Robust Data-driven Prescriptiveness Optimization

Mehran Poursoltani
McGill University

Erick Delage
HEC Montreal

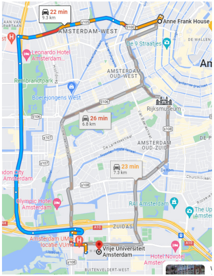
Angelos Georghiou
University of Cyprus

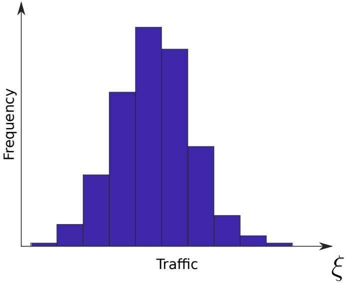
The Forty-first International Conference on Machine Learning (ICML 2024)

STOCHASTIC VS. CONTEXTUAL OPTIMIZATION

Stochastic Programming

$$(\text{SP}) \quad x^* \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} \; \mathbb{E}_F \left[h(x, \xi) \right]$$





- ξ traffic demand with distribution F
- x shortest path route

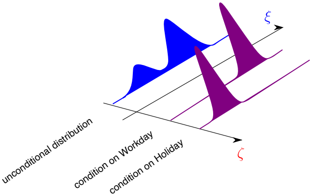
STOCHASTIC VS. CONTEXTUAL OPTIMIZATION

Contextual Stochastic Optimization

$$(\text{CSO})\ \boldsymbol{x}^*(\zeta) \in \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}}\ \mathbb{E}_F\left[h(\boldsymbol{x},\boldsymbol{\xi})\,\middle|\,\zeta\right]$$

Workday

Holiday



- $\boldsymbol{\xi}$ traffic demand
- $\zeta \in \{\text{workday}, \text{holiday}\}$ side information
- F is the joint distribution $(\zeta, \boldsymbol{\xi})$, $F_{\xi|\zeta}$ conditional distribution

STOCHASTIC VS. CONTEXTUAL OPTIMIZATION

Contextual Stochastic Optimization

$$(\text{CSO})\ \boldsymbol{x}^*(\zeta) \in \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}}\ \mathbb{E}_F\left[h(\boldsymbol{x},\boldsymbol{\xi})\middle|\zeta\right]$$

- $F(\boldsymbol{\xi}|\zeta)$ not known in practice
- Estimate conditional distribution $\hat{F}(\boldsymbol{\xi}|\zeta)$, e.g., KDE, random forest, etc.
- Can we trust the estimates?

STOCHASTIC VS. CONTEXTUAL OPTIMIZATION

Contextual Stochastic Optimization

$$(\text{CSO}) \; \boldsymbol{x}^*(\zeta) \in \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \; \mathbb{E}_F\left[h(\boldsymbol{x},\boldsymbol{\xi})\middle|\zeta\right]$$

- $F(\xi|\zeta)$ not known in practice
- Estimate conditional distribution $\hat{F}(\xi|\zeta)$, e.g., KDE, random forest, etc.
- Can we trust the estimates?

Distributionally Robust Contextual Stochastic Optimization

$$(\text{DRCSO}) \; \boldsymbol{x}^*(\zeta) \in \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \sup_{F\in\mathcal{D}} \; \mathbb{E}_F\left[h(\boldsymbol{x},\boldsymbol{\xi})\middle|\zeta\right]$$

where $\mathcal{D}$ is admissible set of distributions (ambiguity set)

Relevant Literature

Ban and Rudin [2019] The big data newsvendor: Practical insights from machine learning

- Conditional Stochastic Optimization (CSO)
- Nadaraya-Watson Kernel regression
- Decision rules

Hannah et al. [2010] Nonparametric density estimation for stochastic optimization with an observable state variable

- Conditional Stochastic Optimization (CSO)
- Nadaraya-Watson Kernel regression
- Dirichlet process mixture models

Bertsimas and Van Parys [2022] Bootstrap robust prescriptive analytics

- Distributionally Robust Conditional Stochastic Optimization (DRCSO)
- Nadaraya-Watson Kernel regression
- Nearest neighbors learning

Wang et al. [2021] Distributionally robust prescriptive analytics with Wasserstein distance

- Distributionally Robust Conditional Stochastic Optimization (DRCSO)
- Nadaraya-Watson Kernel regression

Relevant Literature

Ban and Rudin [2019] The big data newsvendor: Practical insights from machine learning

- Conditional Stochastic Optimization (CSO)
- Nadaraya-Watson Kernel regression
- Decision rules

Hannah et al. [2010] Nonparametric density estimation for stochastic optimization with an observable state variable

- Conditional Stochastic Optimization (CSO)
- Nadaraya-Watson Kernel regression
- Dirichlet process mixture models

Bertsimas and Van Parys [2022] Bootstrap robust prescriptive analytics

- Distributionally Robust Conditional Stochastic Optimization (DRCSO)
- Nadaraya-Watson Kernel regression
- Nearest neighbors learning

Wang et al. [2021] Distributionally robust prescriptive analytics with Wasserstein distance

- Distributionally Robust Conditional Stochastic Optimization (DRCSO)
- Nadaraya-Watson Kernel regression

$\Longrightarrow$ How to compare different methods?

Coefficient of prescriptiveness[1]

"Recently proposed performance measure"

Given a data-driven policy $\boldsymbol{x}(\cdot)$ and distribution F

$$\mathcal{P}_F(\boldsymbol{x}(\cdot)) := 1 - \frac{\mathbb{E}_F[h(\boldsymbol{x}(\boldsymbol{\zeta}), \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]}{\mathbb{E}_F[h(\hat{\boldsymbol{x}}, \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]},$$

where $\hat{\boldsymbol{x}} \in \operatorname{argmin}_{\boldsymbol{x}} \mathbb{E}_{\hat{F}}[h(\boldsymbol{x}, \boldsymbol{\xi})]$ with $\hat{F}$ as the in-sample empirical distribution that puts equal weights on each observed data point (i.e. the solution of SAA)

[1] Bertsimas and Kallus, MS, 2020

COEFFICIENT OF PRESCRIPTIVENESS[1]

Given a data-driven policy $\boldsymbol{x}(\cdot)$ and distribution F

$$\mathcal{P}_F(\boldsymbol{x}(\cdot)) := 1 - \frac{\overbrace{\mathbb{E}_F[h(\boldsymbol{x}(\zeta), \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]}^{\textbf{distance from full information}}}{\underbrace{\mathbb{E}_F[h(\hat{\boldsymbol{x}}, \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]}_{\textbf{distance from no to full information}}},$$

where $\hat{\boldsymbol{x}} \in \operatorname{argmin}_{\boldsymbol{x}} \mathbb{E}_{\hat{F}}[h(\boldsymbol{x}, \boldsymbol{\xi})]$ with $\hat{F}$ as the in-sample empirical distribution that puts equal weights on each observed data point (i.e. the solution of SAA)

[1] Bertsimas and Kallus, MS, 2020

Coefficient of prescriptiveness

$$\mathcal{P}_F(x(\cdot)) := 1 - \frac{\overbrace{\mathbb{E}_F[h(x(\zeta),\xi)] - \mathbb{E}_F[\min_{x'\in\mathcal{X}} h(x',\xi)]}^{\text{distance from full information}}}{\underbrace{\mathbb{E}_F[h(\hat{x},\xi)] - \mathbb{E}_F[\min_{x'\in\mathcal{X}} h(x',\xi)]}_{\text{distance from no to full information}}}$$

Properties

- $\mathcal{P}_F = 1$:
  $x(\cdot)$ is fully anticipative
  in terms of ξ.

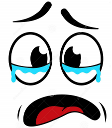
Coefficient of prescriptiveness

$$\mathcal{P}_F(\boldsymbol{x}(\cdot)) := 1 - \frac{\overbrace{\mathbb{E}_F[h(\boldsymbol{x}(\boldsymbol{\zeta}), \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]}^{\textbf{distance from full information}}}{\underbrace{\mathbb{E}_F[h(\hat{\boldsymbol{x}}, \boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}' \in \mathcal{X}} h(\boldsymbol{x}', \boldsymbol{\xi})]}_{\textbf{distance from no to full information}}}$$

Properties

- $\mathcal{P}_F = 1$:
  $\boldsymbol{x}(\cdot)$ is fully anticipative in terms of $\boldsymbol{\xi}$.

- Small $\mathcal{P}_F \approx 0$:
  $\boldsymbol{x}(\cdot)$ is not able to exploit information.



Rising Popularity of the Coefficient of Prescriptiveness

Recent papers exploiting $\mathcal{P}_F$ for evaluating the superiority of the contextual optimization methods:

- Bertsimas et al. [2016]
  Inventory management in the era of big data
- Bertsimas and Kallus [2020]
  From predictive to prescriptive analytics
- Notz and Pibernik [2022]
  Prescriptive analytics for flexible capacity management
- Kallus and Mao [2022]
  Stochastic optimization forests

Rising Popularity of the Coefficient of Prescriptiveness

Recent papers exploiting $\mathcal{P}_F$ for evaluating the superiority of the contextual optimization methods:

- Bertsimas et al. [2016]
  Inventory management in the era of big data
- Bertsimas and Kallus [2020]
  From predictive to prescriptive analytics
- Notz and Pibernik [2022]
  Prescriptive analytics for flexible capacity management
- Kallus and Mao [2022]
  Stochastic optimization forests

Can we optimize directly the coefficient of prescriptiveness in a way that is robust to distribution misspecification?

Distributionally Robust Prescriptiveness competitive ratio (DRPCR)

$$\max_{\boldsymbol{x}(\cdot)\in\mathcal{H}} \inf_{F\in\mathcal{D}} \mathcal{P}_F(\boldsymbol{x}(\cdot)) :=$$

$$\max_{\boldsymbol{x}(\cdot)\in\mathcal{H}} \inf_{F\in\mathcal{D}} 1 - \frac{\mathbb{E}_F[h(\boldsymbol{x}(\boldsymbol{\zeta}),\boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}'\in\mathcal{X}} h(\boldsymbol{x}',\boldsymbol{\xi})]}{\mathbb{E}_F[h(\hat{\boldsymbol{x}},\boldsymbol{\xi})] - \mathbb{E}_F[\min_{\boldsymbol{x}'\in\mathcal{X}} h(\boldsymbol{x}',\boldsymbol{\xi})]}$$

- Under weak conditions the optimal value of DRPCR is necessarily in the interval $[0,\ 1]$.

EPIGRAPH FORMULATION FOR DRPCR

DRPCR is equivalent to

$$\max_{\gamma} \quad \gamma \tag{1a}$$

$$\text{subject to} \quad \min_{\boldsymbol{x}(\cdot)\in\mathcal{H}} Q(\boldsymbol{x}(\cdot),\gamma) \leq 0 \tag{1b}$$

$$0 \leq \gamma \leq 1, \tag{1c}$$

where

$$Q(\boldsymbol{x}(\cdot),\gamma) := \sup_{F\in\mathcal{D}} \mathbb{E}_F\left[h(\boldsymbol{x}(\boldsymbol{\zeta}),\boldsymbol{\xi}) - \left((1-\gamma)h(\hat{\boldsymbol{x}},\boldsymbol{\xi}) + \gamma \min_{x'\in\mathcal{X}} h(\boldsymbol{x}',\boldsymbol{\xi})\right)\right]$$

is a convex increasing function of γ.

Epigraph formulation for DRPCR

DRPCR is equivalent to

$$\max_{\gamma} \quad \gamma \tag{1a}$$

$$\text{subject to} \quad \min_{x(\cdot)\in\mathcal{H}} Q(x(\cdot),\gamma) \leq 0 \tag{1b}$$

$$0 \leq \gamma \leq 1, \tag{1c}$$

where

$$Q(x(\cdot),\gamma) := \sup_{F\in\mathcal{D}} \mathbb{E}_F\left[h(x(\zeta),\xi) - \left((1-\gamma)h(\hat{x},\xi) + \gamma \min_{x'\in\mathcal{X}} h(x',\xi)\right)\right]$$

is a convex increasing function of γ.

Idea to solve the problem: use the bisection method to bisect over γ and solve the LHS of (1b) *to see whether it satisfies the constraint!*

Choice of the ambiguity set

Assumption

There is a discrete distribution $\bar{F}$, with $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta}$ and $\{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$ as the set of distinct scenarios for $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ respectively, such that the distribution set $\mathcal{D}$ takes the form of the "nested CVaR ambiguity set" with respect to $\mathbb{P}_{\bar{F}}$ and defined as

$$\bar{\mathcal{D}}(\bar{F},\alpha) := \left\{ \begin{array}{c} F \in \\ \mathcal{M}(\Omega_\zeta \times \Omega_\xi) \end{array} \middle| \begin{array}{r} \mathbb{P}_F(\boldsymbol{\zeta} = \boldsymbol{\zeta}_\omega) = \mathbb{P}_{\bar{F}}(\boldsymbol{\zeta} = \boldsymbol{\zeta}_\omega)\ \forall \omega \in \Omega_\zeta, \\ \mathbb{P}_F(\boldsymbol{\xi} = \boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \leq (1/(1-\alpha))\mathbb{P}_{\bar{F}}(\boldsymbol{\xi} = \boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \\ \forall\, \omega \in \Omega_\zeta, \omega' \in \Omega_\xi \end{array} \right\}$$

where $\mathcal{M}(\Omega_\zeta \times \Omega_\xi)$ is the set of all distributions supported on over the joint space $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta} \times \{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$.

Choice of the ambiguity set

Assumption

There is a discrete distribution $\bar{F}$*, with* $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta}$ *and* $\{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$ *as the set of distinct scenarios for* $\boldsymbol{\zeta}$ *and* $\boldsymbol{\xi}$ *respectively, such that the distribution set* $\mathcal{D}$ *takes the form of the "nested CVaR ambiguity set" with respect to* $\mathbb{P}_{\bar{F}}$ *and defined as*

$$\bar{\mathcal{D}}(\bar{F},\alpha) := \left\{ \begin{array}{c} F \in \\ \mathcal{M}(\Omega_\zeta \times \Omega_\xi) \end{array} \middle| \begin{array}{r} \mathbb{P}_F(\boldsymbol{\zeta} = \boldsymbol{\zeta}_\omega) = \mathbb{P}_{\bar{F}}(\boldsymbol{\zeta} = \boldsymbol{\zeta}_\omega)\ \forall \omega \in \Omega_\zeta, \\ \mathbb{P}_F(\boldsymbol{\xi} = \boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \leq (1/(1-\alpha))\mathbb{P}_{\bar{F}}(\boldsymbol{\xi} = \boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \\ \forall\, \omega \in \Omega_\zeta, \omega' \in \Omega_\xi \end{array} \right\}$$

where $\mathcal{M}(\Omega_\zeta \times \Omega_\xi)$ *is the set of all distributions supported on over the joint space* $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta} \times \{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$.

- No ambiguity in the marginal distribution of the observed random variable ζ

Choice of the ambiguity set

Assumption

There is a discrete distribution $\bar{F}$, with $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta}$ and $\{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$ as the set of distinct scenarios for $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ respectively, such that the distribution set $\mathcal{D}$ takes the form of the "nested CVaR ambiguity set" with respect to $\mathbb{P}_{\bar{F}}$ and defined as

$$\bar{\mathcal{D}}(\bar{F},\alpha) := \left\{ \begin{array}{c} F \in \\ \mathcal{M}(\Omega_\zeta \times \Omega_\xi) \end{array} \left| \begin{array}{r} \mathbb{P}_F(\boldsymbol{\zeta}=\boldsymbol{\zeta}_\omega) = \mathbb{P}_{\bar{F}}(\boldsymbol{\zeta}=\boldsymbol{\zeta}_\omega)\ \forall\omega\in\Omega_\zeta, \\ \mathbb{P}_F(\boldsymbol{\xi}=\boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \leq (1/(1-\alpha))\mathbb{P}_{\bar{F}}(\boldsymbol{\xi}=\boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}_\omega) \\ \forall\,\omega\in\Omega_\zeta,\omega'\in\Omega_\xi \end{array} \right. \right\}$$

where $\mathcal{M}(\Omega_\zeta \times \Omega_\xi)$ is the set of all distributions supported on over the joint space $\{\boldsymbol{\zeta}_\omega\}_{\omega\in\Omega_\zeta} \times \{\boldsymbol{\xi}_\omega\}_{\omega\in\Omega_\xi}$.

- No ambiguity in the marginal distribution of the observed random variable $\boldsymbol{\zeta}$
- Ambiguity solely on the unobserved random variable $\boldsymbol{\xi}$ and is sized using the parameter α

DRPCR under nested CVaR

Corollary

Under the nested CVaR ambiguity set we have

$$\min_{x(\cdot)\in\mathcal{H}} Q(\boldsymbol{x}(\cdot),\gamma) = \sum_{\omega\in\Omega_\zeta} \mathbb{P}_{\bar{F}}(\boldsymbol{\zeta}=\boldsymbol{\zeta}_\omega)\phi_\omega(\gamma)$$

where the optimal value of $\phi_\omega(\gamma)$ can be obtained through solving the following optimization problem

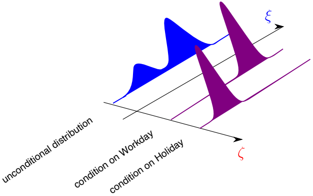
$$\begin{aligned} \min_{x\in\mathcal{X},t,s\geq 0} \quad & t+\frac{1}{1-\alpha}\sum_{\omega'\in\Omega_\xi}\mathbb{P}_{\bar{F}}(\boldsymbol{\xi}=\boldsymbol{\xi}_{\omega'}|\boldsymbol{\zeta}=\boldsymbol{\zeta}_\omega)s_{\omega'} \\ \text{subject to} \quad & s_{\omega'} \geq h(\boldsymbol{x},\boldsymbol{\xi}_{\omega'}) - \left((1-\gamma)h(\bar{\boldsymbol{x}},\boldsymbol{\xi}_{\omega'}) + \gamma\min_{x'\in\mathcal{X}} h(\boldsymbol{x}',\boldsymbol{\xi}_{\omega'})\right) \\ & \qquad -t\,,\ \forall\omega'\in\Omega_\xi. \end{aligned}$$

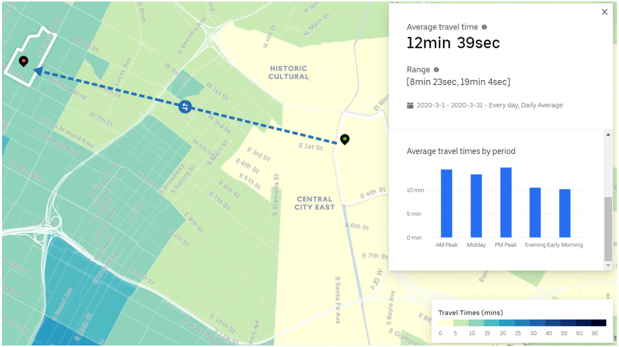
This problem can be reduced to a linear program when $\mathcal{X}$ is polyhedral and $h(\boldsymbol{x},\boldsymbol{\xi}_{\omega'})$ is linear programming representable for all $\omega'\in\Omega_\xi$.

Shortest Path Problem

Workday

Holiday





Shortest Path Problem with CSO objective

$$x^*(\zeta) \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}_{\hat{F}_{\xi|\zeta}}[x^\top \xi],$$

$$\mathcal{X} = \left\{ x \in \mathbb{R}^{|\mathcal{A}|} \;\middle|\; \begin{array}{ll} x_{(i,j)} \in \{0,1\} & \forall (i,j) \in \mathcal{A} \\ \sum_{j:(i,j)\in\mathcal{A}} x_{(i,j)} - \sum_{j:(j,i)\in\mathcal{A}} x_{(j,i)} = 1 & \text{if } i = o \\ \sum_{j:(i,j)\in\mathcal{A}} x_{(i,j)} - \sum_{j:(j,i)\in\mathcal{A}} x_{(j,i)} = -1 & \text{if } i = d \\ \sum_{j:(i,j)\in\mathcal{A}} x_{(i,j)} - \sum_{j:(j,i)\in\mathcal{A}} x_{(j,i)} = 0 & \forall i \in \mathcal{V} \setminus \{o, d\} \end{array} \right\},$$

- A directed graph defined as $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{A} \in \mathcal{V} \times \mathcal{V}$ is the set of arcs.
- $\xi_{(i,j)}$ denotes the travel time of a directed path from node i to node j.
- $x_{(i,j)} = 1$ if we decide to travel from node i to node j and $x_{(i,j)} = 0$ otherwise.
- $\hat{F}_{\xi|\zeta}$ denotes the conditional distribution inferred from the training dataset.
- Adapt to the graph ($\mathcal{G}$) structure employed in Kallus and Mao (2022)

Alternative methods to DRPCR

- *Contextual Stochastic Optimization (CSO)*

$$x^*(\zeta) \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}_{\hat{F}_{\xi|\zeta}}[x^\top \xi]$$

Alternative methods to DRPCR

- *Contextual Stochastic Optimization (CSO)*

$$x^*(\zeta) \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} \, \mathbb{E}_{\hat{F}_{\xi|\zeta}}[x^\top \xi]$$

- *Distributionally Robust Contextual Stochastic Optimization (DRCSO)*

$$x^*(\zeta) \in \arg\min_{x \in \mathcal{X}} \sup_{F_{\xi|\zeta} \in \bar{\mathcal{D}}(\hat{F}_{\xi|\zeta}, \alpha)} \mathbb{E}_{F_{\xi|\zeta}}[x^\top \xi]$$

Alternative methods to DRPCR

- *Contextual Stochastic Optimization (CSO)*

$$x^*(\zeta) \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}_{\hat{F}_{\xi|\zeta}}[x^\top \xi]$$

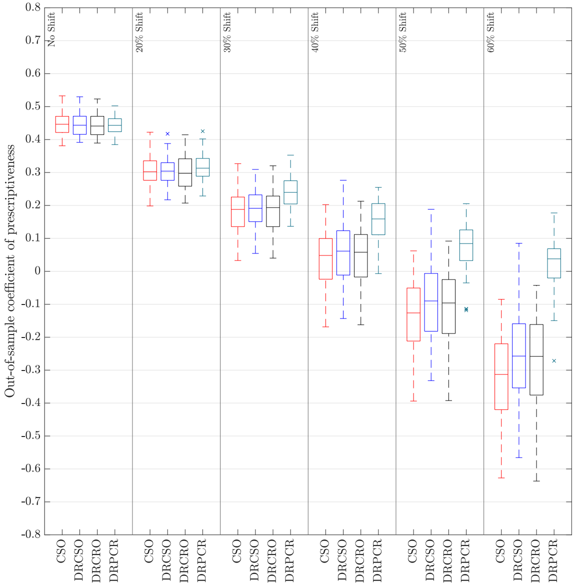
- *Distributionally Robust Contextual Stochastic Optimization (DRCSO)*

$$x^*(\zeta) \in \arg\min_{x \in \mathcal{X}} \sup_{F_{\xi|\zeta} \in \bar{\mathcal{D}}(\hat{F}_{\xi|\zeta}, \alpha)} \mathbb{E}_{F_{\xi|\zeta}}[x^\top \xi]$$

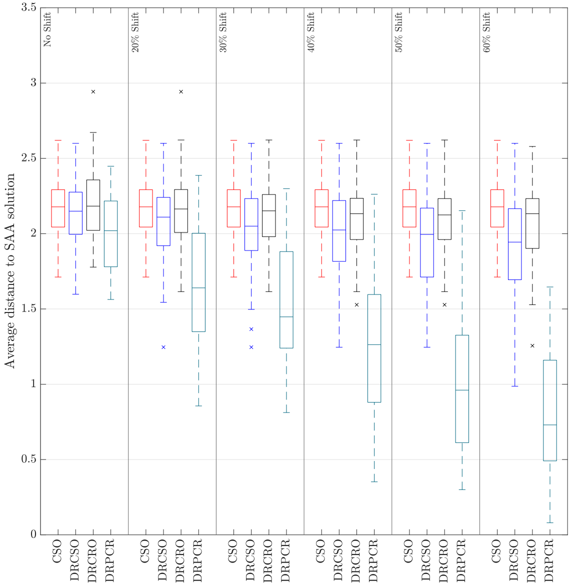
- *Distributionally Robust Contextual Regret Optimization (DRCRO)*

$$x^*(\zeta) \in \arg\min_{x \in \mathcal{X}} \sup_{F_{\xi|\zeta} \in \bar{\mathcal{D}}(\hat{F}_{\xi|\zeta}, \alpha)} \mathbb{E}_{F_{\xi|\zeta}}[x^\top \xi - \min_{x' \in \mathcal{X}} x'^\top \xi]$$

Out-of-Sample Coefficient of Prescriptiveness



AVERAGE L-1 NORM DISTANCE TO SAA SOLUTION



Take-away

- Under the nested CVaR ambiguity set, optimization of the coefficient of prescriptiveness in the DRO context leads to the special case of solving a series of linear programs.
- Roughly speaking, when the mean of the unobserved random variable is exposed to a distribution shift, the out-of-sample coefficients of prescriptiveness achieved by DRPCR policies are higher than those obtained by the alternative methods.