Data driven with Features

Runyu Tang

March 8, 2023

Usage of big data

"4V" characteristics of big data:

- Volume: large amount of data.
- Value: low value of data.
- Velocity: fast data processing.
- Variety: different types of data.

Newsvendor problem

The conventional Newsvendor problem:

$$\min_{q \geq 0} \quad EC(q) := \mathbb{E}[C(q; D)],$$

where

$$C(q; D) := b(D - q)^{+} + h(q - D)^{+}.$$

Newsvendor problem

The conventional Newsvendor problem:

$$\min_{q\geq 0} \quad EC(q) := \mathbb{E}[C(q; D)],$$

where

$$C(q; D) := b(D - q)^{+} + h(q - D)^{+}.$$

The optimal decision:

$$q^{*} = \inf\{y : F(y) \geq \frac{b}{b+h}\}.$$

Data-driven Newsvendor

Sample Average Approximation (SAA):
(With historical demand data $\mathbf{d}(n) = [d_1, d_2, \ldots, d_n]$.)

$$\min_{q \geq 0} \hat{R}(q; \mathbf{d}(n)) = \frac{1}{n} \sum_{i=1}^{n} \left[b\left(d_i - q\right)^+ + h\left(q - d_i\right)^+ \right], \qquad \text{(SAA)}$$

Data-driven Newsvendor

Sample Average Approximation (SAA):
(With historical demand data $\mathbf{d}(n) = [d_1, d_2, \ldots, d_n]$.)

$$\min_{q \geq 0} \hat{R}(q; \mathbf{d}(n)) = \frac{1}{n} \sum_{i=1}^{n} \left[b \left(d_i - q \right)^+ + h \left(q - d_i \right)^+ \right], \qquad \text{(SAA)}$$

The optimal decision:

$$\hat{q}^* = \inf\{y : \hat{F}_n(y) \geq \frac{b}{b+h}\},$$

which is the $\lceil n \frac{b}{b+h} \rceil$th largest demand observation.

SAA

Asymptotic convergence: As the number of data points $N \to \infty$, both the optimal value z^{SAA} and the optimal solution x^{SAA} converge to the optimal value z^* and the optimal solution x^* almost surely.
Tractability: For many cost functions $c(x;\xi)$ and sets $\mathcal{X}$, finding the optimal value of and an optimal solution SAA is computationally tractable.

But we haven't considered the "big" data yet.

The Feature-based Newsvonder

$$\min_{q(\cdot)\in\mathscr{Q},\{q:\mathscr{X}\to\mathbb{R}\}} \mathbb{E}[C(q(\mathbf{x});D(\mathbf{x}))\mid\mathbf{x}],$$

where $\mathbf{x}$ is the feature, and the decision is a map from the feature space to the reals.

Empirical Risk Minimization

The ERM approach to solving the newsvendor problem with feature data is

$$\min_{q(\cdot)\in\mathscr{Q},\{q:\mathscr{X}\to\mathbb{R}\}} \hat{R}\left(q(\cdot);S_n\right)=\frac{1}{n}\sum_{i=1}^{n}\left[b\left(d_i-q\left(\mathbf{x}_i\right)\right)^{+}+h\left(q\left(\mathbf{x}_i\right)-d_i\right)^{+}\right], \quad (\mathrm{NV}-\mathrm{ERM})$$

where $\hat{R}$ is called the empirical risk of function q with respect to the data set S_n.
Problem: function class $\mathscr{Q}$?

Empirical Risk Minimization

The ERM approach to solving the newsvendor problem with feature data is

$$\min_{q(\cdot)\in\mathscr{Q},\{q:\mathscr{X}\to\mathbb{R}\}} \hat{R}\left(q(\cdot);S_n\right)=\frac{1}{n}\sum_{i=1}^{n}\left[b\left(d_i-q\left(\mathbf{x}_i\right)\right)^{+}+h\left(q\left(\mathbf{x}_i\right)-d_i\right)^{+}\right], \quad (\mathrm{NV-ERM})$$

where $\hat{R}$ is called the empirical risk of function q with respect to the data set S_n.

Problem: function class $\mathscr{Q}$?

$\Rightarrow$ **Linear Decision Rule!**

$$\mathscr{Q}=\left\{q:\mathscr{X}\to\mathbb{R}:q(\mathbf{x})=\mathbf{q}'\mathbf{x}=\sum_{j=1}^{p}q^{j}x^{j}\right\}.$$

ERM algorithm 1

$$\begin{aligned}
\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \quad & \hat{R}\left(q(\cdot);S_n\right)=\frac{1}{n}\sum_{i=1}^{n}\left[b\left(d_i-q\left(\mathbf{x}_i\right)\right)^{+}+h\left(q\left(\mathbf{x}_i\right)-d_i\right)^{+}\right]\\
\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad & \frac{1}{n}\sum_{i=1}^{n}\left(bu_i+ho_i\right)\\
\text{s.t.} \quad & u_i \geq d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j\\
& o_i \geq q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i\\
& u_i, o_i \geq 0,
\end{aligned}$$

ERM algorithm 2

Add a regularization term:

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \hat{R}\left(q(\cdot);S_n\right)+\lambda\|\mathbf{q}\|_k^2=\frac{1}{n}\sum_{i=1}^{n}\left[b\left(d_i-q\left(\mathbf{x}_i\right)\right)^{+}+h\left(q\left(\mathbf{x}_i\right)-d_i\right)^{+}\right]+\lambda\|\mathbf{q}\|_k^2$$

$$\begin{aligned}
\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad & \frac{1}{n}\sum_{i=1}^{n}\left(bu_i+ho_i\right)+\lambda\|\mathbf{q}\|_k^2 \\
\text{s.t.} \quad & u_i \geq d_i-q^1-\sum_{j=2}^{p} q^j x_i^j \\
& o_i \geq q^1+\sum_{j=2}^{p} q^j x_i^j-d_i \\
& u_i, o_i \geq 0,
\end{aligned}$$

Kernel Optimization

predictive prescription: Any function $q(x)$ that prescribes a decision in anticipation of the future given the observation $\mathbf{x}$.

$$\tilde{q}(\mathbf{x}) = \arg\min \sum_{i=1}^{n} w_{n,i}(\mathbf{x}) C(q, d_i)$$

based on Nadaraya-Watson kernel regression

$$\min_{q \geq 0} \tilde{R}\left(q; S_n, \mathbf{x}_{n+1}\right) = \min_{q \geq 0} \frac{\sum_{i=1}^{n} K_w\left(\mathbf{x}_{n+1} - \mathbf{x}_i\right) C\left(q, d_i\right)}{\sum_{i=1}^{n} K_w\left(\mathbf{x}_{n+1} - \mathbf{x}_i\right)}$$

where $K_w(\cdot)$ is a kernel function with bandwidth w.

Kernel Regression

General regression: Given past data $(x_1, y_1), \ldots, (x_n, y_n)$, one wants to estimate the conditional expectation function (CEF): $E[Y|X]$;
In statistics, kernel regression is a non-parametric technique to estimate the conditional expectation of a random variable. The objective is to find a non-linear relation between a pair of random variables X and Y.

Kernel regression

```
# R code
install.packages("np")
library(np) # non parametric library
data(cps71)
attach(cps71)

m <- npreg(logwage˜age)

plot(m, plot.errors.method="asymptotic",
    plot.errors.style="band",
    ylim=c(11, 15.2))

points(age, logwage, cex=.25)
```

Typical kernel functions

- Uniform kernel:

$$K(\mathbf{u}) = \frac{1}{2}\mathbb{I}(\|\mathbf{u}\|_2 \leq 1)$$

- Gaussian kernel:

$$K(\mathbf{u}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\mathbf{u}\|_2^2}{2}\right).$$

-

$$K_w(\cdot) := K(\cdot/w)/w$$

where w is the bandwidth of the kernel.

NV Kernel Optimization

Proposition.

The optimal feature-based newsvendor decision $\hat{q}_n^k$ obtained by solving (NV-KO) is given by

$$\hat{q}_n^k = \hat{q}_n^k(\mathbf{x}_{n+1}) = \inf\left\{q : \frac{\sum_{i=1}^n k_i \mathbb{I}(d_i \leq q)}{\sum_{i=1}^n k_i} \geq \frac{b}{b+h}\right\} \tag{1}$$

where $k_i = K_w(\mathbf{x}_{n+1} - \mathbf{x}_i)$ and $K_w(\cdot)$ is a kernel function with bandwidth w. In other words, we can find $\hat{q}_n^k$ by ranking the past demand in increasing order and choosing the smallest value at which the inequality in (1) is satisfied.

Value of feature information

Two population example:

Demand model

$$D = D_0(1 - x) + D_1 x$$

where D_0 and D_1 are nonnegative continuous random variables such that the corresponding critical newsvendor fractiles q_0^* and q_1^* follow $q_0^* < q_1^*$.

We have n historical observations: $[(x_1, d_1), \dots, (x_n, d_n)]$, of which $n_0 = np_0$ when $x = 0$ and $n_1 = n - n_0$ when $x = 1$ (assume rounding effects are negligible).

Value of feature information

Two population example:

Theorem 1 (Asymptotic Optimality of (NV-ERM1)).

We can show

$$\left|\mathbb{E}\left[\hat{q}_n^i\right] - F_0^{-1}(r)\right| \leq O\left(\frac{\log n_i}{n_i}\right), i = 0, 1$$

i.e. the finite-sample decision of the feature-based decision is biased by at most $O(\log n_i / n_i), i = 0, 1,$ *and*

$$\lim_{n\to\infty} \hat{q}_n^i \overset{a.s.}{=} F_0^{-1}(r) =: q_i^*, \quad i = 0, 1$$

i.e. the feature-based decision is asymptotically optimal, correctly identifying the case when $x = 0$ *or* 1 *as the number of observations goes to infinity.*

Value of feature information

Two population example:

Theorem 2 (Asymptotic (Sub)Optimality of (SAA)).

The finite-sample bias of the SAA decision is given by

$$\left|\mathbb{E}\left[\hat{q}_n^{SAA}\right] - (F^{mix})^{-1}(r)\right| \leq O\left(\frac{\log n}{n}\right)$$

we also have

$$\left|\mathbb{E}\left[\hat{q}_n^{SAA} - \hat{q}_n^{0}\right]\right| = \left|\left(F^{mix}\right)^{-1}(r) - F_0^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1)$$

$$\left|\mathbb{E}\left[\hat{q}_n^{1} - \hat{q}_n^{SAA}\right]\right| = \left|F_1^{-1}(r) - \left(F^{mix}\right)^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1).$$

That is, on average, if $x = 0$ in the next decision period, the SAA decision orders too much, and if $x = 1$, the SAA decision orders too little. .

Value of feature information

Linear demand example:
Demand model

$$D|(\mathbf{X} = \mathbf{x}) = \beta^T \mathbf{x} + \epsilon$$

where ϵ is independent of the (random) feature vector $\mathbf{X}$, is continuous with probability density function $f_\epsilon(\cdot)$.
A DM without the feature information only has access to past demand data: $D = \{d_1, \ldots, d_n\}$; and a DM who has both past feature and demand data has the information: $D_{\mathbf{x}} = (\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_n, d_n)$.

Value of feature information

Linear demand example:

Theorem 3.

Under the linear demand model, given features $\mathbf{X} = \tilde{\mathbf{x}}$,

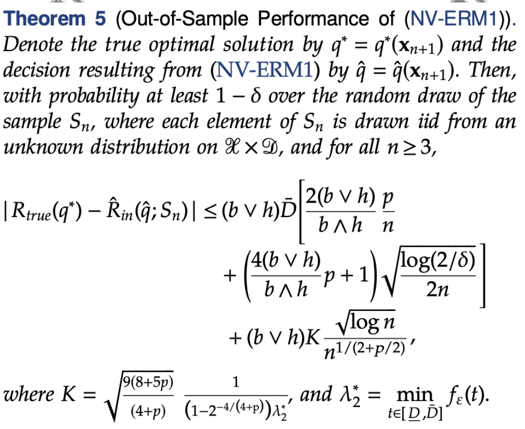
$$\begin{aligned}\hat{q}_n^{SAA}(\tilde{\mathbf{x}}) &\overset{a.s.}{\to} Q_\varepsilon\left(\frac{b}{b+h}\right) + \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_\varepsilon[D \mid \mathbf{X}]\right] \\ &= Q_\varepsilon\left(\frac{b}{b+h}\right) + \beta^\top \mathbb{E}[\mathbf{X}]\end{aligned}$$

and

$$\hat{q}^{DM2}(\tilde{\mathbf{x}}) \overset{a.s.}{\to} Q_\varepsilon\left(\frac{b}{b+h}\right) + \beta^\top \tilde{\mathbf{x}} = q^*(\tilde{\mathbf{x}})$$

as n *tends to infinity.*

Bounds



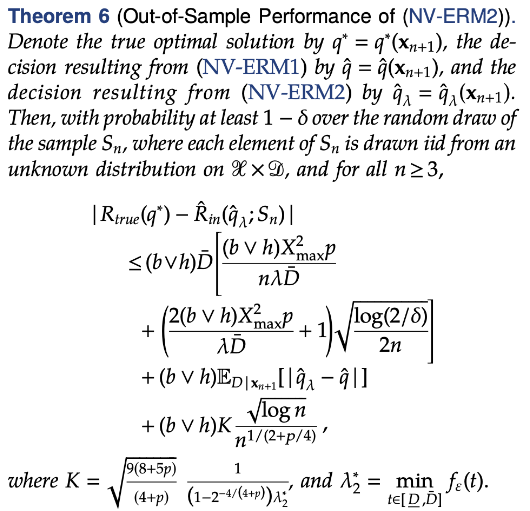
Theorem 5 (Out-of-Sample Performance of (NV-ERM1)). *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$ and the decision resulting from (NV-ERM1) by $\hat{q} = \hat{q}(\mathbf{x}_{n+1})$. Then, with probability at least $1-\delta$ over the random draw of the sample S_n, where each element of S_n is drawn iid from an unknown distribution on $\mathcal{X}\times\mathcal{D}$, and for all $n\ge 3$,*

$$|R_{true}(q^*) - \hat{R}_{in}(\hat{q};S_n)| \le (b\vee h)\bar{D}\left[\frac{2(b\vee h)}{b\wedge h}\frac{p}{n} + \left(\frac{4(b\vee h)}{b\wedge h}p+1\right)\sqrt{\frac{\log(2/\delta)}{2n}}\right] + (b\vee h)K\frac{\sqrt{\log n}}{n^{1/(2+p/2)}},$$

where $K = \sqrt{\frac{9(8+5p)}{(4+p)}}\,\frac{1}{(1-2^{-4/(4+p)})\lambda_2^*}$, *and* $\lambda_2^* = \min_{t\in[\underline{D},\bar{D}]} f_\varepsilon(t)$.

- The first term: generalization error. Scales as $O(p/\sqrt{n})$, decays exponentially fast in n.
- The second term: finite sample bias. The rate $n^{-1/(2+p/2)}\sqrt{\log n}$ is optimal.

Bounds



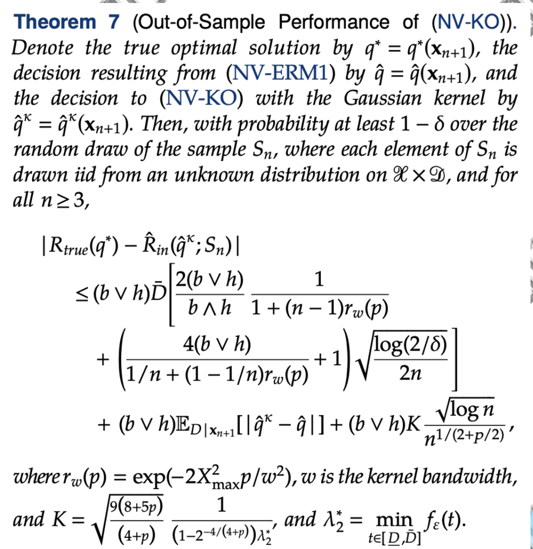
Theorem 6 (Out-of-Sample Performance of (NV-ERM2)). *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$, the decision resulting from* (NV-ERM1) *by $\hat{q} = \hat{q}(\mathbf{x}_{n+1})$, and the decision resulting from* (NV-ERM2) *by $\hat{q}_\lambda = \hat{q}_\lambda(\mathbf{x}_{n+1})$. Then, with probability at least $1 - \delta$ over the random draw of the sample S_n, where each element of S_n is drawn iid from an unknown distribution on $\mathcal{X} \times \mathcal{D}$, and for all $n \geq 3$,*

$$
\begin{aligned}
&|R_{true}(q^*) - \hat{R}_{in}(\hat{q}_\lambda; S_n)| \\
&\quad \leq (b \vee h)\bar{D}\left[\frac{(b \vee h)X_{\max}^2 p}{n\lambda\bar{D}}\right. \\
&\qquad \left. + \left(\frac{2(b \vee h)X_{\max}^2 p}{\lambda\bar{D}} + 1\right)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \\
&\qquad + (b \vee h)\mathbb{E}_{D|\mathbf{x}_{n+1}}[|\hat{q}_\lambda - \hat{q}|] \\
&\qquad + (b \vee h)K\frac{\sqrt{\log n}}{n^{1/(2+p/4)}},
\end{aligned}
$$

where $K = \sqrt{\frac{9(8+5p)}{(4+p)}}\frac{1}{(1-2^{-4/(4+p)})\lambda_2^*}$, *and* $\lambda_2^* = \min_{t \in [\underline{D}, \bar{D}]} f_\varepsilon(t)$.

- The first term: generalization error. Scales as $O(p/(\lambda\sqrt{n}))$. $\lambda = O(1/p^2)$ is a good starting point, because it gives the same error rate as ERM-1.
- The second term: in-sample decision resulting from regularization — the bias resulting from having perturbed the optimization problem away from the true problem of interest.
- The third term: finite-sample bias.

Bounds



Theorem 7 (Out-of-Sample Performance of (NV-KO)). *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$, the decision resulting from (NV-ERM1) by $\hat{q} = \hat{q}(\mathbf{x}_{n+1})$, and the decision to (NV-KO) with the Gaussian kernel by $\hat{q}^\kappa = \hat{q}^\kappa(\mathbf{x}_{n+1})$. Then, with probability at least $1-\delta$ over the random draw of the sample S_n, where each element of S_n is drawn iid from an unknown distribution on $\mathscr{X} \times \mathfrak{D}$, and for all $n \geq 3$,*

$$\begin{aligned}
&|R_{true}(q^*) - \hat{R}_{in}(\hat{q}^\kappa; S_n)| \\
&\quad \leq (b \vee h)\bar{D}\left[\frac{2(b \vee h)}{b \wedge h} \frac{1}{1+(n-1)r_w(p)}\right. \\
&\qquad + \left.\left(\frac{4(b \vee h)}{1/n + (1-1/n)r_w(p)} + 1\right)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \\
&\qquad + (b \vee h)\mathbb{E}_{D|\mathbf{x}_{n+1}}[|\hat{q}^\kappa - \hat{q}|] + (b \vee h)K\frac{\sqrt{\log n}}{n^{1/(2+p/2)}},
\end{aligned}$$

where $r_w(p) = \exp(-2X_{\max}^2 p/w^2)$, w is the kernel bandwidth, and $K = \sqrt{\frac{9(8+5p)}{(4+p)}} \frac{1}{(1-2^{-4/(4+p)})\lambda_2^*}$, and $\lambda_2^* = \min_{t \in [\underline{D}, \bar{D}]} f_\varepsilon(t)$.

- The first term: generalization error. Scales as $O(p/(r_w(p)\sqrt{n}))$, which can be controlled by the bandwidth w. Setting $w = O(\sqrt{p})$ gives an error of $O(1/\sqrt{n})$ which is as good as demand without features.
- The second term: the bias resulting from optimizing with a scalar decision.
- The third term: finite-sample bias.

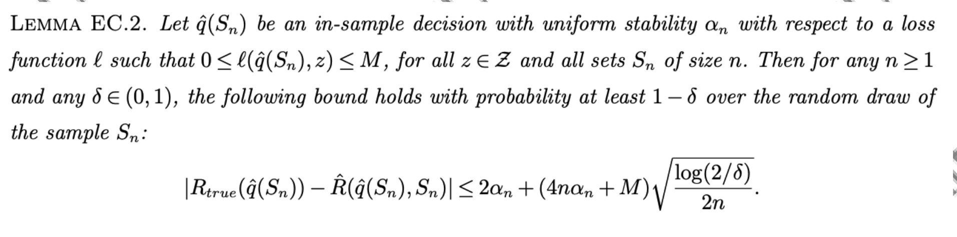
Bounds

$$\mathbb{P}_{S_n}\left[\left|R_{\text{true}}\ \left(q^*\right)-\hat{R}\left(\hat{q};S_n\right)\right|\leq O\left(\frac{p\sqrt{\log(1/\delta)}}{\sqrt{n}}+\frac{\sqrt{\log n}}{n^{\frac{1}{2+p/2}}}\right)\right]\geq 1-\delta$$

$$\mathbb{P}_{S_n}\left[\left|R_{\text{true}}\ \left(q^*\right)-\hat{R}\left(\hat{q}_\lambda;S_n\right)\right|\leq O\left(\frac{p\sqrt{\log(1/\delta)}}{\lambda\sqrt{n}}+\frac{\sqrt{\log n}}{n^{\frac{1}{2+p/4}}}+\mathbb{E}_{D|\mathbf{x}_{n+1}}\left|\hat{q}_\lambda-\hat{q}\right|\right)\right]\geq 1-\delta$$

$$\mathbb{P}_{S_n}\left[\left|R_{\text{true}}\ \left(q^*\right)-\hat{R}\left(\hat{q}^\kappa;S_n\right)\right|\leq O\left(\frac{\sqrt{\log(1/\delta)}}{r_w(p)\sqrt{n}}+\frac{\sqrt{\log n}}{n^{\frac{1}{2+p/2}}}+\mathbb{E}_{D|\mathbf{x}_{n+1}}\left|\hat{q}^\kappa-\hat{q}\right|\right)\right]\geq 1-\delta$$

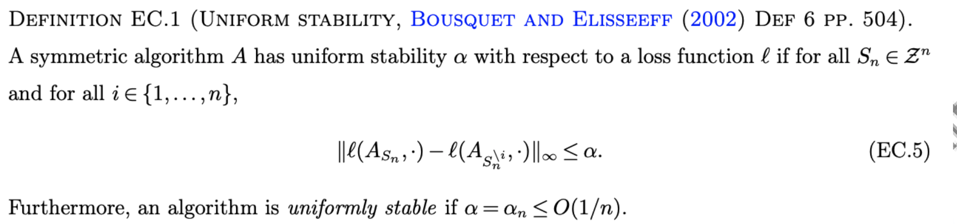
How to prove such bounds?



Lemma EC.2. *Let $\hat{q}(S_n)$ be an in-sample decision with uniform stability α_n with respect to a loss function ℓ such that $0 \le \ell(\hat{q}(S_n), z) \le M$, for all $z \in \mathcal{Z}$ and all sets S_n of size n. Then for any $n \ge 1$ and any $\delta \in (0,1)$, the following bound holds with probability at least $1-\delta$ over the random draw of the sample S_n:*

$$|R_{true}(\hat{q}(S_n)) - \hat{R}(\hat{q}(S_n), S_n)| \le 2\alpha_n + (4n\alpha_n + M)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Uniform stability:



Definition EC.1 (Uniform stability, Bousquet and Elisseeff (2002) Def 6 pp. 504). A symmetric algorithm A has uniform stability α with respect to a loss function ℓ if for all $S_n \in \mathcal{Z}^n$ and for all $i \in \{1, \ldots, n\}$,

$$\|\ell(A_{S_n}, \cdot) - \ell(A_{S_n^{\setminus i}}, \cdot)\|_\infty \le \alpha. \qquad \text{(EC.5)}$$

Furthermore, an algorithm is *uniformly stable* if $\alpha = \alpha_n \le O(1/n)$.

How to prove such bounds?

Uniform stability:

- ERM1: $\alpha_n = \frac{\bar{D}(b \vee h)^2}{(b \wedge h)} \frac{p}{n}$.
- ERM2: $\alpha_n^r = \frac{X_{\max}^2 (b \vee h)^2}{2\lambda} \frac{p}{n}$.
- KO: $\alpha_\kappa = \frac{\bar{D}(b \vee h)^2}{(b \wedge h)} \frac{1}{1+(n-1)r_w}$, where $r_w = \exp(-2X_{\max}^2/w^2)$.

Bousquet O, Elisseeff A (2002) Stability and generalization. J. Mach.Learn. Res. 2(Mar):499-526.

Data source: the emergency room of a large teaching hospital in the United Kingdom from July 2008 to June 2009.

- Optimal staffing levels of nurses for a hospital emergency room.
- Agency nurse v.s. regular nurse.
- Features:
  - the first set being the day of the week, time of the day, and m number of days of past demands
  - the second set being the first set plus the sample average of past demands and the differences in the order statistics of past demands. (*operational statistics* (OS) *features*.)

Numerical Experiments

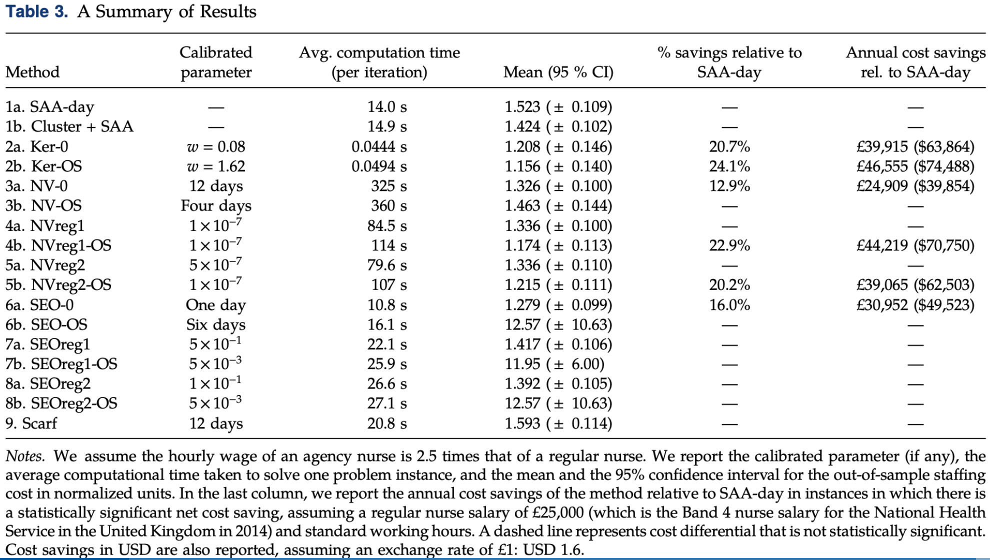


Table 3. A Summary of Results

Method	Calibrated parameter	Avg. computation time (per iteration)	Mean (95 % CI)	% savings relative to SAA-day	Annual cost savings rel. to SAA-day
1a. SAA-day	—	14.0 s	1.523 (± 0.109)	—	—
1b. Cluster + SAA	—	14.9 s	1.424 (± 0.102)	—	—
2a. Ker-0	$w = 0.08$	0.0444 s	1.208 (± 0.146)	20.7%	£39,915 ($63,864)
2b. Ker-OS	$w = 1.62$	0.0494 s	1.156 (± 0.140)	24.1%	£46,555 ($74,488)
3a. NV-0	12 days	325 s	1.326 (± 0.100)	12.9%	£24,909 ($39,854)
3b. NV-OS	Four days	360 s	1.463 (± 0.144)	—	—
4a. NVreg1	1×10^{-7}	84.5 s	1.336 (± 0.100)	—	—
4b. NVreg1-OS	1×10^{-7}	114 s	1.174 (± 0.113)	22.9%	£44,219 ($70,750)
5a. NVreg2	5×10^{-7}	79.6 s	1.336 (± 0.110)	—	—
5b. NVreg2-OS	1×10^{-7}	107 s	1.215 (± 0.111)	20.2%	£39,065 ($62,503)
6a. SEO-0	One day	10.8 s	1.279 (± 0.099)	16.0%	£30,952 ($49,523)
6b. SEO-OS	Six days	16.1 s	12.57 (± 10.63)	—	—
7a. SEOreg1	5×10^{-1}	22.1 s	1.417 (± 0.106)	—	—
7b. SEOreg1-OS	5×10^{-3}	25.9 s	11.95 (± 6.00)	—	—
8a. SEOreg2	1×10^{-1}	26.6 s	1.392 (± 0.105)	—	—
8b. SEOreg2-OS	5×10^{-3}	27.1 s	12.57 (± 10.63)	—	—
9. Scarf	12 days	20.8 s	1.593 (± 0.114)	—	—

Notes. We assume the hourly wage of an agency nurse is 2.5 times that of a regular nurse. We report the calibrated parameter (if any), the average computational time taken to solve one problem instance, and the mean and the 95% confidence interval for the out-of-sample staffing cost in normalized units. In the last column, we report the annual cost savings of the method relative to SAA-day in instances in which there is a statistically significant net cost saving, assuming a regular nurse salary of £25,000 (which is the Band 4 nurse salary for the National Health Service in the United Kingdom in 2014) and standard working hours. A dashed line represents cost differential that is not statistically significant. Cost savings in USD are also reported, assuming an exchange rate of £1: USD 1.6.

From predictive to prescriptive analytics

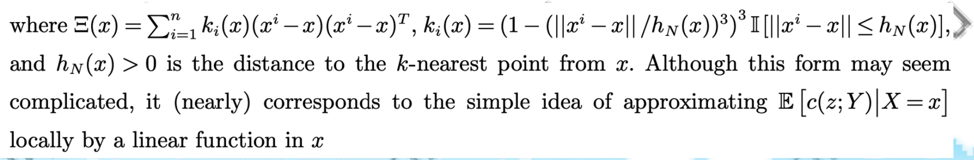
- $k-$nearest-neighbors (kNN) regression:

$$\hat{q}_n^{k\text{NN}}(x) = \arg\min \sum_{i\in\mathcal{N}_k(x)} C(q, d_i)$$

where $\mathcal{N}_k(x)$ is the neighborhood of the k data points that are close to x.

- local linear regression:

$$\hat{q}_n^{\text{LOESS}}(x) = \arg\min \sum_{i}^{n} k_i(x) \max\left\{1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^\top \Xi(x)^{-1}(x^i - x), 0\right\} C(q, d$$



where $\Xi(x) = \sum_{i=1}^{n} k_i(x)(x^i - x)(x^i - x)^T$, $k_i(x) = (1 - (||x^i - x|| / h_N(x))^3)^3\, \mathbb{I}\left[||x^i - x|| \le h_N(x)\right]$, and $h_N(x) > 0$ is the distance to the k-nearest point from x. Although this form may seem complicated, it (nearly) corresponds to the simple idea of approximating $\mathbb{E}\left[c(z;Y)\middle|X = x\right]$ locally by a linear function in x

From predictive to prescriptive analytics

- classification and regression tree (CART):

$$\hat{q}_n^{\text{CART}}(x) = \arg\min \sum_{i:R(x^i)=R(x)} C(q, d_i),$$

where $R(x)$ is the binning rule implied by a regression tree.

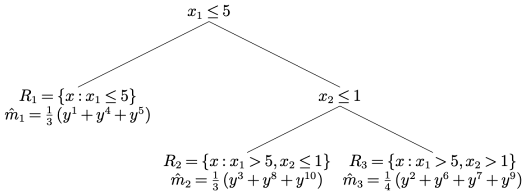
- random forests:

$$\hat{q}_n^{\text{RF}}(x) = \arg\min \sum_t \frac{1}{\{j : R^t(x^j) = R^t(x)\}} \sum_{i:R^t(x^i)=R^t(x)} C(q, d_i),$$

where $R^t(x)$ is the binning rule implied by the t^{th} tree in a random forest

From predictive to prescriptive analytics

Figure 1 **A regression tree is trained on data $\{(x^1, y^1), \ldots, (x^{10}, y^{10})\}$ and partitions the X data into regions defined by the leaves. The Y prediction $\hat{m}(x)$ is $\hat{m}_j$, the average of Y data at the leaf in which $X = x$ ends up. The implicit binning rule is $R(x)$, which maps x to the identity of the leaf in which it ends up.**



$x_1 \leq 5$

$R_1 = \{x : x_1 \leq 5\}$
$\hat{m}_1 = \frac{1}{3}(y^1 + y^4 + y^5)$

$x_2 \leq 1$

$R_2 = \{x : x_1 > 5, x_2 \leq 1\}$
$\hat{m}_2 = \frac{1}{3}(y^3 + y^8 + y^{10})$

$R_3 = \{x : x_1 > 5, x_2 > 1\}$
$\hat{m}_3 = \frac{1}{4}(y^2 + y^6 + y^7 + y^9)$

Implicit binning rule:

$$R(x) = (j \text{ s.t. } x \in R_j)$$

From predictive to prescriptive analytics

Asymptotic optimality:

Theorem.

Under some mild conditions, kNN, Kernel Methods, Local Linear Methods are asymptotic optimal and consistent.

Although no firm theoretical results on the asymptotic optimality of the predictive prescriptions based on CART and RF, we have observed them to converge empirically.

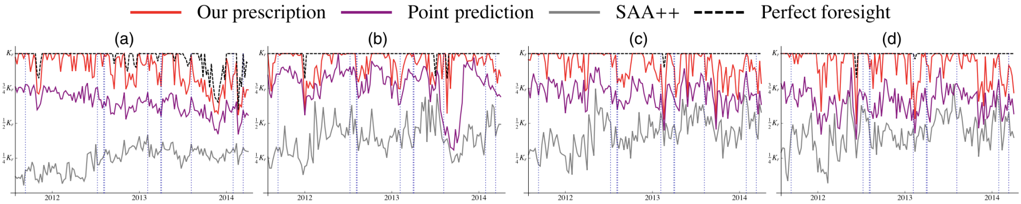
From predictive to prescriptive analytics

Coefficient of prescriptiveness (in the test set):

$$P = 1 - \frac{\text{PredPres} - \text{TrueOpt}}{\text{SAA} - \text{TrueOpt}}$$

Numerical experiments:

Figure 9. Performance of Our Prescription over Time



Munich, $P = 0.89$ Paris, $P = 0.90$ Waterloo, $P = 0.85$ The Hague, $P = 0.86$

Notes. Vertical dashes indicate major release dates. The vertical axis is shown in terms of the location's capacity, K_r.

Smart “predict, then optimize”

Contextual optimization problem:

$$\min_{q \in S} E[c^\top q | x]$$

which is equivalent to (by linearity of expectation)

$$\min_{q \in S} E[c|x]^\top q$$

We need to estimate $E[c|x]$.

Video:
`https://www.youtube.com/watch?v=Hot26kyykaI&ab_channel=AdamElmachtoub`

Smart "predict, then optimize"

Standard solution approach "predict, then optimize":

- 1. Predict parameters using a machine learning model.
- 2. Plug in predictions into optimization model and solve it.

Key gradients:

- 1.Nominal (downstream) optimization problem

$$P(c) : z^*(c) := \min_{q \in S} c^\top q.$$

- 2. Training data: $(x_1, c_1), (x_2, c_2), \ldots, (x_n, c_n)$.
- 3. Hypothesis class: $\hat{c} = f(x)$.
- 4. Loss function: $l(\hat{c}, c)$.

Smart “predict, then optimize”

Prediction: Find f^* using ERM principal:

$$\min_f \frac{1}{N}\sum_{i=1}^{N} l(f(x_i), c_i).$$

Optimization: Given a new x, make decision

$$q^*(f^*(x))$$

For common linear regression (least squares):

$$\min_\beta \frac{1}{N}\sum_{i=1}^{N} \|\beta^\top x_i, c_i\|^2.$$

Optimization: Given a new x, make decision

$$q^*(\beta^{*\top} x)$$

Smart “predict, then optimize”

SPO Goal: Minimize decision error rather than prediction error.
SPO Loss function:

$$l_{\text{SPO}}(\hat{c}, c) := c^\top q^*(\hat{c}) - c^\top q^*(c).$$

Then the prediction part becomes:

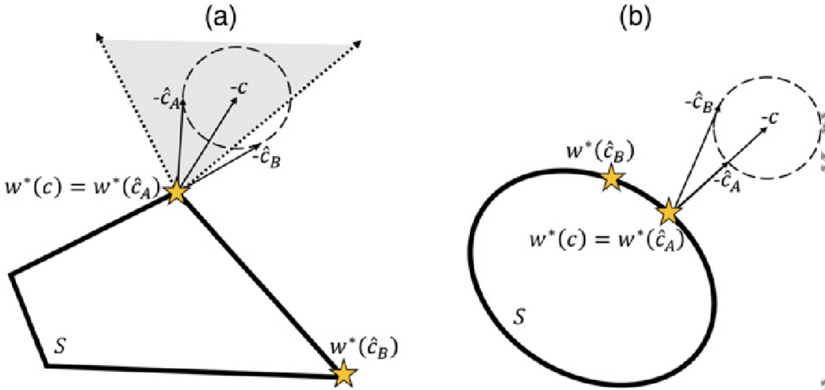
$$\min_f \frac{1}{N} \sum_{i=1}^{N} l_{\text{SPO}}(f(x_i), c_i).$$

Recall in big data newsvendor:

$$\min_q \frac{1}{N} \sum_{i=1}^{n} (c_i^\top | \boldsymbol{x}) q$$

Smart “predict, then optimize”

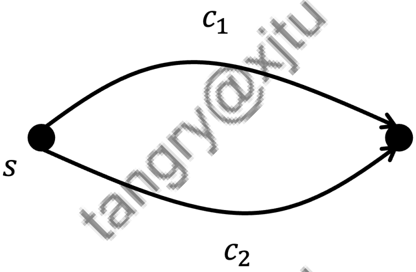
Figure 1. Geometric Illustration of SPO Loss



Notes. (a) Polyhedral feasible region. (b) Elliptic feasible region.

An example

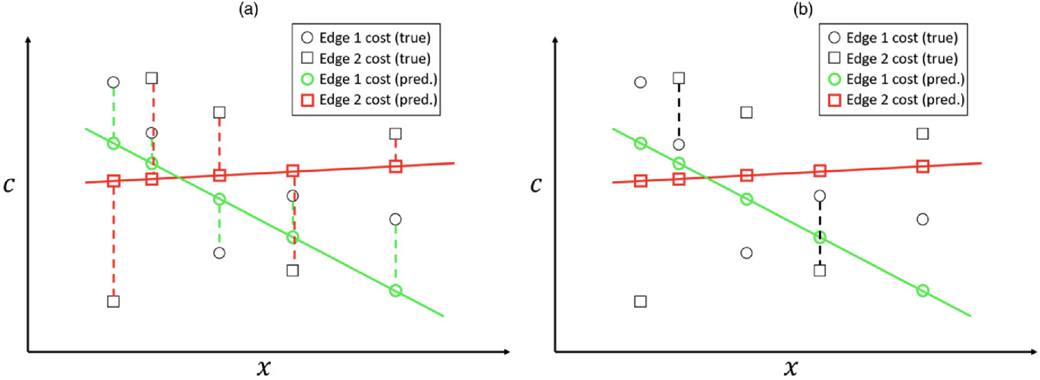
Consider a shortest-path problem with two nodes, s and t. There are two edges that go from s to t, edge 1 and edge 2. Thus the cost vector c is two-dimensional.



Our data consist of (x_i, c_i) pairs, and c_i are generated nonlinearly as a function of x_i.

An example

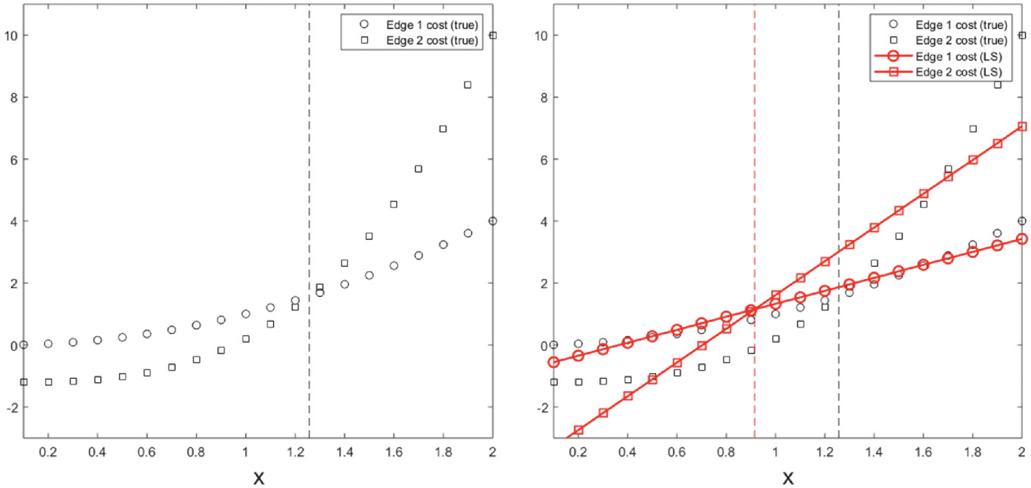
Figure 2. Difference Between Prediction and Decision Residuals



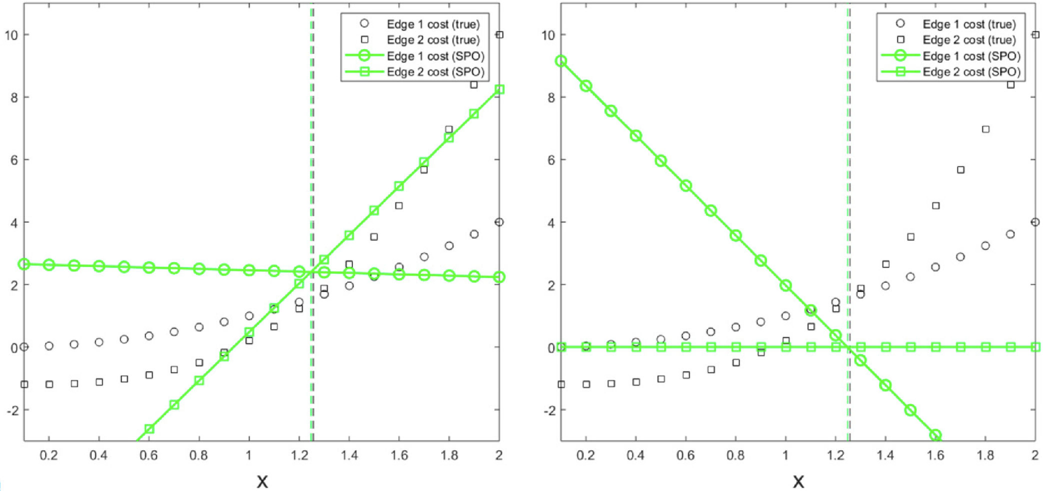
Notes. (a) Prediction residuals. (b) Decision residuals. Pred., prediction.

An example

Figure 3. Illustrative Example



An example



10
8
6
4
2
0
-2
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
X
○ Edge 1 cost (true)
□ Edge 2 cost (true)
Edge 1 cost (SPO)
Edge 2 cost (SPO)
10
8
6
4
2
0
-2
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
X
○ Edge 1 cost (true)
□ Edge 2 cost (true)
Edge 1 cost (SPO)
Edge 2 cost (SPO)

Smart "predict, then optimize"

$$l_{\text{SPO}}(\hat{c}, c) := c^\top w^*(\hat{c}) - c^\top w^*(c).$$

The SPO loss function is nonconvex and can be discontinuous.
SPO+ loss function (A convex approximation):

$$l_{\text{SPO+}}(\hat{c}, c) := \max_{w \in S}\{(c - 2\hat{c})^\top w\} + 2\hat{c}^\top w^*(\hat{c}) - c^\top w^*(c).$$

Then $l_{\text{SPO+}}(\cdot, c)$ is convex and

$$l_{\text{SPO}}(\cdot, c) \leq l_{\text{SPO+}}(\cdot, c)$$

Smart “predict, then optimize”

(Fisher) Consistency of the SPO+ loss function:

Theorem.

Assume $c|x$ is continuous and is symmetric around its mean. Then minimizing expected SPO+ loss also minimizes expected SPO loss.

$$f^*_{SPO+}(x) = \mathbb{E}[c|x] \in f^*_{SPO}(x).$$

Minimizing the SPO+ loss is equivalent to minimizing the SPO loss.

Reformulation

Suppose $f(x) = Bx$, and $S = \{w : Aw \geq b\}$ is a polytope. Then, the regularized SPO+ ERM Problem is equivalent to the following optimization problem:

$$
\begin{aligned}
\min_{B,p} \frac{1}{n} \sum_{i=1}^{n} & \left[-b^T p_i + 2 \left(w^* \left(c_i \right) x_i^T \right) \bullet B - z^* \left(c_i \right) \right] + \lambda \Omega(B) \\
\text{s.t.} \quad & A^T p_i = 2Bx_i - c_i \quad \text{for} \quad \text{all } i \in \{1, \ldots, n\} \\
& p_i \in \mathbb{R}^m, p_i \geq 0 \quad \text{for all } i \in \{1, \ldots, n\} \\
& B \in \mathbb{R}^{d \times p}.
\end{aligned}
$$

Numerical experiments

We consider a shortest-path problem on a 5×5 grid network, where the goal is to go from the northwest corner to the southeast corner, and the edges only go south or east.

Data generation:

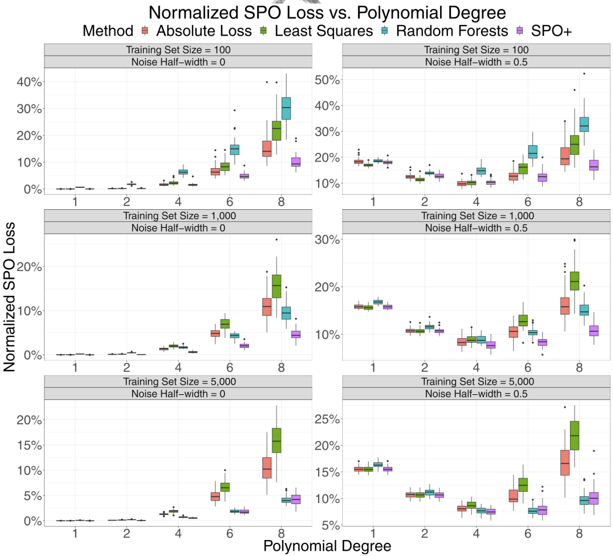
- x_i is generated from a multivariate Gaussian distribution.
- c_i is generated according to

$$c_{ij} = \left[\left(\frac{1}{\sqrt{p}} \left(B^* x_i \right)_j + 3 \right)^{\text{deg}} + 1 \right] \cdot \varepsilon_i^j,$$

where $\varepsilon_i^j \in [1 - \bar{\varepsilon}, 1 + \bar{\varepsilon}]$.

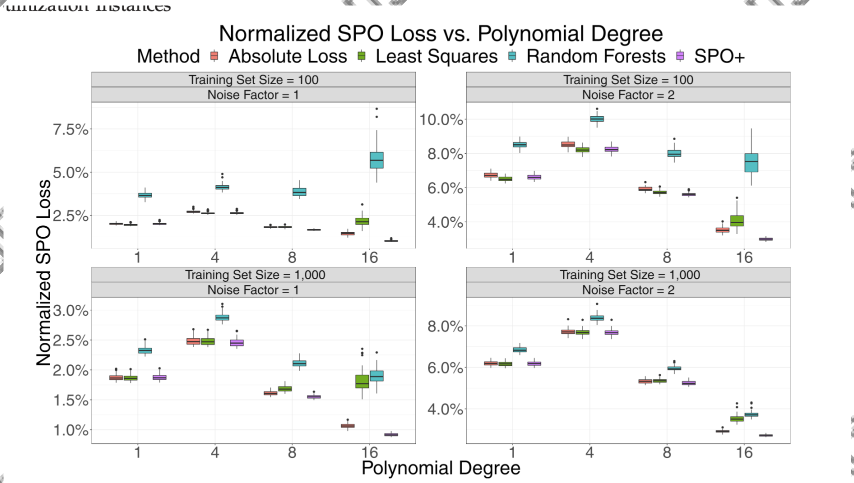
The codes: `https://github.com/paulgrigas/SmartPredictThenOptimize`

Smart “predict, then optimize”



Smart “predict, then optimize”

Portfolio optmization problem:



Reference

- Gah-Yi Ban, Cynthia Rudin. 2018. The Big Data Newsvendor: Practical Insights from Machine Learning. *Operations Research*. (337 citations)
- Dimitris Bertsimas, Nathan Kallus. 2019. From Predictive to Prescriptive Analytics. *Management Science*.
- Adam N. Elmachtoub, Paul Grigas. 2021. Smart “Predict, then Optimize”. *Management Science*. (347 citations)

Applications

Multi-shift staffing problem (MSSP): a company has to staff multiple shifts for each workday in the presence of uncertain arrival rates that vary throughout the day and patient “customers” that do not abandon the queue while waiting for a service, but who must be served by some pre-defined time.

$$\min_{\vec{b}=\{b_s\}} \; C(\vec{b}) := \frac{1}{\tau_{\max}} \int_0^{\tau_{\max}} c_1 b_\tau \, d\tau + c_2 \mathbb{E}\left[N_{\tau_{\max}}\right]$$

$$\text{s.t. } b_\tau := b(\tau) = b_s \text{ for } \tau \in [\tau_{s-1}, \tau_s) \; \forall s = 1, 2, \ldots, S, \quad \text{(MSSP)}$$

Pascal M. Notz, Peter K. Wolf, Richard Pibernika (2023) Prescriptive analytics for a multi-shift staffing problem. *European Journal of Operational Research*. 305 (2023) 887-901.

Applications

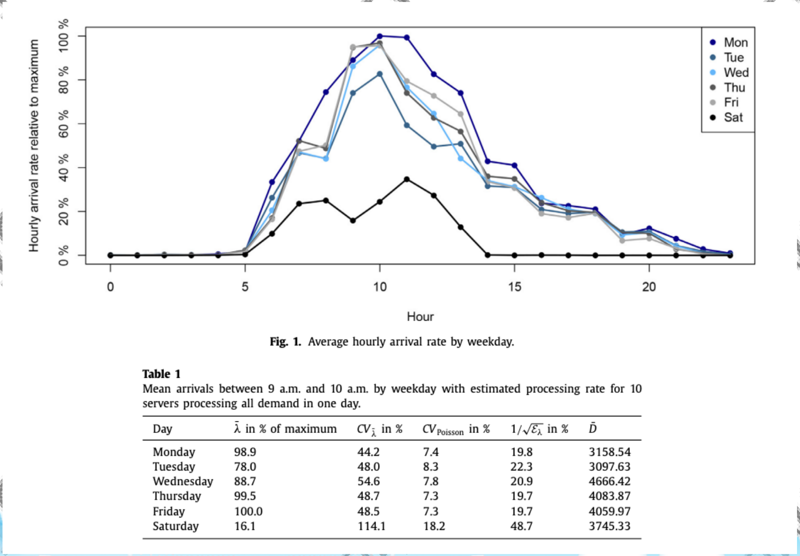


Fig. 1. Average hourly arrival rate by weekday.

Table 1
Mean arrivals between 9 a.m. and 10 a.m. by weekday with estimated processing rate for 10 servers processing all demand in one day.

Day	$\bar{\lambda}$ in % of maximum	$CV_{\bar{\lambda}}$ in %	CV_{Poisson} in %	$1/\sqrt{\mathcal{E}_{\lambda}}$ in %	$\bar{D}$
Monday	98.9	44.2	7.4	19.8	3158.54
Tuesday	78.0	48.0	8.3	22.3	3097.63
Wednesday	88.7	54.6	7.8	20.9	4666.42
Thursday	99.5	48.7	7.3	19.7	4083.87
Friday	100.0	48.5	7.3	19.7	4059.97
Saturday	16.1	114.1	18.2	48.7	3745.33

Applications

AMSSP:

$$\min_{\vec{q}\in\mathcal{Q}} C(\vec{q}) := \frac{1}{T}\sum_{s=1}^{S} c_q \left(t_{s+1}-t_s\right) q_s + c_2\mathbb{E}\left[N_T\right]$$

$$\text{s.t.} N_t = \left(N_{t-1}+D_t-q_s\right)^{+}, \quad \forall t\in\left[t_s,t_{s+1}\right), \forall s=1,\ldots,S.$$

Prescriptive analytics approaches:

- Weighted SAA
- Kernelized ERM
- Optimization prediction approach: which relies only on solving a deterministic optimization problem once and applying a standard machine learning method to predict optimal decision.

Applications

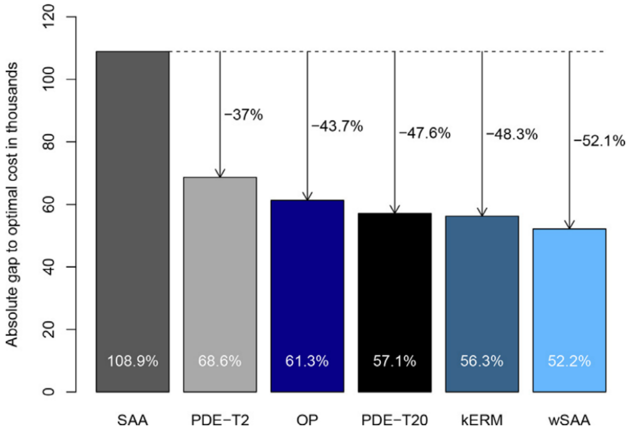


Fig. 2. Absolute gap to optimal cost for realistic cost parameters (Percentage numbers within the bars represent relative difference to ex-post optimal cost).

Suggested readings

- Nathan Kallus, Xiaojie Mao. 2022. Stochastic Optimization Forests. *Management Science* 0(0).
- Luhao Zhang, Jincheng Yang, Rui Gao. 2022. Optimal Robust Policy for Feature-Based Newsvendor, *Management Science* 0(0).
- Ningyuan Chen, Andre A. Cire, Ming Hu, Saman Lagzi (2023) Model-Free Assortment Pricing with Transaction Data. *Management Science* 0(0).
- Huang, Jingkai and Shang, Kevin and Yang, Yi and Zhou, Weihua, 2023. Taylor Approximation of Inventory Policies for One-Warehouse, Multi-Retailer Systems with Demand Feature Information. *SSRN*