The Big Data Newsvendor: Practical Insights from Machine Learning

Cynthia Rudin
Sloan School of Management, Operations Research Center, and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 100 Main St Cambridge MA 02142.
rudin@mit.edu

Gah-Yi Vahn
Management Science & Operations, London Business School, Regent's Park, London, NW1 4SA, United Kingdom.
gvahn@london.edu

We investigate the newsvendor problem when one has n observations of p features related to the demand as well as historical demand data. Both low-dimensional ($p/n = o(1)$) and high-dimensional ($p/n = O(1)$) data are considered. We propose two approaches to finding the optimal order quantity in this new setting — that of Machine Learning (ML) and Kernel Optimization (KO). We show how the feature-based model and solution approaches can be extended naturally to other realistic, "Big Data" situations, such as when one has data on prices, sales, competition, bidding and marketing; when data is censored; when ordering for multiple, similar items or when ordering for a new product with limited data. We show that both solution approaches yield decisions that are algorithmically stable, and derive tight bounds on their performance. We apply the feature-based algorithms for nurse staffing problem in a hospital emergency room and find that (i) the best KO and ML algorithms beat the best practice benchmark by 23% and 24% respectively in out-of-sample cost with statistical significance at the 5% level, and (ii) the best KO algorithm is faster than the best ML algorithm by three orders of magnitude and the best practice benchmark by two orders of magnitude.

Key words: big data, newsvendor, machine learning, Sample Average Approximation, statistical learning theory, quantile regression

History: February 1, 2015

1. Introduction

The classical newsvendor problem assumes that the probability distribution of the demand is fully known. It is clear, however, that one almost never knows the true distribution of the demand. In reality, one would instead have past demand data, as well as information on other factors that are associated with the demand. In this paper, we investigate the newsvendor problem when one has access to past demand observations as well as a potentially large number of *features* about the demand. By features we mean exogenous variables (factors) that are predictors of the demand and are available to the decision maker before the ordering occurs. Examples of relevant features are: the weather forecast, features related to seasonality (e.g. day of the week, month of the year and

season), various economic indicators (e.g. the interest rate and the consumer price index), as well as past demand itself. With plummeting costs of data storage and processing, many organizations are systematically collecting or purchasing such information, and thus our investigation brings the classical newsvendor problem in line with current "Big Data" trends in industry.

Formally, we assume that an unknown joint probability distribution exists between the demand and the p features used to predict the demand, and that we have a sample of size n drawn from this distribution. We consider both low dimensional data, i.e. the feature-to-observation ratio is small ($p/n = o(1)$) and high dimensional data, i.e. the feature-to-observation ratio is large ($p/n = O(1)$). In this paper, we illustrate how the decision maker can choose an order quantity with guaranteed out-of-sample performance, given a new decision period and a new set of features by learning from past data, in both the low and high dimensional regimes.

There have been many efforts to relax the assumption that the demand distribution is known. One main perspective has been the nonparametric ("data-driven") approach, whereby instead of the full knowledge of the demand distribution, the decision maker has access to independent and identically distributed (iid) demand data to estimate the expected newsvendor cost. Levi et al. (2007) considered the Sample Average Approximation (SAA) approach to the newsvendor problem as well as its multiperiod extension. There they derived a sample size bound; that is, a calculation of the minimal number of observations required in order for the SAA solution to be near-optimal with high probability. Whereas Levi et al. (2007) provided bounds for when only past demands are available, we provide bounds for the cases where past demand data and feature data are available.

Other perspectives on the data-driven newsvendor include those of Liyanage and Shanthikumar (2005), who proposed ordering according to a statistic of past demand data whose form is cleverly chosen based on a priori assumptions on the class of distributions the demand belongs to, Huh et al. (2011) and Besbes and Muharremoglu (2013) who provided theoretical insights into the newsvendor problem with iid censored demand data, and Levi et al. (2012), who improved upon the bound of Levi et al. (2007) by incorporating more information about the (featureless) demand distribution, namely through the weighted mean spread.

Alternatively, Scarf et al. (1958) and Gallego and Moon (1993) considered a minimax approach, whereby the decision maker maximizes the worst-case profit over a set of (one dimensional) distributions with the same mean and standard deviation. Perakis and Roels (2008) considered a minimax regret approach for the newsvendor with partial information about the (featureless) demand distribution.

None of the above mentioned works, however, consider the presence of feature data. As far as we are aware, this is the first paper to derive insights about the data-driven newsvendor problem when feature information is available.

The closest work to ours is possibly that of He et al. (2012), an empirical paper that modelled booking a hospital operating room as a newsvendor problem with just two features (number and type of cases). The algorithms, theoretical insights, and experimental investigation that we propose, using a large number of features, are directly applicable to the problem studied by He et al. (2012).

Summary of Contributions

In Sec. 2, we provide a new model for the newsvendor problem when the decision-maker has access to past feature information as well as the demand. We propose two approaches to solve the problem, one based on the machine learning (ML) principle of empirical risk minimization [for an in-depth discussion of this principle, see Vapnik (1998)], and the other based on minimizing the kernel estimate of the objective function (hereafter, KO for kernel optimization). Under the ML approach, the optimal order quantity can be learned via a linear programming (LP) algorithm in the case of low dimensional data (when $p/n = o(1)$) and a regularization-based algorithm in the case of high dimensional data (when $p/n = O(1)$). We also provide a sorting-based algorithm to find the optimal order quantity under the KO approach.

In Sec. 3, we formulate several new models that describe other realistic, Big Data situations based on the newsvendor model introduced in Sec. 2. First of all, we consider having data on product prices, sales, competition, bundling and marketing, and argue that they can simply be considered as features. Hence the algorithms of Sec. 2 do not need to be modified to incorporate such information. Next, we show how to modify the original model when the demand data are censored due to a constraint on the maximal order quantity. We further show that extensions of the original model can handle ordering for multiple related items, and "cold start" ordering when there is a new product on the market with limited demand information.

In Sec. 4, we show that the algorithms described in Sec. 2 possess a strong stability property known as *uniform stability*, and use these results to derive bounds on the performance of the in-sample decision. In particular, we derive tight generalization bounds (probabilistic bounds that quantify how the in-sample cost of the in-sample decision deviates from the true expected cost of the in-sample decision) as well as bounds on how the in-sample cost of the in-sample decision deviates from the true expected cost of the true optimal decision. Our bounds do not make any assumption about the feature-demand relationship, or the distribution of the demand beyond the existence of finite mean. The only assumption is that the feature-demand pairs are drawn independently from an (unknown) distribution. The bounds show how the out-of-sample cost (the "generalization error") of the in-sample decision deviates from its in-sample cost by a complexity term that scales gracefully as $1/\sqrt{n}$ and as $\sqrt{\ln(1/\delta)}$, where $1-\delta$ is the probabilistic accuracy of our bound, and how the finite-sample bias from the true optimal decision scales as $\sqrt{\log n}/n^{s/(2s+p)}$, where s is a smoothness

parameter, which is optimal in the asymptotic minimax sense. From a practical perspective, our bounds explicitly show the tradeoffs between the generalization error ("variance") and bias, and how they depend on the respective control of the decision model (the features and/or the regularization parameter in the case of the ML approach and the kernel function and its bandwidth in the case of the KO approach.). From a practical perspective, our bounds show how prediction quality depends on the parameters of the algorithms, where these parameters control the complexity of the model space.

Finally, in Sec. 5, we evaluate our algorithms against other known benchmarks through an extensive empirical investigation. Specifically, we apply our algorithms and other methods to a nurse staffing problem in a hospital emergency room. The best result using the ML approach was with ℓ_1 regularization, with a cost improvements of 23% [a saving £44,219 per annum (p.a.)] relative to the best practice benchmark (Sample Average Approximation (SAA) clustered by day of the week), and the best result using the KO approach had a cost improvement of 24% (a saving of £46,555 p.a.) relative to the same benchmark. Both results were statistically significant at the 5% level. The best KO method was also very computationally efficient, taking just 0.05 seconds to compute the optimal staffing level for the next period, which is three orders of magnitude faster than the best ML method and two orders of magnitude faster than the best practice benchmark and other benchmarks.

Before proceeding, we also mention that the vanilla feature-based newsvendor algorithms are equivalent to non-parametric and kernel quantile regressions, just as the featureless newsvendor algorithm performs quantile estimation. The results in this paper thus add to the literature on quantile regression [see Koenker (2005) for a general reference on quantile regression, Takeuchi et al. (2006) and Steinwart and Christmann (2011) for current results on non-parametric quantile regression].

2. Solving the Newsvendor with Feature Data

2.1. The Newsvendor Problem

A company sells perishable goods and needs to make an order before observing the uncertain demand. For repetitive sales, a sensible goal is to order a quantity that minimizes the total expected cost according to:

$$\min_{q\geq 0} \; EC(q) := \mathbb{E}[C(q;D)], \tag{1}$$

where q is the order quantity, $D \in \mathcal{D}$ is the uncertain (random) future demand,

$$C(q;D) := b(D-q)^+ + h(q-D)^+ \tag{2}$$

is the random cost of order q and demand D, and b and h are respectively the unit backordering and holding costs. If the demand distribution, F, is known, one can show the optimal decision is given by the $b/(b+h)$ quantile, that is:

$$q^* = \inf\left\{y:\ F(y) \geq \frac{b}{b+h}\right\}. \tag{3}$$

2.2. The Data-Driven Newsvendor Problem

In practice, the decision maker does not know the true distribution. Again assume that no external covariates are available to predict the demand. If one has access to historical demand observations $\mathbf{d}(n) = [d_1, \ldots, d_n]$, then the sensible approach is to substitute the true expectation with a sample average expectation and solve the resulting problem:

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

where we use the $\hat{}$ notation to emphasize quantities estimated from data. This approach is called the Sample Average Approximation (SAA) approach in stochastic optimization [for an excellent general reference, see Shapiro et al. (2009)]. One can show the optimal SAA decision is given by

$$\hat{q}_n = \inf\left\{y:\ \hat{F}_n(y) \geq \frac{b}{b+h}\right\}, \tag{4}$$

where $\hat{F}_n(\cdot)$ is the empirical cdf of the demand from the n observations. Note that if F is continuous, and we let $r = b/(b+h)$, then $\hat{q}_n = d_{\lceil nr \rceil}$, the $\lceil nr \rceil$-th largest demand observation.

2.3. The Feature-Based Newsvendor Problem

In reality, the demand depends on many observable *features* (equivalently, independent/explanatory variables, attributes or characteristics), such as seasonality (day, month, season), weather, location and economic indicators, which are available prior to making the order. In other words, the real newsvendor problem is

$$\min_{q(\cdot) \in \mathcal{Q}, \{q: \mathcal{X} \to \mathbb{R}\}}\ \mathbb{E}[C(q(\mathbf{x}); D(\mathbf{x}))|\mathbf{x}], \tag{5}$$

where the decision is now a function that maps the feature space $\mathcal{X} \subset \mathbb{R}^p$ to the reals and the expected cost that we minimize is now conditional on the feature vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$.

The decision-maker intent on finding an optimal order quantity in this new setting has three issues to address. The first issue is in knowing what features the demand depends on, which prescribes what data to collect. As this is application-specific, we assume that the decision maker has already collected appropriate historical data $S_n = [(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_n, d_n)]$. The data may be low-dimensional, where the number of features p is negligible compared to the number of observations n, i.e. $p/n =$

$o(1)$, or high dimensional, where the number of features is of comparable order to the number of observations, i.e. $p/n = O(1)$. The second issue is how to solve the problem (5) in an efficient manner given the data set. We address this issue in this section by proposing two alternative approaches of solving (5) — a machine learning (ML) approach and a kernel optimization (KO) approach. Both approaches are direct, in that the decision-maker solves for the (in-sample) optimal order quantity in a single step. As such, our proposed algorithms are customized for the feature-based newsvendor problem, and are distinct from other approaches known in the literature and in practice, such as the SAA (which does not assume feature information) and the separated estimation and optimization (SEO) approach. In Appendix B, we demonstrate the existence of simple, realistic scenarios where SAA and SEO are provably suboptimal, and in Sec. 5, we carry out an extensive empirical comparison with our proposed approaches. The final concern is what performance guarantee is possible prior to observing the demand in the next period. We address this in Sec. 4.

Before we begin, we reiterate that the setting of interest is one in which the decision-maker observes the features $\mathbf{x}_{n+1}$ before making the next ordering decision.

2.3.1. Machine Learning Methods The machine learning approach to solving the newsvendor problem with feature data is:

$$\min_{q(\cdot)\in\mathcal{Q},\{q:\mathcal{X}\to\mathbb{R}\}} \hat{R}(q(\cdot);S_n) = \frac{1}{n}\sum_{i=1}^{n}[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+], \qquad \text{(NV-ML)}$$

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

To solve (NV-ML), one needs to specify the function class $\mathcal{Q}$. The size or the complexity of $\mathcal{Q}$ controls overfitting or underfitting: for instance, if $\mathcal{Q}$ is too large, it will contain functions that fit the noise in the data, leading to overfitting. Let us consider linear decision rules of the form

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

where $x^1 = 1$, to allow for a feature-independent term (an intercept term). This is not restrictive, as one can easily accommodate nonlinear dependencies by considering nonlinear transformations of basic features. We might, for instance, consider polynomial transformations of the basic features, e.g., $[x_1, \ldots, x_p, x_1^2, \ldots, x_p^2, x_1x_2, \ldots, x_{p-1}x_p]$. Such transformations can be motivated from generative models of the demand (but do not need to be); for instance, assume: $D = f(\mathbf{x}) + \varepsilon$, where $\mathbf{x}$ is a p-dimensional vector of features. If we also assume that $f(\cdot)$ is analytic, we can express the demand function by its Taylor expansion:

$$\begin{aligned} D &\approx f(\mathbf{0}) + \partial f(\mathbf{0})'\mathbf{x} + \mathbf{x}'[D^2 f(\mathbf{0})]\mathbf{x} + \ldots + \varepsilon \\ &= f(\mathbf{0}) + \sum_{i=1}^{p} \partial f_i(\mathbf{0}) x_i + \sum_{i=1}^{p}\sum_{j=1}^{p} [D^2 f(\mathbf{0})]_{ij} x_i x_j + \ldots + \varepsilon, \end{aligned}$$

which means that the demand function of a basic feature vector $\mathbf{x}$ can be approximated by a linear demand model with a much larger feature space. For example, the second-order Taylor approximation of the demand model can be considered to be a linear demand model with the $(p+p^2)$ features mentioned earlier: $[x_1,\ldots,x_p,x_1^2,\ldots,x_p^2,x_1x_2,\ldots,x_{p-1}x_p]$. Regardless of the motivation for the transformations of basic features, we can choose them to be arbitrarily complex; hence our choice of decision functions that depend linearly on the feature vector is not at all restrictive. The choice of $\mathcal{Q}$ can be made more or less complex depending on which transformations are included.

We can thus solve (NV-ML) via the following linear program:

Machine Learning Algorithm 1

$$\begin{aligned}\min_{q:q(\mathbf{x})=\sum_{j=1}^p q^j x^j} \quad & \hat{R}(q(\cdot);S_n) = \frac{1}{n}\sum_{i=1}^n [b(d_i-q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i)-d_i)^+] \\ \equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad & \frac{1}{n}\sum_{i=1}^n (bu_i+ho_i) \\ s.t.\ \forall\, i=1,\ldots,n: & \\ & u_i \geq d_i - q^1 - \sum_{j=2}^p q^j x_i^j \\ & o_i \geq q^1 + \sum_{j=2}^p q^j x_i^j - d_i \\ & u_i, o_i \geq 0, \end{aligned} \tag{NV-ML1}$$

where the dummy variables u_i and o_i represent, respectively, underage and overage costs in period i. This is an LP with a $p+2n$-dimensional decision vector and $4n$ constraints. We will see in Sec. 4 that while (NV-ML1) yields decisions that are *algorithmically stable*, the performance guarantee relative to the true optimal decision is loose for high dimensional data, i.e. when p/n is large. Thus, in the case of high dimensional data, one can solve the LP (NV-ML1) by selecting a subset of the most relevant features according to some criterion, for example via cross validation or via model selection criteria such as the Akaike Information Criterion [Akaike (1974)] or Bayesian Information Criteria [Schwarz (1978)]. Alternatively, one can automate feature selection by solving the following *regularized* version of (NV-ML1).

Machine Learning Algorithm 2 (with regularization)

$$\begin{aligned}\min_{q:q(\mathbf{x})=\sum_{j=1}^p q^j x^j} \quad & \hat{R}(q(\cdot);S_n) + \lambda\|q\|_2^2 = \frac{1}{n}\sum_{i=1}^n [b(d_i-q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i)-d_i)^+] + \lambda\|\mathbf{q}\|_k^2 \\ \equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad & \frac{1}{n}\sum_{i=1}^n (bu_i+ho_i) \\ s.t.\ \forall\, i=1,\ldots,n: & \end{aligned}$$

$$
\begin{aligned}
u_i &\geq d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j \\
o_i &\geq q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i \\
u_i, o_i &\geq 0,
\end{aligned} \tag{NV-ML2}
$$

where $\lambda > 0$ is the regularization parameter and $\|\mathbf{q}\|_k$ denotes the ℓ_k-norm of the vector $\mathbf{q} = [q^1, \ldots, q^p]$. If we regularize by the ℓ_2 norm, the problem becomes a quadratic program (QP) and can be solved efficiently using widely available conic programming solvers. If we believe that the number of features involved in predicting the demand is very small, we can choose to regularize by the ℓ_0 semi-norm or the ℓ_1 norm to encourage sparsity in the coefficient vector. The resulting problem then becomes, respectively, a mixed-integer program (MIP) or an LP.

Let us consider variations. We may want a set of coefficients to be either all present or all absent, for instance if they fall into the same category (e.g., all are weather-related features). We can accommodate this with a regularization term $\sum_{g=1}^{G} \|q_{\mathcal{I}_g}\|_2$, with $\mathcal{I}_g$ being the indicator of group g. This regularization term is an intermediate between ℓ_1 and ℓ_2 regularization, where sparsity at the group level is encouraged by the sum (ℓ_1 norm) over groups. We will see in Sec. 4 that regularization leads to stable decisions with good finite-sample performance guarantees.

2.3.2. Kernel Optimization (KO) Method Here we introduce an alternative approach that can take feature information into account. We call this approach the Kernel Optimization (KO) method because it is based on Nadaraya-Watson kernel regression [Nadaraya (1964), Watson (1964)].

One of the goals of nonparametric regression is to estimate the expectation of a dependent variable (e.g. demand) conditional on independent variables taking on a particular value. That is, given past data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ one wants to estimate

$$m(\mathbf{x}_{n+1}) = \mathbb{E}[Y|\mathbf{x}_{n+1}],$$

where $Y \in \mathbb{R}$ is the dependent variable and $\mathbf{x}_{n+1} \in \mathbb{R}^p$ is a vector of new independent variables. In 1964, Nadaraya and Watson proposed to estimate this quantity by the locally weighted average

$$m_h(\mathbf{x}_{n+1}) = \frac{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1} - \mathbf{x}_i) y_i}{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1} - \mathbf{x}_i)},$$

where $K_h(\cdot)$ is a kernel function with bandwidth h. In this paper, we consider the Gaussian kernel

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

Now for an order quantity q, the feature-dependent newsvendor expected cost after observing features $\mathbf{x}_{n+1}$ is given by

$$\mathbb{E}[C(q; D)|\mathbf{x}_{n+1}], \tag{6}$$

which depends (implicitly) on the demand distribution at $\mathbf{x}_{n+1}$. Thus if we consider the newsvendor cost to be the dependent variable, we can estimate (6) by the Nadaraya-Watson estimator

$$\frac{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)C(q,d_i)}{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)}.$$

This gives rise to a new approach to feature-data-driven newsvendor, which we call the Kernel Optimization (KO) Method.

$$\min_{q\geq 0} \ \tilde{R}(q;S_n,\mathbf{x}_{n+1}) = \min_{q\geq 0} \ \frac{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)C(q,d_i)}{\sum_{i=1}^{n} K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)}. \tag{NV-KO}$$

Note that there are no edge effects in the objective estimate if the kernel is smooth, which is the case for the Gaussian kernel. Notice that the optimization is over the non-negative reals, and the optimal decision implicitly depends on $\mathbf{x}_{n+1}$. (NV-KO) is a one-dimensional piecewise linear optimization problem, and we can find its solution according to the following proposition.

PROPOSITION 1. *The optimal feature-based newsvendor decision $\hat{q}_n^\kappa$ obtained by solving (NV-KO) is given by*

$$\hat{q}_n^\kappa = \hat{q}_n^\kappa(\mathbf{x}_{n+1}) = \inf\left\{q: \ \frac{\sum_{i=1}^{n}\kappa_i \mathbb{I}(q\leq d_i)}{\sum_{i=1}^{n}\kappa_i} \geq \frac{b}{b+h}\right\}, \tag{7}$$

where for simplicity we introduce $\kappa_i = K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)$. In other words, we can find $\hat{q}_n^\kappa$ by plugging-in the past demand in increasing order, and choosing the smallest value at which the inequality in (7) is satisfied.

Notice that the left hand side (lhs) of the inequality in (7) is similar to the empirical cdf of the demand, except that each past demand observation d_i is re-weighted by the distance of its corresponding feature $\mathbf{x}_i$ to the current feature $\mathbf{x}_{n+1}$.

3. New Operational Models with Big Data Newsvendor

In this section, we develop new operational models for situations that are based on the big data newsvendor.

3.1. Pricing, Sales, Competition, Bundling and Marketing

The major benefit of employing a feature-based approach is that anything that affects the demand can be included as a feature. The price of the product, along with nonlinear transformations of it, can be used as features in the model. To encode whether the item is on sale (of a certain type - say 10% off) we can use an indicator variable (1 if sale, 0 otherwise). The prices of competitors could also be included directly as features. Features can be created to encode discounts offered for bundling the item with other items. Further, the amount and type of marketing of the item can be included as features. This flexibility to naturally model scenarios that have not arisen in the past is the core of the feature-based newsvendor problem investigated in this paper.

3.2. Censored Data: Limited Ordering Capacity

We would have censored demand data if the ordering capacity is limited. In other words, if the i-th historical demand is equal to the maximum capacity q_{max}, then we would only know that the actual demand was greater than or equal to q_{max}. For demands that hit this limit, i.e. $d_i \geq q_{max}$, we can change the objective function to penalize our in-sample estimate $q(\mathbf{x}_i)$ if it is less than $q_{\max}$. Our objective then becomes

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \quad \sum_{i:\text{demand} < \text{capacity}} [b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+] + \sum_{i:\text{demand} = \text{capacity}} [b(q_{\max} - q(\mathbf{x}_i))^+] + (\lambda \|\mathbf{q}\|_k).$$

if adopting the ML method and

$$\min_{q \geq 0} \quad \frac{\sum_{i:\text{demand} < \text{capacity}} \kappa_i [b(d_i - q)^+ + h(q - d_i)^+]}{\sum_i \kappa_i} + \frac{\sum_{i:\text{demand} = \text{capacity}} \kappa_i [b(q_{\max} - q)^+]}{\sum_i \kappa_i}.$$

if adopting the KO method.

3.3. Similar Influences for Multiple Items

We consider the situation where we have multiple items, and we believe that some of the features play a similar role in predicting the demand for all of these items. In that case, we can create a joint objective for both items, and regularize the decisions to be close together. An example for two items (denoted by $^{(1)}$ and $^{(2)}$) is shown below:

$$\min_{\mathbf{q}^{(1)},\mathbf{q}^{(2)}} \quad \sum_{i=1}^{n} [b^{(1)}(d_i^{(1)} - q^{(1)}(\mathbf{x}_i))^+ + h^{(1)}(q^{(1)}(\mathbf{x}_i) - d_i^{(1)})^+] + \sum_{i'=1}^{n'} [b^{(2)}(d_{i'}^{(2)} - q^{(2)}(\mathbf{x}_{i'}))^+ + h^{(2)}(q^{(2)}(\mathbf{x}_{i'}) - d_{i'}^{(2)})^+] + \lambda \|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\|_2,$$

if adopting the ML method, and

$$\min_{q^{(1)} \geq 0, q^{(2)} \geq 0} \quad \sum_{i=1}^{n} \kappa_i [b^{(1)}(d_i^{(1)} - q^{(1)})^+ + h^{(1)}(q^{(1)} - d_i^{(1)})^+] / \sum_{i=1}^{n} \kappa_i + \sum_{i'=1}^{n'} \kappa_{i'} [b^{(2)}(d_{i'}^{(2)} - q^{(2)})^+ + h^{(2)}(q^{(2)} - d_{i'}^{(2)})^+] / \sum_{i'=1}^{n} \kappa_{i'} + \lambda |q^{(1)} - q^{(2)}|,$$

if adopting the KO method.

3.4. The 'Cold Start' Problem with New Items

When a new item becomes available, we may not have sufficient historical demand data to draw inferences about its future demand. In our framework, we can accommodate this by using data about related products to inform our predictions. We do this by training our model on a combination of historical data from the new item and from the existing items. This way, the data from the existing items can act as a form of regularization. In a related work, Chang et al. (2012) use data from other items for predicting quality rankings for product categories that contain few products.

If we regularize using one additional product, the objective becomes:

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \quad \frac{1}{n}\sum_{i=1}^{n}[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+] \\ +\lambda_{\text{existing}}\frac{1}{n_{\text{existing}}}\sum_{i'=1}^{\text{existing}}[b(d_{i'} - q(\mathbf{x}_{i'}))^+ + h(q(\mathbf{x}_{i'}) - d_{i'})^+] + \lambda\|\mathbf{q}\|_2$$

if adopting the ML method, and

$$\min_{q\geq 0} \quad \sum_{i=1}^{n}\kappa_i[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+]/\sum_{i=1}^{n}\kappa_i \\ +\lambda_{\text{existing}}\sum_{i'=1}^{n_{\text{existing}}}\kappa_{i'}[b(d_{i'} - q(\mathbf{x}_{i'}))^+ + h(q(\mathbf{x}_{i'}) - d_{i'})^+]/\sum_{i'=1}^{n_{\text{existing}}}\kappa_{i'}$$

if adopting the KO method.

Here we would use b and h for the new item rather than the existing items, even if the backorder and holding costs for the existing items were different. This is because we want to estimate the order quantity for the new items and not the existing items. We choose $\lambda_{\text{existing}}$ based on our belief of the similarity of the new product to the existing product. As n increases, $\lambda_{\text{existing}}$ should decrease so that the influence of the existing product fades. For instance, if desired, $\lambda_{\text{existing}}$ could be set to $\alpha n_{\text{existing}}/n$ where $\alpha < 1$ so that each observation from an existing item is worth fraction α of an observation from a new item.

4. Bounds on the Out-of-Sample Cost

In this section, we provide theoretical guarantees on the out-of-sample cost of the ordering decisions chosen by (NV-ML1), (NV-ML2) and (NV-KO). We start by first showing that the optimal (in-sample) decisions obtained by solving the three optimization problems are *algorithmically stable*, which means that the true expected cost of the decision is not sensitive to changes in the data set. The stability of a decision is thus a desirable property and is intimately linked to how well the decision performs on new observations. We then quantify the performance of the three decisions on

a new observation in the form of *generalization bounds*, and extend them to quantify how well the in-sample decisions perform relative to the true optimal decision in terms of the expected out-of-sample cost.

Generalization bounds are probabilistic bounds on the out-of-sample performance of in-sample predictions, and are useful in highlighting quantities that are important in prediction. There are several types of generalization bounds that are known in the statistical learning theory literature. The type of bounds we employ are *stability* bounds, because as algorithm-specific bounds, they generate the most insightful results for the feature-based newsvendor problem. In contrast, *uniform* generalization bounds, which are more common in statistical learning theory and perhaps more familiar to the reader, are algorithm-independent. Uniform generalization bounds do not consider the way in which the algorithm searches the space of possible models, and as such they lack specific insights about prediction.

Stability bounds can be thought of as a form of Hoeffding's inequality for algorithms. (Recall, for instance, that the sample-size bounds of Levi et al. (2007) and Levi et al. (2012) are critically based on Hoeffding's inequality.) To apply Hoeffding's inequality, we need to know a bound on the values of a random variable, whereas to apply stability bounds, we need to know a bound on the stability of an algorithm to a random data set. The challenging part is to prove a stability result that is as strong as possible for the algorithm, which we here achieve for the newsvendor problem. Stability bounds have origins in the 1970's [Rogers and Wagner (1978), Devroye and Wagner (1979a,b)], and recent work includes that of Bousquet and Elisseeff (2002).

To show that our newsvendor algorithms are stable, we first show that the newsvendor cost on the training set does not change very much when one of the training examples changes. Stability of a decision is a desirable property, and these bounds show that stability is intimately linked to how well the decision performs on new observations. One of the main contributions of this section is thus in showing that (NV-ML1), (NV-ML2) and (NV-KO) are strongly stable, which are important results in their own right.

We start by defining what it means for the algorithm to generalize, and then define stability.

The *true risk* is the expected out-of-sample cost, where the expectation is taken over an unknown distribution over $\mathcal{X} \times \mathcal{D}$, where $\mathcal{X} \subset \mathbb{R}^p$. Specifically,

$$R_{true}(q) := \mathbb{E}_{D(\mathbf{x})}[C(q; D(\mathbf{x}))].$$

No assumptions are made about the form of this distribution. We are interested in minimizing this cost, but we cannot measure it as the distribution is unknown. The empirical risk is the average cost over the training sample:

$$\hat{R}(q; S_n) := \frac{1}{n} \sum_{i=1}^{n} C(q, d_i(\mathbf{x}_i)).$$

The empirical risk can be calculated, whereas the true risk cannot; the empirical risk alone, however, is an incomplete picture of the true risk. We must have some additional property of the algorithm to ensure that the method does not overfit. If the algorithm is stable, it is less likely to overfit, which we quantify in the results in this section. Specifically, we provide probabilistic upper bounds on the true risk in terms of the empirical risk and the algorithmic stability of the method. Since we desire the true risk to be low, a combination of low empirical risk and sufficient stability ensures this.

The training set is, as before, $S_n = \{z_1 = (\mathbf{x}_1, d_1), \ldots, z_n = (\mathbf{x}_n, d_n)\}$, $z \in \mathcal{Z}$, and we also define the modified training set

$$S_n^{\setminus i} := \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\},$$

which leaves one observation out.

A *learning algorithm* is a function A from $\mathcal{Z}^n$ into $\mathcal{Q} \subset \mathcal{D}^{\mathcal{X}}$, where $\mathcal{D}^{\mathcal{X}}$ denotes the set of all functions that map from $\mathcal{X}$ to $\mathcal{D}$. A learning algorithm A maps the training set S_n onto a function $A_{S_n} : \mathcal{X} \to \mathcal{D}$. A learning algorithm A is *symmetric with respect to* S_n if for all permutations $\pi : S_n \to S_n$ of the set S_n,

$$A_{S_n} = A_{\pi(S_n)} = A_{\{\pi(z_1),\ldots,\pi(z_n)\}}.$$

In other words, a symmetric learning algorithm does not depend on the order of the elements in the training set S_n. The *loss* of the decision rule $q \in \mathcal{Q}$ with respect to a sample $z = (\mathbf{x}, d)$ is defined as

$$\ell(q, z) := c(q, d(\mathbf{x})),$$

for some cost function c, which in our work is the newsvendor cost C. In what follows, we assume that all functions are measurable and all sets are countable. We also assume $\mathcal{Q}$ is a convex subset of a linear space.

Our algorithms for the learning newsvendor problem turns out to have a very strong stability property, namely it is *uniformly stable*. We define stability and uniform stability below.

Definition 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 $i \in \{1, \ldots, n\}$,

$$\|\ell(A_{S_n}, \cdot) - \ell(A_{S_n^{\setminus i}}, \cdot)\|_\infty \leq \alpha. \tag{8}$$

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

In what follows, the random demand is denoted by D, and is assumed to be bounded: $D \in \mathcal{D} := [0, \bar{D}]$, but we relax this assumption later so that the demand is bounded with high probability. The feature domain is also bounded; in particular, we assume all feature vectors live in a ball:

$\|\mathbf{x}\|_2^2 \leq X_{\max}^2$ for all $\mathbf{x}$. As before, the historical ('training') set of data is given by $S_n = \{(\mathbf{x}_i, d_i)\}_{i=1}^n$. We defer all proofs to Appendix A.

We start with the results for (NV-ML1).

PROPOSITION 2 **(Uniform stability of (NV-ML1))**. *The learning algorithm (NV-ML1) with iid data is symmetric and uniformly stable with respect to the newsvendor cost function $C(\cdot,\cdot)$ with stability parameter*

$$\alpha_n = \frac{\bar{D}(b \vee h)^2}{(b \wedge h)} \frac{p}{n}. \tag{9}$$

Here the notation $b \vee h$ indicates the maximum value of b and h, and $b \wedge h$ indicates the minimum of the two. This bound illuminates that if one of b or h is too large and the other too small, in other words if the backordering and holding costs are highly asymmetric, then the algorithm is very sensitive to small changes in the data set. The algorithm is more stable when the target order quantity is closer to the median of the demand distribution.

The stability result, coupled with a Hoeffding/McDiarmid based argument [in our case, we are using modernized versions of the 1970's results due to Bousquet and Elisseeff (2002)], yields the following:

THEOREM 1 **(Generalization Bound for (NV-ML1))**. *Let $\hat{q}$ be the model produced by Algorithm (NV-ML1). The following bound holds 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 $\mathcal{X} \times \mathcal{D}$:*

$$|R_{true}(\hat{q}) - \hat{R}(\hat{q}; S_n)| \leq \frac{2(b \vee h)^2 \bar{D}}{b \wedge h} \frac{p}{n} + \left(\frac{4(b \vee h)^2 \bar{D}}{b \wedge h} p + \bar{D} \right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{10}$$

For small p, Theorem 1 suggests that the generalization error of the newsvendor cost scales gracefully as $O(1/\sqrt{n})$. In addition, if $p = 0$, we retrieve the well-known bound of Hoeffding (1963). This bound also contains the intuition about quantities in the learning process that we had hoped for: that for high dimensional data (large p/n), one could solve (NV-ML1) by selecting a subset of the most relevant features, and that a highly asymmetric b and h could hurt generalization ability.

We can relax the assumption of a bounded demand by considering instead demand that is bounded above by $\bar{D}$ with probability at least $1-\gamma$, in which case the bounds hold now with probability $1-\delta-n\gamma$. Note that some value of the demand is, in some sense, necessary on the right hand side of the inequality so that the bound is not scale invariant. In other words, if the risks on the left side of the inequality were doubled, it would not make sense for the right side of the inequality to stay the same (rather it should also scale as the left side does).

For large p/n, we had suggested finding the optimal order quantity by solving (NV-ML2) instead. In this case, the generalization is driven by the regularization parameter, rather than the ratio p/n, as well as b and h as before. First is the stability result.

PROPOSITION 3 **(Uniform stability of (NV-ML2))**. *The learning algorithm (NV-ML2) is symmetric, and is uniformly stable with respect to the newsvendor cost function C with stability parameter*

$$\alpha_n^r = \frac{(b \vee h)^2}{2X_{\max}^{-2}} \frac{1}{n\lambda}. \tag{11}$$

The new stability parameter does not depend on p explicitly, but through $X_{\max}^2$. Thus the stability of the algorithm can be controlled via the regularization parameter λ, which must be chosen in relation to $X_{\max}^2$.

It is also interesting that this stability only depends on the maximum of b and h and not the minimum, which was the case for the unregularized problem. Intuitively, the regularization compensates for small values in b and h. If either b or h is very small, for small dimensions p, estimates for the optimal quantile can be very uncertain. The regularization helps to control that by encouraging quantiles to be estimated through coefficients that stay small. This can also be seen in the generalization bound.

THEOREM 2 **(Generalization Bound for (NV-ML2))**. *Let $\hat{q}$ be the model produced by Algorithm (NV-ML2) with ℓ_2 regularization. The following bound holds 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 $\mathcal{X} \times \mathcal{D}$:*

$$|R_{true}(\hat{q}) - \hat{R}(\hat{q}; S_n)| \leq \frac{(b \vee h)^2}{X_{\max}^{-2}} \frac{1}{n\lambda} + \left(\frac{2(b \vee h)^2}{X_{\max}^{-2}} \frac{1}{\lambda} + \bar{D} \right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{12}$$

The generalization bound of Theorem 2 makes the advantage of regularization clear for high-dimensional data. By employing the regularized algorithm one can obtain an order quantity that is not only stable but also *agnostic* to which feature is ultimately selected by the decision. In other words, the decision-maker can simply apply (NV-ML2) to the entire "Big Data" set, as long as λ is chosen appropriately in relation to $X_{\max}^2$ (in practice, one would search for the best λ via cross-validation; see Sec. 5). In contrast, to obtain a stable decision from (NV-ML1) one would need to pre-select the features.

The generalization error for Theorem 2 also scales appropriately in δ, as $O(\sqrt{\ln(1/\delta)})$.

We now build upon the generalization bounds of Theorem 1 and 2 to compare the optimal in-sample decisions of (NV-ML1) and (NV-ML2) to the true optimal decisions.

THEOREM 3 **(Comparison of (NV-ML1) with True Optimal)**. *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$, and assume it is continuously differentiable up to order k on some fixed open neighborhood of zero in $\mathbb{R}^p$ and its k-th derivative is uniformly Hölder continuous at zero with exponent γ (see Condition 1 in Appendix A for details). Then with probability at least $1-\delta$ over*

the random draw of the sample S_n, where each element of S_n is drawn i.i.d. from an unknown distribution on $\mathcal{X} \times \mathcal{D}$:

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

where $s = k + \gamma$ is the order of smoothness of q^* and $\hat{q}$ is the solution to (NV-ML1) and M is a constant that depends on the demand distribution.

The first term in the bound above is the generalization error. The second term is the finite-sample bias of the in-sample decision. The bound is tight in the sense that the dependence on n of the bias term is optimal in the asymptotic minimax sense. The bound above shows explicitly how the performance of the decision depends on the size of the data set used in the decision model.

THEOREM 4 **(Comparison of (NV-ML2) with True Optimal)**. *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$, and assume it is continuously differentiable up to order k on some fixed open neighborhood of zero in $\mathbb{R}^p$ and its k-th derivative is uniformly Hölder continuous at zero with exponent γ (see Condition 1 in Appendix A for details). Then with probability at least $1-\delta$ over the random draw of the sample S_n, where each element of S_n is drawn i.i.d. from an unknown distribution on $\mathcal{X} \times \mathcal{D}$:*

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

where $s = k + \gamma$ is the order of smoothness of q^*, $\hat{q}_\lambda$ is the solution to (NV-ML2) with ℓ_2 regularization, $\hat{q}$ is the solution to (NV-ML1) and M is a constant that depends on the demand distribution.

The first term in the bound above is the generalization error. The second and the third terms are, respectively, the bias of the in-sample decision due to regularization and due to having a finite number of observations. The bound shows explicitly how there is a trade-off between the generalization error and the regularization bias because while larger λ decreases the generalization error, it would increase the bias. Ultimately, however, regularization gives the decision-maker an extra degree of control while being agnostic to which feature is important a priori.

We now state stability, generalization bound and true risk bound results for (NV-KO).

PROPOSITION 4 **(Uniform stability of (NV-KO))**. *The algorithm (NV-KO) with iid data and the Gaussian kernel is symmetric with respect to the newsvendor cost function $C(\cdot,\cdot)$ with uniform stability parameter*

$$\alpha_\kappa = \frac{\bar{D}(b \vee h)^2}{(b \wedge h)}\frac{1}{1 + (n-1)r_h}, \tag{13}$$

where $r_h = \exp(-2X_{\max}^2/h^2)$.

The first term in the kernel stability parameter is the same as for (NV-ML1), hence the insight that highly asymmetric backordering and holding costs leads to unstable decisions still holds. The second term depends on n and r_h, which in turn depends on the bandwidth h and the number of features p through $X_{\max}$. r_h goes from zero to one as h increases from zero to infinity; this shows that greater stability is achieved with a larger bandwidth. This is an intuitive result, as a larger bandwidth is associated with bunching feature observations closer together. With this in mind we state the generalization bound for (NV-KO).

THEOREM 5 **(Generalization Bound for (NV-KO))**. *Let $\hat{q}^\kappa$ be the optimal decision of (NV-KO). The following bound holds 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 $\mathcal{X}\times\mathcal{D}$:*

$$|R_{true}(\hat{q}^\kappa) - \hat{R}(\hat{q}^\kappa; S_n)| \leq \frac{2(b\vee h)^2\bar{D}}{b\wedge h}\frac{1}{1+(n-1)r_h} + \left(\frac{4(b\vee h)^2}{b\wedge h}\frac{1}{1+(n-1)r_h}+1\right)\bar{D}\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{14}$$

As for (NV-ML1) and (NV-ML2), the generalization error scales gracefully as $O(1/\sqrt{(n)})$. The bound does depend on p implicitly through r_h, however it also shows that the error can be controlled by the bandwidth h. In this regard, the bandwidth parameter plays the same role as the regularization parameter in (NV-ML2).

Finally, we have the following result that parallels Theorems 3 and 4.

THEOREM 6 **(Comparison of (NV-KO) with True Optimal)**. *Denote the true optimal solution by $q^* = q^*(\mathbf{x}_{n+1})$, and assume it is continuously differentiable up to order k on some fixed open neighborhood of zero in $\mathbb{R}^p$ and its k-th derivative is uniformly Hölder continuous at zero with exponent γ (see Condition 1 in Appendix A for details). Then with probability at least $1-\delta$ over the random draw of the sample S_n, where each element of S_n is drawn i.i.d. from an unknown distribution on $\mathcal{X}\times\mathcal{D}$:*

$$\begin{aligned}|R_{true}(q^*) - \hat{R}(\hat{q}^\kappa; S_n)| \leq{}& \frac{2(b\vee h)^2\bar{D}}{b\wedge h}\frac{1}{n} + \left(\frac{4(b\vee h)^2}{b\wedge h}+1\right)\bar{D}\sqrt{\frac{\ln(2/\delta)}{2n}}\\ &+ (b\vee h)M\frac{\sqrt{\log n}}{n^{s/(2s+p)}},\end{aligned}$$

where $\hat{q}^\kappa$ is the solution to (NV-KO) for the uniform kernel with bandwidth $h = h_n = O(n^{1/(2s+p)}) \geq 2X_{\max}$, and M is a constant that depends on the demand distribution.

5. Case Study: Nurse Staffing in a Hospital Emergency Room

In this section, we compare the three algorithms introduced in Sec. 2, (NV-ML1), (NV-ML2) and (NV-KO) against the main data-driven benchmarks known in the literature and practice through an

PAGES REMOVED !

is the best in terms of the out-of-sample cost, whereas it takes 114 seconds for the the ML method with ℓ_1 regularization, which is the second-best performing method. The KO method is also faster than SAA-day, SEO methods and Scarf by two orders of magnitude.

5.3. Optimal Staffing Decisions

Let us further investigate the staffing decision of the best method: KO-OS with $h = 1.62$.

In Fig. 2 (a), we display the staffing levels predicted by KO-OS with $h = 1.62$ along with the actual required levels. For comparison, we also provide the staffing levels predicted by the second-best method, NVreg1-OS with $\lambda = 1 \times 10^{-7}$ in Fig. 2 (b). A striking observation is that both KO-OS and NVreg1-OS methods anticipate periods of high demand fairly well, as evidenced by the matching of the peaks in the predicted and actual staffing levels. The two methods are otherwise quite different in the prediction; in particular, the KO-OS method balances both over-staffing and under-staffing, whereas NVreg1-OS method seems to systematically over-predict the staffing level.

Let us now suppose the hospital indeed implements our algorithm for its nurse staffing decisions. We wish to gain some insight into the predictions made by the algorithm. In particular, we would like to know when the hospital is over- or under-staffed, assuming the hospital chooses to implement the best possible method, provided by KO-OS with $h = 1.62$. In Figs. 3 and 4, we show the conditional probability (frequency) of under- and over-staffing by day of the week and by time period. We derive the following insights from these plots, which could be useful for patients and managers directly: (i) mid-week days are more likely to be under-staffed then weekends, thus, given the choice to visit the emergency room on a weekday or weekend, we would choose a weekend, (ii) the period from noon to midnight is substantially more likely to be over-staffed then the period from midnight to noon, thus, given the choice of time to visit the emergency room, we would choose visiting in the afternoon, and (iii) the algorithm is most likely to over-staff by at least 50% of the required level on a Monday then any other day of the week, hence, given the flexibility, we would choose to visit the emergency room on a Monday.

6. Conclusion

This work shows how a newsvendor decision-maker who has access to past information about various features about the demand as well as demand itself can make a sensible ordering decision. We proposed tractable algorithms, two based on the empirical risk minimization principle from machine learning and the other based on minimizing a kernel estimate of the cost function, and derived theoretical error bounds on their performance. We further derived new models based on the "Big Data newsvendor" to capture other realistic settings. Finally, we investigated nurse staffing in a hospital emergency room and showed that our custom-designed, feature-based algorithms compute staffing decisions that yield substantially lower cost than the major benchmarks.

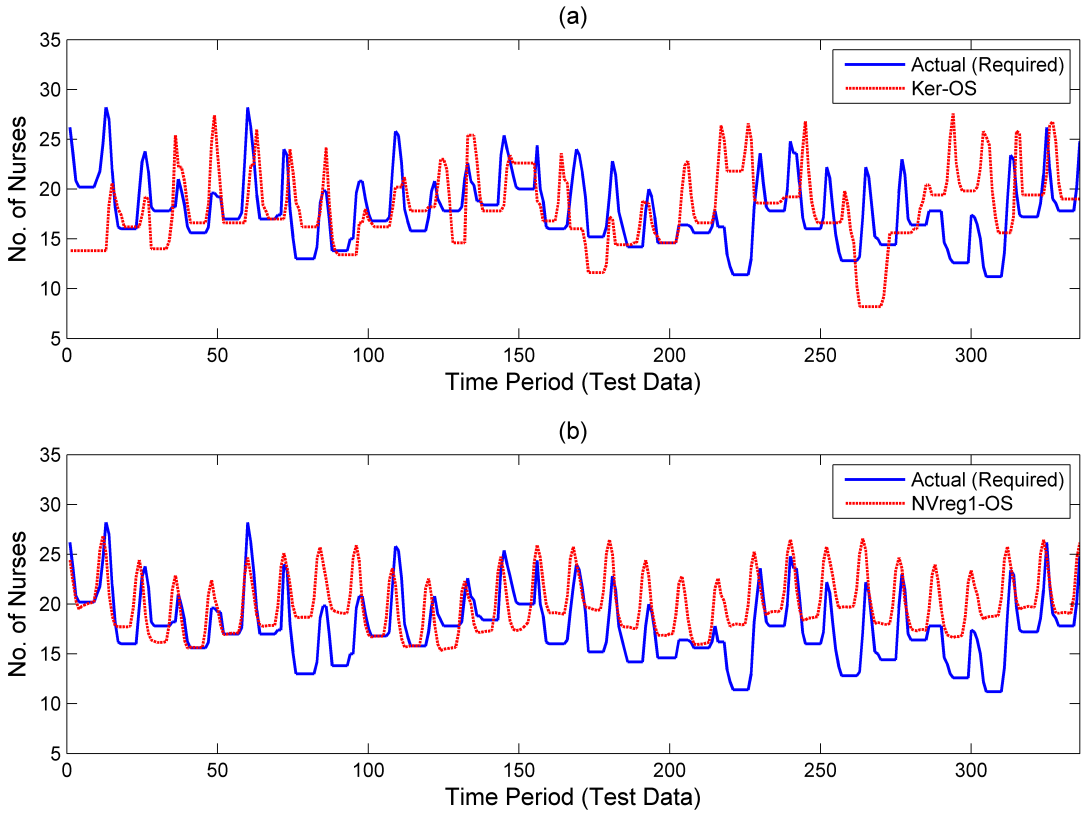

Figure 2 (a) A time-series plot of actual staffing demand (solid blue) versus staffing levels predicted by KO-OS with $h = 1.62$ (best method) on test data set in dotted red. (b) A time-series plot of actual staffing demand (solid blue) versus staffing levels predicted by NVreg1-OS with $\lambda = 1 \times 10^{-7}$ (second best method) on test data set in dotted red.

Appendix A: Proofs of Main Theorems in Sec. 4

We will use the following lemma in the proof of Propositions 2 and 4.

LEMMA 1 **(Exact Uniform Bound on the NV Cost)**. *The newsvendor cost function $C(\cdot,\cdot)$ is bounded by $(b \vee h)\bar{D}$, which is tight in the sense that:*

$$\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} |C(q, D(\mathbf{x}))| = \bar{D}(b \vee h).$$

Proof. (Of Lemma 1) Clearly, $\bar{D}(b \vee h)$ is an upper bound on $|C(q,d)|$ for all $q, d \in [0, \bar{D}]$. Now if $d = 0$ and $q = \bar{D}$, $|C(q,d)| = \bar{D}h$. Conversely, if $d = \bar{D}$ and $q = 0$, $|C(q,d)| = \bar{D}b$. Hence the upper bound is attained. □

Now for the proof of the Proposition 2.

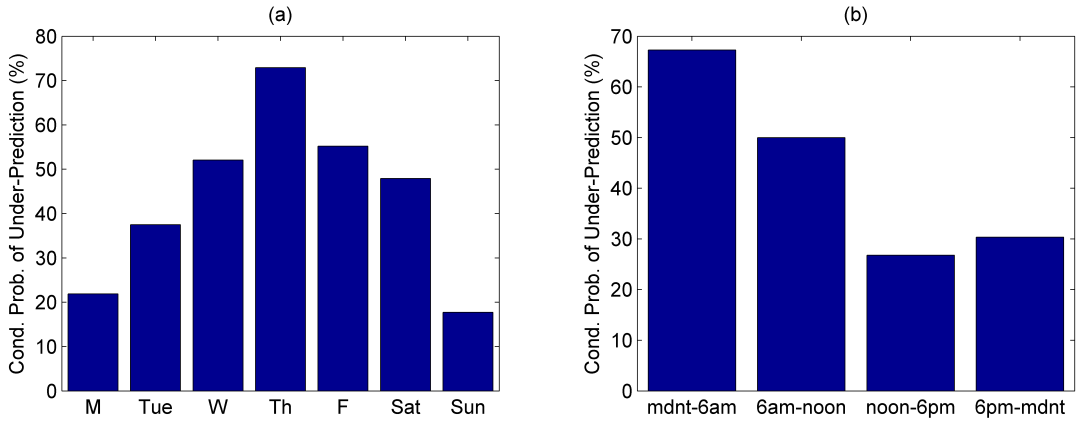

Figure 3 A plot of the conditional probabilities of under-staffing (a) by day and (b) by time period for KO-OS with $h = 1.62$ (best method). The conditioning is done by the particular day or the time period, i.e. the probability of under-staffing given it is a Monday.

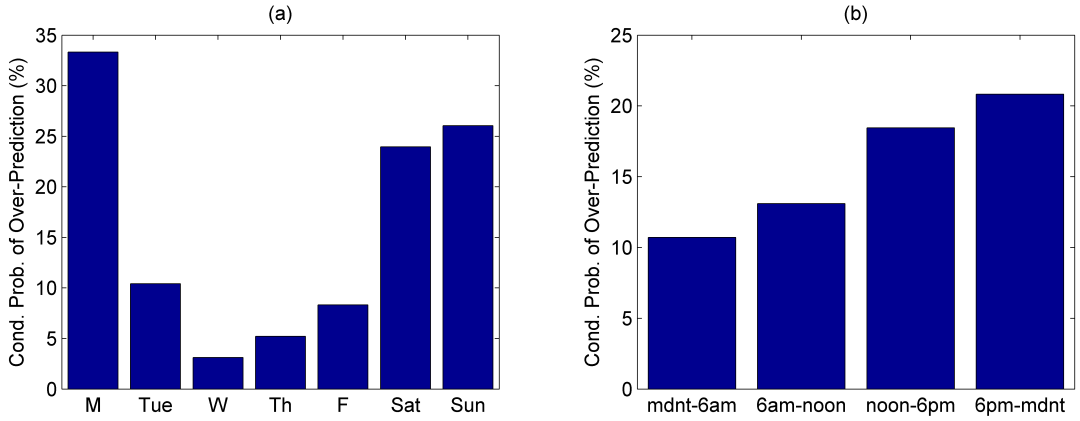

Figure 4 A plot of the conditional probabilities of over-staffing by at least 50% (a) by day and (b) by time period for KO-OS with $h = 1.62$ (best method). The conditioning is done by the particular day or the time period, i.e. the probability of over-staffing given it is a Monday.

Proof. (Of Proposition 2) Symmetry follows from the fact that the data-generating process is iid. For stability, we will change our notation slightly to make the dependence on n and S_n explicit. Let

$$q_n(\mathbf{x}) := \mathbf{q}_n^\top \mathbf{x} = \sum_{j=1}^{p} q_n^j x_j$$

and

$$q_{n\setminus i}(\mathbf{x}) := \mathbf{q}_{n\setminus i}^\top \mathbf{x} = \sum_{j=1}^{p} q_{n\setminus i}^j x_j$$

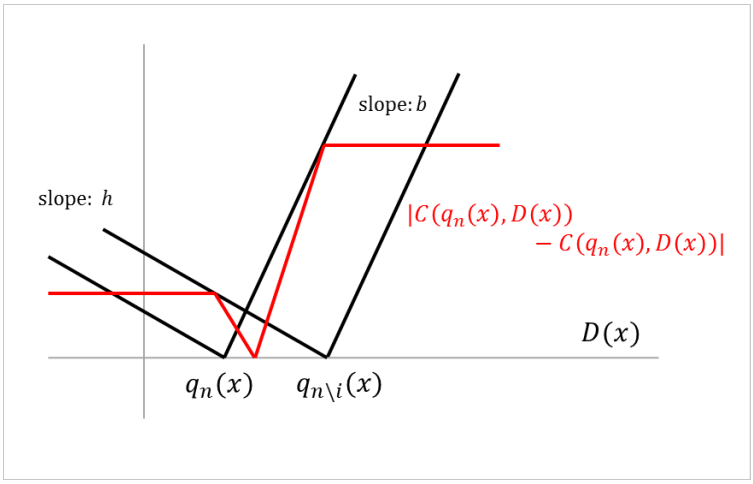

Figure 5 A plot illustrating that the difference $|C(q_n(\mathbf{x}), D(\mathbf{x})) - C(q_{n\setminus i}(\mathbf{x}), D(\mathbf{x}))|$ is bounded.

where

$$[q_n^1, \ldots, q_n^p] = \underset{\mathbf{q}=[q^1,\ldots,q^p]}{\arg\min} \hat{R}(\mathbf{q}; S_n) = \frac{1}{n}\sum_{j=1}^{n}\left[b\left(d_j - \sum_{j=1}^{p} q^j x_j\right)^+ + h\left(\sum_{j=1}^{p} q^j x_j - d_j\right)^+\right]$$

is the solution to (NV-ML1) for the set S_n, and

$$(q_{n\setminus i}^1, q_{n\setminus i}^1) = \underset{\mathbf{q}=[q^1,\ldots,q^p]}{\arg\min} \hat{R}(\mathbf{q}; S_n^{\setminus i}) = \frac{1}{n}\sum_{j=1}^{n}\left[b\left(d_j - \sum_{j=1}^{p} q^j x_j\right)^+ + h\left(\sum_{j=1}^{p} q^j x_j - d_j\right)^+\right]$$

is the solution to (NV-ML1) for the set $S_n^{\setminus i}$. Note that:

$$\hat{R}(\mathbf{q}; S_n) = \frac{n-1}{n}\hat{R}(\mathbf{q}; S_n^{\setminus i}) + \frac{1}{n}\tilde{R}(\mathbf{q}; S_i),$$

where $S_i = (\mathbf{x}_i, d_i)$.

For a fixed $\mathbf{x}$, we have, by the Lipschitz property of $C(q;\cdot)$,

$$\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} |C(q_n(\mathbf{x}), D(\mathbf{x})) - C(q_{n\setminus i}(\mathbf{x}), D(\mathbf{x}))| \le \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)|q_n(\mathbf{x}) - q_{n\setminus i}(\mathbf{x})|.$$

So we want to bound

$$\boxed{|q_n(\mathbf{x}) - q_{n\setminus i}(\mathbf{x})|} = \left|\sum_{j=1}^{p} q_n^j x_j - \sum_{j=1}^{p} q_{n\setminus i}^j x_j\right|.$$

By the convexity of the function $\hat{R}_n(\cdot, S)$, we have (see Section 23 of Rockafellar (1997)):

$$\sum_{j=1}^{p} \nu_j (q_{n\setminus i}^j - q_n^j) \le \hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)$$

for all $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_m] \in \partial \hat{R}(q_n; S_n)$ (set of subgradients of $\hat{R}(\cdot, S_n)$ at q_n). Furthermore, because $0 \in \partial \hat{R}(q_n; S_n)$ by the optimality of q_n, we have

$$0 \leq \max_{\boldsymbol{\nu} \in \partial \hat{R}(q_n; S_n)} \sum_{j=1}^{p} \nu_j (q^j_{n\setminus i} - q^j_n) \leq \hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)$$

where the max over $\boldsymbol{\nu}$ can be attained because $\partial \hat{R}(q_n; S_n)$ is a compact set. Denote this maximum $\boldsymbol{\nu}^*$. We thus have

$$\begin{aligned}
\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n) &\geq |\boldsymbol{\nu}^{*\top}(\mathbf{q}_{n\setminus i} - \mathbf{q}_n)| = \sum_{j=1}^{p} \nu^*_j (q^j_{n\setminus i} - q^j_n) \\
&\geq |\nu^*_j (q^j_{n\setminus i} - q^j_n)| = |\nu^*_j| |q^j_{n\setminus i} - q^j_n| \quad \text{for all } j = 1, \ldots p
\end{aligned}$$

where the second inequality is because $\nu^*_j (q^j_{n\setminus i} - q^j_n) > 0$ for all j because $\hat{R}(\cdot; S_n)$ is piecewise linear and nowhere flat. Thus we get, for all $j = 1, \ldots, p$,

$$|q^j_{n\setminus i} - q^j_n| \leq \frac{\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)}{|\nu^*_j|}.$$

Let us bound $\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)$. Note

$$\begin{aligned}
\hat{R}(\mathbf{q}_n; S_n) &= \frac{n-1}{n} \hat{R}(\mathbf{q}_n; S_n^{\setminus i}) + \frac{1}{n} \hat{R}(\mathbf{q}_n; S_i) \\
&\geq \frac{n-1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_n^{\setminus i})
\end{aligned}$$

since $\mathbf{q}_{n\setminus i}$ is the minimizer of $\hat{R}(\cdot; S_n^{\setminus i})$. Also, $\hat{R}(\mathbf{q}_n; S_n) \leq \hat{R}(\mathbf{q}_{n\setminus i}; S_n)$ since q_n is by definition the minimizer of $\hat{R}(\cdot; S_n)$. Putting these together, we get

$$\begin{aligned}
&\frac{n-1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_n^{\setminus i}) - \hat{R}(\mathbf{q}_{n\setminus i}; S_n) \leq \hat{R}(\mathbf{q}_n; S_n) - \hat{R}(\mathbf{q}_{n\setminus i}; S_n) \leq 0 \\
&\implies |\hat{R}(\mathbf{q}_n; S_n) - \hat{R}(\mathbf{q}_{n\setminus i}; S_n)| \leq \left| \frac{n-1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_n^{\setminus i}) - \hat{R}(\mathbf{q}_{n\setminus i}; S_n) \right| \\
&\qquad = \left| \frac{n-1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_n^{\setminus i}) - \frac{n-1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_n^{\setminus i}) - \frac{1}{n} \hat{R}(\mathbf{q}_{n\setminus i}; S_i) \right| \\
&\qquad = \frac{1}{n} |\hat{R}(\mathbf{q}_{n\setminus i}; S_i)|.
\end{aligned}$$

Thus

$$\begin{aligned}
\sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} (b \vee h) |q_n(\mathbf{x}) - q_{n\setminus i}(\mathbf{x})| &\leq \sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} (b \vee h) \left(\sum_{j=1}^{p} |q^j_n - q^j_{n\setminus i}| |x_j| \right) \\
&\leq \sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} (b \vee h) \cdot \sum_{j=1}^{p} \frac{|x_j|}{|\nu^*_j|} \cdot (\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)) \\
&\leq \sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} \frac{b \vee h}{n} \cdot \sum_{j=1}^{p} \frac{|x_j|}{|\nu^*_j|} \cdot |\hat{R}(\mathbf{q}_{n\setminus i}; S_i)|. \qquad (16)
\end{aligned}$$

We can further simplify the upper bound (16) as follows. Recall that $\boldsymbol{\nu}^*$ is the subgradient of $\hat{R}(\cdot; S_n)$ at $\mathbf{q}_n$ that maximizes $\sum_{j=1}^p \nu_j(q^j_{n\setminus i} - q^j_n)$; and as $\partial\hat{R}(\mathbf{q}_n; S_n)$ is compact (by the convexity of $\hat{R}(\cdot; S_n)$), we can compute $\boldsymbol{\nu}^*$ exactly. It is straightforward to show:

$$\nu^*_j = \begin{cases} -bx_j & \text{if } q^j_{n\setminus i} - q^j_n \le 0 \\ hx_j & \text{if } q^j_{n\setminus i} - q^j_n \ge 0 \ \ \forall\ j. \end{cases}$$

We can thus bound $1/|\nu^*_j|$ by $1/[(b \wedge h)|x_j|]$. By using the tight uniform upper bound $(b \vee h)\bar{D}$ on each term of $|\hat{R}(\cdot,\cdot)|$ from Lemma 1, we get the desired result. $\square$

We move onto the main result needed to prove Proposition 3. First, we build some terminology.

Definition 2 (σ-admissible loss function). A loss function ℓ defined on $\mathcal{Q}\times\mathcal{D}$ is σ-admissible with respect to $\mathcal{Q}$ if the associated convex function c is convex in its first argument and the following condition holds:

$$\forall\ y_1, y_2 \in \mathcal{Y},\ \forall\ d \in \mathcal{D},\ |c(y_1, d) - c(y_2, d)| \le \sigma|q_1 - q_2|,$$

where $\mathcal{Y} = \{y :\ \exists\ q \in \mathcal{Q},\ \exists\ \mathbf{x} \in \mathcal{X} : q(\mathbf{x}) = y\}$ is the domain of the first argument of c.

Theorem 7 **(Bousquet and Elisseeff (2002) Theorem 22 pp. 514)**. *Let $\mathcal{F}$ be a reproducing kernel Hilbert space with kernel k such that $\forall\ x \in \mathcal{X}$, $k(x,x) \le \kappa^2 < \infty$. Let ℓ be σ-admissible with respect to $\mathcal{F}$. The learning algorithm A defined by*

$$A_{S_n} = \operatorname*{arg\,min}_{g\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(g, z_i) + \lambda\|g\|_k^2$$

has uniform stability α_n wrt ℓ with

$$\alpha_n \le \frac{\sigma^2\kappa^2}{2\lambda n}.$$

Note that $\mathbb{R}^p$ is a reproducing kernel Hilbert space where the kernel is the standard inner product. Thus, κ in our case is $X_{\max}$.

Proof. (Of Proposition 3) By the Lipschitz property of $C(\cdot; d)$,

$$\sup_{d\in\mathcal{D}} |C(q_1(\mathbf{x}), d) - C(q_2(\mathbf{x}), d)| \le (b \vee h)|q_1(\mathbf{x}) - q_2(\mathbf{x})|, \quad \forall q_1(\mathbf{x}), q_2(\mathbf{x}) \in \mathcal{Q}$$

as before, hence $C : \mathcal{X}\times\mathcal{D} \to \mathbb{R}$ is $(b \vee h)$-admissible. Hence by Theorem 7 the algorithm (NV-ML2) has uniform stability with parameter α^r_n as given. $\square$

We have thus far established the stability of the big-data newsvendor algorithms (NV-ML1) and (NV-ML2), which lead to the risk bounds provided in Theorems 1 and 2, as follows.

Denote the generic true and empirical risks for general algorithm A as:

$$R_{true}(A, S_n) := \mathbb{E}_{z_{n+1}}[\ell(A_{S_n}, z_{n+1})] \text{ and } \hat{R}(A, S_n) := \frac{1}{n}\sum_{i=1}^n \ell(A_{S_n}, z_i).$$

LEMMA 2. *Let A be an algorithm with uniform stability α_n with respect to a loss function ℓ such that $0 \le \ell(A_{S_n}, z) \le M$, 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:*

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

Proof. (Of Lemma 2) The result is obtained by extending Theorem 12 of Bousquet and Elisseeff (2002) on pp. 507 by using the two-sided version of McDiarmid's inequality. □

We can now put the results together to prove Corollaries 1 and 2.

Proof. (Of Theorem 1) The result follows from Proposition 2 and Lemma 2. □

Proof. (Of Theorem 2) By Lemma 1, $0 \le \ell(A_S, z) \le \bar{D}(b \vee h)$ for all $z \in \mathcal{Z}$ and all sets S. The result then follows from Proposition 3 and Lemma 2. □

We now prove results for (NV-KO).

Proof. (Of Proposition 4) The proof parallels that of Proposition 2. Symmetry follows from the fact that the data-generating process is iid. For stability, we will change out notation slightly to make the dependence on n and S_n explicit. Let

$$q_n^\kappa = \arg\min_{q \ge 0} \ \tilde{R}(q; S_n, \mathbf{x}_{n+1}) = \operatorname*{argmin}_{q \ge 0} \ \frac{\sum_{j=1}^n \kappa_j [b(d_j - q)^+ + h(q - d_j)^+]}{\sum_{j=1}^n \kappa_j}$$

be the solution to (NV-KO) for the set S_n, and

$$q_{n\setminus i}^\kappa = \arg\min_{q \ge 0} \ \tilde{R}(q; S_n^{\setminus i}, \mathbf{x}_{n+1}) = \operatorname*{argmin}_{q \ge 0} \ \frac{\sum_{j \ne i} \kappa_j [b(d_j - q)^+ + h(q - d_j)^+]}{\sum_{j \ne i} \kappa_j}$$

be the solution to (NV-KO) for the set $S_n^{\setminus i}$. Note that:

$$\tilde{R}(q; S_n, \mathbf{x}_{n+1}) = \frac{\sum_{j \ne i} \kappa_j}{\sum_j \kappa_j} \tilde{R}(q; S_n^{\setminus i}, \mathbf{x}_{n+1}) + \frac{1}{\sum_j \kappa_j} \hat{R}(q; S_i, \mathbf{x}_{n+1}),$$

where $S_i = (\mathbf{x}_i, d_i)$.

By definition, the algorithm is stable if for all $S_n \in \mathcal{Z}^n$ and $i \in \{1, \dots, n\}$,

$$\sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} |C(q_n^\kappa, D(\mathbf{x})) - C(q_{n\setminus i}^\kappa, D(\mathbf{x}))| \le \alpha_n,$$

where $\alpha_n \le O(1/n)$. Now for a fixed $\mathbf{x}$, we have, by the Lipschitz property of $C(q; \cdot)$,

$$\sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} |C(q_n^\kappa, D(\mathbf{x})) - C(q_{n\setminus i}^\kappa, D(\mathbf{x}))| \le \sup_{(\mathbf{x}, D) \in \mathcal{X} \times \mathcal{D}} (b \vee h) |q_n^\kappa - q_{n\setminus i}^\kappa|.$$

(See Fig. A). So we want to bound $|q_n^\kappa - q_{n\setminus i}^\kappa|$.

By the convexity of the function $\tilde{R}_n(\cdot; S_n, \mathbf{x}_{n+1})$, we have (see Section 23 of Rockafellar (1997)):

$$\nu(q_{n\setminus i}^\kappa - q_n^\kappa) \le \tilde{R}(q_{n\setminus i}^\kappa; S_n, \mathbf{x}_{n+1}) - \tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})$$

for all $\nu \in \partial\tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})$ (set of subgradients of $\tilde{R}(\cdot; S_n, \mathbf{x}_{n+1})$ at q_n^κ). Further, because $0 \in \partial\tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})$ by the optimality of q_n^κ, we have

$$0 \leq \max_{\nu \in \partial\tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})} \nu(q_{n\setminus i}^\kappa - q_n^\kappa) \leq \tilde{R}(q_{n\setminus i}^\kappa; S_n, \mathbf{x}_{n+1}) - \tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})$$

where the max over ν can be attained because $\partial\tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})$ is a compact set. Denote this maximum ν^*.

Following arguments parallel those of the proof for (2),

$$\begin{aligned}\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)|q_n^\kappa - q_{n\setminus i}^\kappa| &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{b\vee h}{|\nu^*|}\cdot(\tilde{R}(q_{n\setminus i}^\kappa; S_n, \mathbf{x}_{n+1}) - \tilde{R}(q_n^\kappa; S_n, \mathbf{x}_{n+1})) \\ &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{(b\vee h)\kappa_i}{|\nu^*|\sum_j \kappa_j}\cdot|\tilde{R}(q_{n\setminus i}^\kappa; S_i, \mathbf{x}_{n+1})|. \end{aligned} \tag{17}$$

We can further simplify the upper bound (17) as follows. Consider

$$\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{\kappa_i}{\sum_j \kappa_j} = \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{1}{1+\sum_{j\neq i}\kappa_j/\kappa_i}.$$

The supremum is thus achieved by the infimum of the ratio κ_j/κ_i over $\mathcal{X}\times\mathcal{X}\times\mathcal{X}$. For the Gaussian kernel,

$$\inf_{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_{n+1})\in\mathcal{X}\times\mathcal{X}\times\mathcal{X}} \frac{K_h(\mathbf{x}_{n+1}-\mathbf{x}_j)}{K_h(\mathbf{x}_{n+1}-\mathbf{x}_i)} = \frac{e^{-4X_{\max}^2/2h^2}}{e^0} = e^{-2X_{\max}^2/h^2} := r_h.$$

Finally, we can bound $1/|\nu_j^*|$ by $1/(b\wedge h)$, as in the proof for Proposition 2. By using the tight uniform upper bound $(b\vee h)\bar{D}$ on each term of $|\tilde{R}(\cdot;\cdot,\mathbf{x}_{n+1})|$ from Lemma 1, we get the desired result. □

Proof. (Of Theorem 5) The result follows from Proposition 4 and Lemma 2. □

Before proving Theorems 3, 4 and 6, we first prove the following lemma, that shows the equivalence of the newsvendor objective function to the nonparametric regression loss function, save a constant factor.

LEMMA 3. *The newsvendor objective function equivalent to the nonparametric regression loss function $H_r(\cdot)$, save a constant multiplier.*

$$C(q;D) \equiv \frac{1}{2}H_r(q-D) := \frac{1}{2}[|q-D| + (2r-1)(q-D)],$$

where $r = b/(b+h)$.

Proof.

$$\begin{aligned} H_r(q-D) &:= |q-D| + (2r-1)(q-D) \\ &= (q-D)^+ + (D-q)^+ + 2r(q-D) - (q-D) \\ &= (q-D)^+ + (D-q)^+ + 2r[(q-D)^+ - (D-q)^+] - (q-D)^+ + (D-q)^+ \\ &= 2r(q-D)^+ + 2(1-r)(D-q)^+ \\ &= 2[r(q-D)^+ + (1-r)(D-q)^+]. \quad \square \end{aligned}$$

Let us also state the mild regularity conditions assumed in Theorems 4,4 and 6 before going onto the proofs.

Condition 1. For $u=(u_1,\ldots,u_p)\in\mathbb{Z}^p$, let D^u denote the differential operator $\partial^{[u]}/\partial x_1^{u_1}\ldots\partial x_p^{u_p}$, where $[u]=u_1+\ldots+u_p$. Let V be some fixed open neighborhood of 0 in $\mathbb{R}^p$. Then for a fixed nonnegative integer k and real numbers c and γ such that $c>0$ and $0<\gamma\leq 1$, we assume that the (true) r-th conditional quantile function $q^*(\mathbf{x})$ of the demand is such that

- $D^u q^*(\mathbf{x})$ exists and is continuous in $\mathbf{x}$ for all $\mathbf{x}\in V$ and $[u]\leq k$,
- $|D^u q^*(\mathbf{x})-D^u q^*(0)|\leq c\|\mathbf{x}\|^\gamma$ for all $\mathbf{x}\in V$ and $[u]\leq k$.

In other words, $q^*(\mathbf{x})$ is continuously differentiable up to order k on V and their k-th derivatives are uniformly Hölder continuous at 0 with exponent γ. We refer to $s=k+\gamma$ as the *order of smoothness* of the function $q^*(\mathbf{x})$.

Proof. (Of Theorem 3)

$$\begin{aligned}|R_{true}(q^*)-\hat{R}_{in}(\hat{q};S_n)| &\leq |R_{true}(\hat{q})-\hat{R}_{in}(\hat{q};S_n)|+|R_{true}(q^*)-R_{true}(\hat{q})|\\ &\leq \frac{2(b\vee h)^2\bar{D}}{b\wedge h}\frac{p}{n}+\left(\frac{4(b\vee h)^2\bar{D}}{b\wedge h}p+\bar{D}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}+|R_{true}(q^*)-R_{true}(\hat{q})|,\end{aligned}$$

because the first term in the last inequality is the generalization bound in Theorem 1. As in the proof of Theorem 6, we can bound the second term by

$$|R_{true}(q^*)-R_{true}(\hat{q})|\leq (b\vee h)\mathbb{E}|q^*-\hat{q}|.$$

Finally, the term $\mathbb{E}|q^*-\hat{q}|$ is the finite-sample bias of the decision $\hat{q}$, which is the $b/(b+h)$ quantile of the conditional distribution of $D|\mathbf{x}_{n+1}$ by Lemma 3. The result then follows from Condition 1 and Theorem 3.2. of Chaudhuri et al. (1991), which provides the optimal convergence rate of $\hat{q}$ to q^* in the asymptotic minimax sense. □

Proof. (Of Theorem 4)

$$\begin{aligned}|R_{true}(q^*)-\hat{R}_{in}(\hat{q}_\lambda;S_n)| &\leq |R_{true}(\hat{q}_\lambda)-\hat{R}_{in}(\hat{q}_\lambda;S_n)|+|R_{true}(\hat{q})-R_{true}(\hat{q}_\lambda)|+|R_{true}(q^*)-R_{true}(\hat{q})|\\ &\leq (b\vee h)\left[\frac{(b\vee h)X_{\max}^2}{n\lambda}+\left(\frac{2(b\vee h)X_{\max}^2}{\lambda}+\bar{D}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}\right]\\ &\quad +|R_{true}(\hat{q})-R_{true}(\hat{q}_\lambda)|+|R_{true}(q^*)-R_{true}(\hat{q})|\end{aligned}$$

because the first term in the last inequality is the generalization bound in Theorem 2. As in the proof of Theorem 6, we can bound the second and the third terms by

$$\begin{aligned}&|R_{true}(\hat{q})-R_{true}(\hat{q}_\lambda)|\leq (b\vee h)\mathbb{E}|\hat{q}-\hat{q}_\lambda|, \text{ and}\\ &|R_{true}(q^*)-R_{true}(\hat{q})|\leq (b\vee h)\mathbb{E}|q^*-\hat{q}|.\end{aligned}$$

The first term $\mathbb{E}|\hat{q} - \hat{q}_\lambda|$ is the finite-sample bias that results from regularization, which depends on the exact regularization used and on the demand distribution. The second term $\mathbb{E}|q^* - \hat{q}|$ is the finite-sample bias of the decision $\hat{q}$, which we bound as in the proof for Theorem 3. $\square$

Proof. (Of Theorem 6) For the uniform kernel $K_h(\mathbf{u}) = (2h)^{-1}\mathbf{1}(\|\mathbf{u}\| \le h)$ with $h \ge 2X_{\max}$, $r_h = 1$ and we have

$$\begin{aligned}|R_{true}(q^*) - \hat{R}_{in}(\hat{q}^\kappa; S_n)| &\le |R_{true}(\hat{q}^\kappa) - \hat{R}_{in}(\hat{q}^\kappa; S_n)| + |R_{true}(q^*) - R_{true}(\hat{q}^\kappa)| \\ &\le \frac{2(b \vee h)^2 \bar{D}}{b \wedge h}\frac{1}{n} + \left(\frac{4(b \vee h)^2 \bar{D}}{b \wedge h} + \bar{D}\right)\sqrt{\frac{\ln(2/\delta)}{2n}} \\ &\quad + |R_{true}(q^*) - R_{true}(\hat{q}^\kappa)|,\end{aligned}$$

because the first term in the last inequality is the generalization bound in Theorem 5. As in the proof of Theorem 6, we can bound the second term by

$$|R_{true}(q^*) - R_{true}(\hat{q}^\kappa)| \le (b \vee h)\mathbb{E}|q^* - \hat{q}^\kappa|.$$

Finally, the term $\mathbb{E}|q^* - \hat{q}^\kappa|$ is the finite-sample bias of the decision $\hat{q}^\kappa$, which is a weighted kernel estimator for the $b/(b+h)$-th conditional quantile for the demand, by Lemma 3. The result then follows from Condition 1 and Theorem 3.2. of Chaudhuri et al. (1991) (which applies to the uniform kernel with bandwidth $h = h_n = O(n^{1/(2s+p)})$), which provides the optimal convergence rate of $\hat{q}^\kappa$ to q^* in the asymptotic minimax sense. $\square$

Appendix B: Failure Mechanisms of Previous Data-Driven Methods

In this section, we highlight the conceptual differences between the machine learning algorithms of Sec. 2 to two other main data-driven methods known in the literature, SAA and SEO. We accompany the conceptual differences with examples where the existing methods may not work, which makes the case for the detailed empirical investigation of Sec. 5.

In particular, previous approaches to the newsvendor problem rely on very strong assumptions. In this section we provide very simple cases where those assumptions do not hold, leading to provably suboptimal performance for the previous methods. In particular, the SAA approach does not use features, the SEO approach makes linearity and normality assumptions about the demand. The Big Data Newsvendor makes only the assumption of iid feature-demand pairs, which is a much weaker assumption. In this section, we provide very simple distributions where it is possible to prove that the previous approaches do not provide the correct result, and the new approach does provide the correct result.

B.1. Comparison with SAA

In SAA, one assumes that only past demand observations are available (whereas we consider relevant features about the demand as well). If there is a strong relationship between the demand and some feature, the SAA approach would yield biased and inconsistent decisions, unlike (NV-ML1). We illustrate this point with the following example.

Consider the following demand model:

$$D = D_0 + D_1 x,$$

where D_0 and D_1 are non-negative continuous random variables and $x \in \{0,1\}$ is a binary feature (e.g. 0 for weekday and 1 for weekend). Let p_0 be the proportion of time $x = 0$. We have n historical observations: $[(x_1, d_1), \ldots, (x_n, d_n)]$, of which $n_0 = np_0$ are when $x = 0$ and $n_1 = n - n_0$ are when $x = 1$ (assume rounding effects are negligible). Note the observations d_k can be decomposed into: $\{d_k | x_k = 0\} = d_k^0$ and $\{d_k | x_k = 1\} = d_k^0 + d_k^1$. Also let $r = b/(b+h)$ for ease of notation. Let F_0 and F_1 denote the cumulative distribution functions (cdfs), F_0^{-1} and F_1^{-1} denote the inverse cdfs, and f_0 and f_1 the probability density functions (pdfs) of D_0 and D_1 respectively.

In addition to continuity of F_0 and F_1, we have the following properties for F_0 and F_1.

Condition B1. Assume F_0 and F_1 are twice differentiable (i.e. f_0 and f_1 are differentiable) and that there exists a $0 < \gamma < 2$ such that

$$\sup_{0<y<1} y(1-y)\frac{|J_i(y)|}{f(F_i^{-1}(y))} \leq \gamma, \tag{18}$$

where $J_i(\cdot)$ is the *score function* of distribution F_i defined by

$$J_i(y) = \frac{-f_i'(F_i^{-1}(y))}{f_i(F_i^{-1}(y))} = -\frac{d}{dy} f_i(F_i^{-1}(y)). \tag{19}$$

Condition B1 is satisfied by many standard distributions such as uniform, exponential, logistic, normal, and log normal, for γ between 0 and $\sim$1.24. The critical ratios for the uniform, exponential and logistic distributions can be computed straight-forwardly; for the normal distribution it is easier to compute the critical ratio by using the following equivalent formulation for the critical ratio:

$$\sup_{x \in \mathrm{dom}(D)} F(x)(1-F(x)) \frac{|f'(x)|}{f(x)^2}. \tag{20}$$

It is then tedious but straight-forward to compute the supremum of the critical ratio over $-\infty < x < \infty$ for the normal. For the lognormal distribution, it is tedious but straight-forward to establish the continuity and boundedness of the critical ratio over $0 < x < \infty$, then to compute a bound to this supremum numerically. For more details see Parzen (1979).

Distribution	$f(F^{-1}(y))$	$J(y)$	Is $\sup_{0<y<1} y(1-y)\frac{\lvert J(y)\rvert}{f(F^{-1}(y))} < 2$?
Uniform	1	0	Yes, LHS $= 0$
Exponential	$1-y$	1	Yes, LHS $= 1$
Logistic	$y(1-y)$	$2y-1$	Yes, LHS $= 1$
Normal	$\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\lvert\Phi^{-1}(y)\rvert^2\}$	$\Phi^{-1}(y)$	Yes, LHS $= 1$
Lognormal	$\phi(\Phi^{-1}(y)) \exp\{-\Phi^{-1}(y)\}$	$\exp\{-\Phi^{-1}(y)\}(\Phi^{-1}(y)+1)$	Yes, LHS $\lesssim 1.24$

Table 3 Some standard distributions that satisfy the requirement of Condition B1. The standard normal cdf and pdf are denoted as $\Phi(\cdot)$ and $\phi(\cdot)$ respectively.

LEMMA 4 **(Optimal ordering decision of (NV-ML1))**. *Let $\hat{F}_i$ denote the empirical cdf of $D|x=i$ with n_i iid observations for $i=0,1$. Then the optimal decision that solves (NV-ML1) is given by*

$$\hat{q}_n^0 = \inf\left\{q : \hat{F}_0(q) \geq \frac{b}{b+h}\right\} = d^0_{(\lceil n_0 r \rceil)}, \text{ if } x_{n+1}=0$$
$$\hat{q}_n^0 + \hat{q}_n^1 = \inf\left\{q : \hat{F}_1(q) \geq \frac{b}{b+h}\right\} = d^1_{(\lceil n_1 r \rceil)}, \text{ if } x_{n+1}=1.$$

Put simply, $\hat{q}_n^0$ solves the SAA problem for the subsample of data corresponding to $x=0$ and $\hat{q}_n^0 + \hat{q}_n^1$ solves the SAA problem for the subsample of data corresponding to $x=1$.

Proof. (Of Lemma 4) The feature-based algorithm (NV-ML1) solves

$$\min_{q(x)=q^0+q^1 x} \hat{R}(q(x); S_n) = \frac{1}{n}\sum_{i=1}^{n} [b(d_i(x)-q(x))^+ + h(q(x)-d_i(x))^+]$$

$$= \min_{q(x)=q^0+q^1x} \frac{1}{n_0}\sum_{i:x_i=0}[b(d_i^0-q^0)^+ + h(q^0-d_i^0)^+] + \frac{1}{n_1}\sum_{i:x_i=1}[b(d_i^0+d_i^1-q^0-q^1)^+ + h(q^0+q^1-d_i^0-d_i^1)^+]$$

$$= \min_{q^0\geq 0}\left\{\frac{1}{n_0}\sum_{i:x_i=0}[b(d_i^0-q^0)^+ + h(q^0-d_i^0)^+] + \min_{q^1\geq 0}\left\{\frac{1}{n_1}\sum_{i:x_i=1}[b(d_i^0+d_i^1-q^0-q^1)^+ + h(q^0+q^1-d_i^0-d_i^1)^+]\right\}\right\}, \tag{21}$$

where the outer and inner minimization problems correspond to the SAA problem for the subsample of data corresponding to $x=0$ and $x=1$ respectively. Hence the solutions are the corresponding SAA solutions for the appropriate subsample of data, which is the well-known critical fractile of the inverse empirical cdf as in (4). □

PROPOSITION 5 **(Finite-sample bias and asymptotic optimality of (NV-ML1))**. *We can show*

$$\left|\mathbb{E}[\hat{q}_n^0] - F_0^{-1}(r)\right| \leq O\left(\frac{\log n}{n}\right)$$
$$\left|\mathbb{E}[\hat{q}_n^0+\hat{q}_n^1] - F_1^{-1}(r)\right| \leq O\left(\frac{\log n}{n}\right),$$

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

$$\lim_{n\to\infty} \hat{q}_n^0 \overset{a.s.}{=} F_0^{-1}(r) =: q_{opt}^0$$
$$\lim_{n\to\infty} \hat{q}_n^0+\hat{q}_n^1 \overset{a.s.}{=} F_1^{-1}(r) =: q_{opt}^1$$

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.

Proof. (Of Proposition 5) Under Condition B1, the following strong result holds via Theorem 4.1.2. pp. 31 of Csörgö (1983): there exists, for each n_i, a Brownian Bridge $\{B_{n_i}(y), 0\leq y\leq 1\}$ such that

$$\sup_{0<y<1}\left|f_i(F_i^{-1}(y))(\hat{F}_i^{-1}(y)-F_i^{-1}(y)) - \frac{B_{n_i}(y)}{\sqrt{n_i}}\right| \overset{a.s.}{=} O\left(\frac{\log n_i}{n_i}\right). \tag{22}$$

The above implies, for $y=r$:

$$\left|(\hat{F}_i^{-1}(r)-F_i^{-1}(r)) - \frac{B_{n_i}(r)}{f_i(F_i^{-1}(r))\sqrt{n_i}}\right| \overset{a.s.}{\leq} O\left(\frac{\log n_i}{n_i}\right)$$
$$\implies \left|\hat{F}_i^{-1}(r)-F_i^{-1}(r)\right| \overset{a.s.}{\leq} \frac{B_{n_i}(r)}{f_i(F_i^{-1})\sqrt{n_i}}(r) + O\left(\frac{\log n_i}{n_i}\right)$$
$$\implies \left|\mathbb{E}[\hat{F}_i^{-1}(r)] - F_i^{-1}(r)\right| \leq \mathbb{E}\left|\hat{F}_i^{-1}(r)-F_i^{-1}(r)\right| \leq \frac{\mathbb{E}B_{n_i}(r)}{f_i(F_i^{-1}(r))\sqrt{n_i}} + O\left(\frac{\log n_i}{n_i}\right) = O\left(\frac{\log n_i}{n_i}\right),$$

where the last line uses Jensen's inequality and the fact that the mean of a Brownian Bridge is zero everywhere. Hence we get both the finite-sample bias result and the asymptotic optimality result. □

LEMMA 5 **(Optimal SAA ordering decision)**. *Let F^{mix} denote the cdf of the mixture distribution $D^{mix} = p_0 D_0 + (1-p_0)D_1$ and $\hat{F}_n^{mix}$ its empirical counterpart with n observations. Then the optimal SAA decision is given by*

$$\hat{q}_n^{SAA} = \inf\left\{q : \hat{F}_n^{mix}(q) \geq \frac{b}{b+h}\right\} = d_{(\lceil nr \rceil)}.$$

Proof. (Of Lemma 5) This is simply the SAA solution for the complete data set. □

PROPOSITION 6 **(Finite-sample bias and asymptotic (sub)-optimality of SAA)**. *With probability 1,*

$$\hat{q}_n^0 < \hat{q}_n^{SAA} < \hat{q}_n^0 + \hat{q}_n^1. \tag{23}$$

Moreover,

$$\left|\mathbb{E}[\hat{q}_n^{SAA}] - (F^{mix})^{-1}(r)\right| \leq O\left(\frac{\log n}{n}\right), \tag{24}$$

where $(F^{mix})^{-1}$ is the inverse cdf of D^{mix}. Hence we also have

$$\begin{aligned}\left|\mathbb{E}[\hat{q}_n^{SAA} - \hat{q}_n^0]\right| &= \left|(F^{mix})^{-1}(r) - F_0^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1)\\ \left|\mathbb{E}[\hat{q}_n^1 - \hat{q}_n^{SAA}]\right| &= \left|F_1^{-1}(r) - (F^{mix})^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1).\end{aligned} \tag{25}$$

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. In addition,

$$q_{opt}^0 < \lim_{n\to\infty} \hat{q}_n^{SAA} \overset{a.s.}{=} (F^{mix})^{-1}(r) < q_{opt}^1, \tag{26}$$

hence the SAA decision is not asymptotically optimal (is inconsistent).

Proof. (Of Proposition 6) Proof of (23). By assumption, the demand is almost surely greater when $x=1$ compared to when $x=0$. Hence the r-th quantile of the empirical distribution of D^{mix} is almost surely greater than the r-th quantile of the empirical distribution of $D|x=0$. The same observation holds for the second inequality.

Proof of (24) & (25). Proof of (24) parallels that of Proposition 5. Proof of (25) then follows from (24) and Proposition 5.

Proof of (26). The asymptotic convergence of $\hat{q}_n^{SAA}$ to its true value is again due to the asymptotic convergence of the sample quantile estimator, as shown in Proposition 5. The statement then follows from 23. □

As a final point, we remark that these observations are analogous in spirit to the bias and inconsistency of regression coefficients when there are, in econometric parlance, correlated omitted variables in the model [see for instance Greene (2003)].

B.2. Comparison with Separated Estimation and Optimization

One alternative, common-sense approach to incorporating feature information in the newsvendor decision-making is by first regressing the demand on the features assuming a normally distributed error term (estimation) then applying the appropriate formula for the optimal order quantity (optimization). Let us call this method separated estimation and optimization (SEO). The core of this method (and the key problem with it) is in the normality assumption of the residual error, and how it is constant over $\mathbf{x}$. In the following, we show that this may lead to nonsensical negative ordering decisions if the normality assumption does not hold. This is in contrast to (NV-ML1), a nonparametric method that yields sensible (small finite-sample bias and asymptotically optimal) ordering decisions.

Consider the following demand model:

$$D = \beta_0 + \beta_1 x + \varepsilon,$$

where β_0 and β_1 are non-negative constants, $x \in \{0,1\}$ is a binary feature and ε is a zero mean error term. Let p_0 be the proportion of time $x = 0$. We have n historical observations: $[(x_1, d_1), \ldots, (x_n, d_n)]$, of which $n_0 = np_0$ is the number of observations where $x = 0$ and $n_1 = n - n_0$ is the number of observations where $x = 1$ (assume rounding effects are negligible). Again let $r = b/(b+h)$ for ease of notation.

Under the SEO approach, one would assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, solve the vanilla newsvendor problem (1) under this assumption. One can show that the optimal newsvendor solution to (1) under the assumption $D(x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ is given by

$$q_{opt}(x) = \mu(x) + \sigma\Phi^{-1}(r). \tag{27}$$

To estimate $\mu(x)$, one can employ ordinary least squares (OLS) regression; that is, solve

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (d_i - \beta_0 - \beta_1 x_i)^2,$$

to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, yielding a mean estimate of $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. One can then estimate the variance by the standard error of regression (SER):

$$\hat{s}^2 = \frac{\sum_{i=1}^{n_1} (d_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n-1}.$$

The resulting order quantity under the SEO approach is thus

$$\hat{q}_{sep}(x) = \hat{\mu}(x) + \hat{s}\Phi^{-1}(r). \tag{28}$$

By properties of the OLS estimators $\hat{\mu}(x)$ and $\hat{s}^2(x)$, we make the following observation.

LEMMA 6. *If indeed* $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ *in truth, the order quantity* $\hat{q}_{sep}(x)$ *is unbiased and asymptotically optimal. That is,* $\mathbb{E}[\hat{q}_{sep}] = q_{opt}(x)$ *and* $\hat{q}_{sep} \xrightarrow{P} q_{opt}(x)$ *as* $n \to \infty$.

Proof. (Of Lemma 6) The result follows from the well-known fact that $\hat{\mu}(x)$ and $\hat{s}$ are unbiased and strongly consistent estimators of $\mu(x)$ and σ^2 respectively. For details, we refer the reader to Greene (2003). □

A problem arises, however, if the normality assumption $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ does not hold. In the following, we show that mis-specficiation of the model can lead to a nonsensical negative order quantity.

LEMMA 7 **(Negative order quantity with model misspecification)**. *Suppose* $0 < r < \Phi(-1)$ *and* $\varepsilon \sim \exp(\theta)$*, where*

$$0 < \theta < \frac{(\Phi^{-1}(1-r)-1)}{d_0+d_1}.$$

Then the SEO approach with the incorrect assumption $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ *yields a solution that is negative on average and almost surely in the limit as* n *tends to infinity.*

Proof. (Of Lemma 7) If $\varepsilon \sim \exp(\theta)$, $\mu(x) = \mathbb{E}[D|x] = d_0 + d_1 x + 1/\theta$ and $\sigma^2(x) = Var[D|x] = 1/\theta^2$. We have thus

$$\hat{q}_{sep}(x) = d_0 + d_1 x + \frac{1}{\hat{\theta}} + \frac{1}{\hat{\theta}}\Phi^{-1}(r) = d_0 + d_1 x + \frac{1}{\hat{\theta}} - \frac{1}{\hat{\theta}}\Phi^{-1}(1-r), \tag{29}$$

where $1/\hat{\theta}$ is the OLS estimator of $1/\theta$. Note the last equality is due to the identity $\Phi^{-1}(r) = -\Phi^{-1}(1-r)$, which holds because of the symmetry of the normal cdf. That this quantity is negative on average and in the limit follows from the unbiasedness and strong consistency of OLS estimators. □

B.3. Comparison with Operational Statistics

Our last comparison is with operational statistics (OS), which was first introduced by Liyanage and Shanthikumar (2005). The idea behind OS is to integrate parameter estimation and optimization rather than separate them. Let us illustrate how OS works by an example similar to the one used in Liyanage and Shanthikumar (2005).

Suppose the true demand has an exponential distribution, i.e. $D \sim \exp(1/\theta)$, and that the decision maker has access to $d_1, \ldots, d_n$ observations of past data. Then with straightforward calculations, one can show

$$\hat{q}_{SEO} = \log\left(\frac{b+h}{b}\right)\bar{d}_n,$$

where $\bar{d}_n$ is the sample average of the demand, is the optimal SEO order quantity. Now consider instead the decision

$$\hat{q}^1_{OS}(\alpha) = \alpha \bar{d}_n \tag{30}$$

parameterized by a constant $\alpha > 0$. The OS approach then picks α by the following optimization:

$$\min_{\alpha \geq 0} \mathbb{E}_\theta[C(\hat{q}^1_{OS}(\alpha); D)]. \tag{31}$$

As $\alpha = \log((b+h)/b)$ is a feasible solution of (31), this guarantees the OS decision to yield a true expected cost that is bounded above by the true expected cost of the SEO decision. In other words, by construction we have

$$\mathbb{E}_\theta[C(\hat{q}^1_{OS}(\alpha^*); D)] \leq \mathbb{E}_\theta[C(\hat{q}_{SEO}; D)], \tag{32}$$

where α^* is the optimal parameter in (31). With some computations, one can show

$$\alpha^* = \left[\left(\frac{b+h}{h} \right)^{1/n+1} - 1 \right] n.$$

Liyanage and Shanthikumar (2005) also shows that one can also improve upon the SAA optimal decision in terms of the true expected cost by considering the decision

$$\hat{q}^2_{OS}(\alpha, \beta) = d_{\lceil \beta - 1 \rceil} + \alpha(d_{\lceil \beta \rceil} - d_{\lceil \beta - 1 \rceil}), \tag{33}$$

where $\beta \in \{1, \ldots, n\}$ and $\alpha \geq 0$ are parameters to be chosen via

$$\min_{\alpha \geq 0, \beta \in \{1,\ldots,n\}} \mathbb{E}_\theta[C(\hat{q}^2_{OS}(\alpha, \beta); D)]. \tag{34}$$

As the above example illustrates, OS takes insight from the form of the decision derived by other methods (e.g. SEO and SAA) and constructively improves upon them in terms of the true expected cost simply by considering a decision that is a *function* of past demand data rather than a scalar quantity. In the parlance of our feature-based approach, the OS method is essentially considering meaningful statistics of past demand data as *features*. However, there is an important difference between the OS approach and ours, and this is in the way the unknown coefficients (parameters) of the decision function are chosen. Under our decision-making paradigm, one would simply input the sample average of past demand and differences of order statistics of past demand as features and choose the coefficients that minimize the *in-sample average cost*. In contrast, OS is based on the premise that one knows the distributional family the demand belongs to, and thus is able to compute the coefficients that minimize the *true expected cost*. That one knows the true distributional family is not a weak assumption, however the insights from OS analysis are not trivial. In Sec. 5, we will consider solving (NV-ML1) and (NV-ML2) both without and with OS-inspired features, to evaluate their practical benefit in terms of the out-of-sample cost.

Appendix C: In-sample empirical results

No. of past days	without OS Features		with OS Features	
	Avg. Cost	Total no. of Features (avg. chosen)	Avg. Cost	Total no. of Features (avg. chosen)
0	0.9785	20 (3.0)	0.9785	20 (3.0)
1	0.9882	32 (4.0)	0.9848	32 (6.3)
2	0.9893	44 (5.3)	0.9966	44 (9.1)
3	0.9896	56 (6.2)	0.9335	56 (21.1)
4	0.9716	68 (8.0)	**0.8937**	68 (28.1)
5	0.9711	80 (8.3)	0.9112	80 (36.7)
6	0.9700	92 (8.6)	0.9080	92 (43.4)
7	0.9691	104 (9.5)	0.9270	104 (52.4)
8	0.9688	116 (9.5)	0.9097	116 (56.5)
9	0.9687	128 (9.6)	0.9329	128 (62.6)
10	0.9690	140 (9.6)	0.9195	140 (69.7)
11	0.9693	152 (9.7)	0.9204	152 (73.4)
12	**0.9685**	164 (9.9)	0.9459	164 (82.2)
13	0.9689	176 (10.3)	0.8976	176 (91.8)
14	0.9686	188 (10.3)	0.9002	188 (97.2)

Table 4 Average in-sample cost of the solution to (NV-ML1) with the day of the week and time of the day features and an increasing number of days of past demands (for up to 2 weeks), with and without OS features. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of nonzero elements in the decision vector).

No. of past days	without OS Features		with OS Features	
	Avg. Cost	Total no. of Features (avg. chosen)	Avg. Cost	Total no. of Features (avg. chosen)
0	1.0945	20 (16.7)	1.0945	20 (16.7)
1	**1.0892**	32 (16.8)	1.0381	44 (27.3)
2	1.0960	44 (16.6)	1.0775	68 (37.6)
3	1.1142	56 (16.6)	0.9372	92 (48.5)
4	1.1220	68 (16.8)	0.9393	116 (59.5)
5	1.1360	80 (16.4)	0.9270	140 (71.2)
6	1.1525	92 (16.3)	**0.9202**	164 (83.1)
7	1.1487	104 (16.5)	0.9632	188 (95.4)
8	1.1596	116 (16.4)	1.0301	212 (108.3)
9	1.1638	128 (16.6)	1.0500	236 (118.6)
10	1.1587	140 (16.5)	1.0521	260 (129.7)
11	1.1694	152 (16.4)	1.0899	284 (141.4)
12	1.1601	164 (16.5)	1.0799	308 (153.5)
13	1.1540	176 (27.7)	1.0464	332 (165.5)
14	1.1658	188 (30.7)	1.0785	356 (178.2)

Table 5 Average in-sample cost of the solution to the SEO approach with day of the week and time of the day features and an increasing number of days of past demands (for up to 2 weeks), with and without OS features. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of elements in the decision vector that are at least as large as 1% of the largest element in absolute value).

Regularization Param.	without OS Features		with OS Features	
	Avg Cost	Total no. of Features (avg. chosen)	Avg Cost	Total no. of Features (avg. chosen)
1×10^{-4}	1.1450	188 (9.1)	0.6140	356 (4.7)
5×10^{-5}	1.0769	188 (12.0)	0.5793	356 (5.8)
1×10^{-5}	1.0750	188 (14.6)	0.5837	356 (7.2)
5×10^{-6}	1.0748	188 (15.3)	0.5837	356 (7.8)
1×10^{-6}	1.0117	188 (14.6)	0.5376	356 (10.3)
5×10^{-7}	0.9737	188 (13.0)	0.5165	356 (11.5)
1×10^{-7}	**0.9729**	188 (13.2)	**0.4489**	356 (28.1)

Table 6 Average in-sample cost of the solution to (NV-ML2) solved with ℓ_1 regularization, with the day of the week and time of the day features and 2 weeks of past demands with and without OS features for a range of regularization parameters. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of nonzero elements in the decision vector).

Regularization Param.	without OS Features		with OS Features	
	Avg. Cost	Total no. of Features (avg. chosen)	Avg. Cost	Total no. of Features (avg. chosen)
1×10^{-4}	1.1138	188 (8.7)	1.1153	356 (8.7)
5×10^{-5}	1.0626	188 (11.2)	1.0629	356 (11.2)
1×10^{-5}	1.0462	188 (13.0)	1.0463	356 (13.0)
5×10^{-6}	1.0434	188 (13.4)	1.0438	356 (13.7)
1×10^{-6}	0.9805	188 (13.2)	0.9276	356 (17.7)
5×10^{-7}	**0.9570**	188 (10.1)	0.9270	356 (21.6)
1×10^{-7}	0.9684	188 (10.2)	**0.9153**	356 (42.5)

Table 7 Average in-sample cost of the solution to (NV-ML2) solved with ℓ_2 regularization, with the day of the week and time of the day features and 2 weeks of past demands with and without OS features for a range of regularization parameters. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of elements in the decision vector that are at least as large as 1% of the largest element in absolute value).

Regularization Param.	without OS Features		with OS Features	
	Avg Cost	Total no. of Features (avg. chosen)	Avg Cost	Total no. of Features (avg. chosen)
1×10^{0}	1.1020	188 (9.9)	1.4849	356 (9.9)
5×10^{-1}	**1.0905**	188 (10.4)	1.2418	356 (10.4)
1×10^{-1}	1.0990	188 (14.9)	1.2318	356 (14.9)
5×10^{-2}	1.1023	188 (19.7)	1.1975	356 (15.9)
1×10^{-2}	1.1264	188 (34.0)	1.1307	356 (24.1)
5×10^{-3}	1.1365	188 (44.2)	**1.0636**	356 (28.5)

Table 8 Average in-sample cost of the solution to the SEO approach solved with ℓ_1 regularization, with day of the week and time of the day features and an increasing number of days of past demands (for up to 2 weeks), with and without OS features. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of elements in the decision vector that are at least as large as 1% of the largest element in absolute value).

Acknowledgments

This research was supported by National Science Foundation grant IIS-1053407 (Rudin) and the London Business School Research and Material Development Scheme (Vahn). The authors thank Nicos Savva and Stefan Scholtes for providing the hospital emergency room data. The authors would further like to thank seminar attendees at LBS, Columbia, NYU, MIT, Duke, Berkeley, Stanford, MSOM, Wharton EMPOM and Chung Piaw Teo for helpful suggestions.

References

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**(6) 716–723.

Besbes, Omar, Alp Muharremoglu. 2013. On implications of demand censoring in the newsvendor problem. *Management Science* **59**(6) 1407–1424.

Regularization Param.	without OS Features		with OS Features	
	Avg Cost	Total no. of Features (avg. chosen)	Avg Cost	Total no. of Features (avg. chosen)
1×10^{-1}	**1.1486**	188 (57.6)	1.1584	356 (92.8)
5×10^{-2}	1.1555	188 (61.2)	1.1357	356 (85.8)
1×10^{-2}	1.1613	188 (65.0)	1.0398	356 (88.0)
5×10^{-3}	1.1626	188 (65.7)	**1.0040**	356 (87.1)
1×10^{-3}	1.1669	188 (67.8)	1.0526	356 (76.2)
5×10^{-4}	1.1658	188 (68.6)	1.0636	356 (69.9)
1×10^{-4}	1.1654	188 (69.2)	1.0716	356 (64.3)

Table 9 Average in-sample cost of the solution to the SEO approach solved with ℓ_2 regularization, with day of the week and time of the day features and an increasing number of days of past demands (for up to 2 weeks), with and without OS features. The best in-sample results are highlighted in bold. The total number of features that appear in the decision are also reported (as measured by the number of elements in the decision vector that are at least as large as 1% of the largest element in absolute value).

Bousquet, Olivier, André Elisseeff. 2002. Stability and generalization. *The Journal of Machine Learning Research* **2** 499–526.

Chang, Allison, Cynthia Rudin, Michael Cavaretta, Robert Thomas, Gloria Chou. 2012. How to reverse-engineer quality rankings. *Machine Learning* **88** 369–398.

Chaudhuri, Probal, et al. 1991. Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of statistics* **19**(2) 760–777.

Csörgö, Miklos. 1983. *Quantile processes with statistical applications*. SIAM.

CVX Research, Inc. 2012. CVX: Matlab software for disciplined convex programming, version 2.0. `http://cvxr.com/cvx`.

Devroye, Luc, T Wagner. 1979a. Distribution-free inequalities for the deleted and holdout error estimates. *Information Theory, IEEE Transactions on* **25**(2) 202–207.

Devroye, Luc, T Wagner. 1979b. Distribution-free performance bounds for potential function rules. *Information Theory, IEEE Transactions on* **25**(5) 601–604.

Donnelly, L., M. Mulhern. 2012. Nhs pays £1,600 a day for nurses as agency use soars. `http://www.telegraph.co.uk/news/9400079/NHS-pays-1600-a-day-for-nurses-as-agency-use-soars.html`.

Gallego, Guillermo, Ilkyeong Moon. 1993. The distribution free newsboy problem: review and extensions. *Journal of the Operational Research Society* 825–834.

Grant, M., S. Boyd. 2008. Graph implementations for nonsmooth convex programs. V. Blondel, S. Boyd, H. Kimura, eds., *Recent Advances in Learning and Control*. Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 95–110. `http://stanford.edu/~boyd/graph_dcp.html`.

Green, Linda V, Sergei Savin, Nicos Savva. 2013. Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237–2256.

Greene, William H. 2003. Econometric analysis, 5th. *Ed.. Upper Saddle River, NJ* .

He, Biyu, Franklin Dexter, Alex Macario, Stefanos Zenios. 2012. The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing & Service Operations Management* **14**(1) 99–114.

Hoeffding, Wassily. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**(301) 13–30.

Huh, Woonghee Tim, Retsef Levi, Paat Rusmevichientong, James B Orlin. 2011. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research* **59**(4) 929–941.

Koenker, Roger. 2005. *Quantile regression*. Cambridge University Press.

Levi, Retsef, Georgia Perakis, Joline Uichanco. 2012. The data-driven newsvendor problem: new bounds and insights. *working paper* .

Levi, Retsef, Robin O Roundy, David B Shmoys. 2007. Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* **32**(4) 821–839.

Liyanage, Liwan H, J George Shanthikumar. 2005. A practical inventory control policy using operational statistics. *Operations Research Letters* **33**(4) 341–348.

Nadaraya, Elizbar A. 1964. On estimating regression. *Theory of Probability & Its Applications* **9**(1) 141–142.

Parzen, Emanuel. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association* **74**(365) 105–121.

Perakis, Georgia, Guillaume Roels. 2008. Regret in the newsvendor model with partial information. *Operations Research* **56**(1) 188–203.

Rockafellar, R Tyrell. 1997. *Convex analysis*, vol. 28. Princeton University Press.

Rogers, William H, Terry J Wagner. 1978. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics* 506–514.

Scarf, Herbert, KJ Arrow, S Karlin. 1958. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production* **10** 201–209.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* **6**(2) 461–464.

Shapiro, Alexander, Darinka Dentcheva, Andrzej P Ruszczyński. 2009. *Lectures on stochastic programming: modeling and theory*, vol. 9. SIAM.

Steinwart, Ingo, Andreas Christmann. 2011. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17**(1) 211–225.

Takeuchi, Ichiro, Quoc V Le, Timothy D Sears, Alexander J Smola. 2006. Nonparametric quantile estimation. *The Journal of Machine Learning Research* **7** 1231–1264.

Vapnik, Vladimir N. 1998. *Statistical learning theory*. Wiley.

Watson, Geoffrey S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 359–372.