ and $y_m(\mathbf{x})$ are non-negative, $\forall m \in [M]$. For any $k \in [M]$,*

$$\sum_{m=1}^{M} \frac{e^{-\xi \hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi \hat{y}_j(\mathbf{x})}} y_m(\mathbf{x}) \le y_k(\mathbf{x}) + \left(\hat{y}_k(\mathbf{x}) - \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m(\mathbf{x}) \right) + \xi \left(\frac{1}{M} \sum_{m=1}^{M} \hat{y}_m^2(\mathbf{x}) + \sum_{m=1}^{M} \frac{e^{-\xi \hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi \hat{y}_j(\mathbf{x})}} y_m^2(\mathbf{x}) \right) + \frac{\log M}{\xi}.$$

CT Radiation Overdose Detection

- Identify patients who receive an abnormally high radiation dose.
- Response variable: CTDI (CT Dose Index)
- Predictors: patient characteristics and exam related variables.
  - Patient Gender, Height, Manufacturer, Scanner protocol, Xray-modulation-type.
  - Totally 28 predictors, and 189,959 patients.
- Outlier detection criterion:

$$\text{Outlier} = \begin{cases} \text{YES}, & \text{if } |\text{residual}| > \text{threshold} \times \hat{\sigma}, \\ \text{NO}, & \text{otherwise}. \end{cases}$$

- Specificity=0.85; Sensitivity=0.91; PPV=0.84; NPV=0.92.

		Manual Review		
		True Outlier	False Outlier	Total
Algorithm	True Outlier	84	16	100
	False Outlier	8	92	100
	Total	92	108	

GWGL on a Surgery Dataset

- Predict the post-operative hospital length of stay using patient demographics, pre- and intra-operative variables.
- 2,275,452 patients; 131 numerical predictors.
- 67 groups of predictors, with the following variables grouped together:
  - Dummy variables corresponding to the same categorical predictor.
  - Variables indicating the evidence, and the number of occurrences of the same disease.
  - Variables indicating similar diseases, e.g., cardiac arrest & myocardial infarction.
- The mean and standard deviation of out-of-sample *Median Absolute Deviation (MAD)*:

	Mean	Standard deviation
GLASSO with ℓ_2-loss	0.1716 (6.93%)	0.0013
GWGL	0.1597 (N/A)	6.34×10^{-4}
EN	0.1732 (7.79%)	4.58×10^{-4}
LASSO	0.1732 (7.79%)	4.43×10^{-4}
GSRL	0.1696 (5.84%)	6.58×10^{-4}

Prescribe Optimal Treatments

- Goal: develop optimal prescriptions for patients with type-2 diabetes and hypertension using the EHRs.
- Predictors: demographics, diagnoses, lab tests, and past admission records.
- Response: HbA_{1c}, and systolic blood pressure.
- The reduction in HbA_{1c}/systolic blood pressure, mean (std.):

	Diabetes		Hypertension	
	Deterministic	Randomized	Deterministic	Randomized
LASSO	-0.51 (0.16)	-0.51 (0.16)	-4.71 (1.09)	-4.72 (1.10)
CART	-0.45 (0.13)	-0.42 (0.14)	-4.84 (0.62)	-4.87 (0.66)
OLS+K-NN	-0.53 (0.13)	-0.53 (0.13)	-4.33 (0.46)	-4.33 (0.47)
DRO+K-NN	-0.56 (0.06)	-0.55 (0.08)	-6.98 (0.86)	-7.22 (0.82)
Current therapy	-0.22 (0.04)		-2.52 (0.19)	
Standard of care	-0.22 (0.03)		-2.37 (0.11)	

References

- Chen, R., and Paschalidis, I.C. (2018). A robust learning algorithm for regression models using distributionally robust optimization under the Wasserstein metric, **Journal of Machine Learning Research**, 19, 1-48.
- Chen, R., and Paschalidis, I.C. (2018). Learning optimal personalized treatment rules using robust regression informed K-NN, **NIPS Machine Learning for Health (ML4H) workshop**, Montreal, Canada.
- Chen, R., Paschalidis, I.C., Hatabu, H., Valtchinov, V.I., and Siegelman J. (2019). Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-Data using regularized regression, **European Journal of Radiology Open**.