MIT Sloan School of Management

MIT Sloan School Working Paper 5036-13

The Big Data Newsvendor: Practical Insights from Machine Learning Analysis

Cynthia Rudin, and Gah-Yi Vahn

This version February 6, 2014

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 past demands. Both *small data* ($p/n = o(1)$) and *big data* ($p/n = O(1)$) are considered. For both cases, we propose a machine learning algorithm to solve the problem and derive a tight generalization bound on the expected out-of-sample cost. The algorithms can be extended intuitively to other situations, such as having censored demand data, ordering for multiple, similar items and having a new item with limited data. We show analytically that our custom-designed, feature-based approach can be better than other data-driven approaches such as Sample Average Approximation (SAA) and separated estimation and optimization (SEO). Our method can also naturally incorporate the operational statistics method. We then apply the algorithms to nurse staffing in a hospital emergency room and show that (i) they can reduce the median out-of-sample cost by up to 46% and 16% compared to SAA and SEO respectively, with statistical significance at 0.01, and (ii) this is achieved either by carefully selecting a small number of features and applying the small data algorithm, or by using a large number of features and using the big data algorithm, which automates feature-selection.

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

History: February 6, 2014

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 data on features 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) and

various economic indicators (e.g. the interest rate and the consumer price index). With plummeting costs of data storage and processing, many organizations are systematically collecting or purchasing such information. This paper investigates the newsvendor problem for precisely this type of a situation. 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 "small data", i.e. the feature-to-observation ratio is small (formally, $p/n = o(1)$) and "big data", i.e. the feature-to-observation ratio is large (formally, $p/n = O(1)$). In this paper, we consider how the decision maker can choose an appropriate order quantity given a new decision period and a new set of features by learning from past data, in both the small and big data regimes.

In the classical newsvendor problem, one assumes the true demand distribution is known. Then the optimal order quantity is the critical fractile of the inverse cumulative distribution of the demand. In practice, however, it is quite restrictive to assume that the demand distribution is known, and in recent years there have been many efforts to relax this assumption. 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) first 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. In this paper, we build on Levi et al. (2007) by deriving a bound on the out-of-sample cost when feature data is available.

Other perspectives on the data-driven newsvendor include those of Liyanage and Shanthikumar (2005), who proposed ordering according to a statistic (function) 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 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 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.

One work that does consider feature information is He et al. (2012), who modeled booking a hospital operating room with two features (number and type of cases) as a feature-based newsvendor problem. Motivated by this work, we also investigate a case study in a healthcare setting in Sec. 6. Our work is, however, fundamentally different from He et al. (2012) in that (i) we investigate the effects of having a large number of features, whereas He et al. (2012) consider just two, and (ii) we focus primarily on probabilistic guarantees and theoretical insights whereas the work of He et al. (2012) is an empirical paper.

Summary of Contributions

In Sec. 2, we investigate the newsvendor problem when the decision-maker has access to past feature information as well as the demand. We show that the optimal order quantity can be learned via a linear programming (LP) algorithm in the case of small data (small p/n) and a regularization-based algorithm in the case of big data (large p/n). The latter algorithm is a quadratic program (QP) with L_2 regularization, an LP with L_1 regularization and a mixed-integer program (MIP) with L_0 regularization. Both algorithms can be used broadly for both iid as well as time-dependent data. The algorithms are based on the *empirical risk minimization principle* that has been widely adopted by the Machine Learning community for classification and regression problems [for an in-depth discussion of this principle, see Vapnik (1998)].

In Sec. 3, we provide generalization bounds on the out-of-sample cost of the data-driven newsvendor algorithms described in Sec. 2. Our bounds do not make any assumption about the feature-demand relationship, or the distribution of the demand beyond the existence of finite mean. Both results show how the out-of-sample cost (the "generalization error") of a decision deviates from the 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. The small data bound depends on p, hence it demonstrates the curse of dimensionality in generalizing the in-sample result to out-of-sample result when p is too large. On the other hand, the big data bound does not depend explicitly on p (it depends inversely on the regularization parameter instead). The practical implication is that given a large number of feature information at hand, the decision-maker is advised to either use a small subset of the available dataset by carefully choosing a small number of features, or use the whole dataset but regularize, which automates feature selection.

In Sec. 4, we show how the feature-based newsvendor model introduced in Sec. 2 can be extended to other realistic situations. First of all, we consider having data on product prices, sales, competition, budling 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 is censored, due to a constraint

on the maximal order quantity. We then consider a situation where one has a prior knowledge on the sign of the feature-demand information, and show that this can be incorporated by adding extra linear constraints to the original model. We further show that the following scenarios can be handled with fairly intuitive modifications of the original model: ordering for multiple items, where some features may play a similar role in predicting the demand for all products; and having a new product on the market with limited demand information, but where the new product is similar to old products for which one has sufficient data.

In Sec. 5, we give theoretical justifications for using our feature-based approach over other main data-driven approaches known in the literature. First, we compare our method with using the SAA method without any feature information. We prove that using the featureless SAA approach can lead to an ordering decision that is biased and asymptotically sub-optimal whereas our feature-based algorithm yields a near-unbiased (in the sense that the bias is bounded by $O(\log n/n)$) and asymptotically optimal decision. Second, we consider the separated estimation and optimization approach — a common sense approach of first estimating and then deciding. In this method, the decision maker performs a regression to estimate the conditional demand distribution, then applies the corresponding optimal newsvendor ordering formula. The key to this approach, however, is in the distributional assumption required for the regression, and we show that this approach can lead to a nonsensical negative order quantity if the model is mis-specified. In practice, it is easy to mis-specify such a model. Lastly, we show that our feature-based algorithms can incorporate the operational statistics (OS) approach of Liyanage and Shanthikumar (2005) quite naturally.

Finally, in Sec. 6, we demonstrate that our algorithms can be effective on a real dataset. Specifically, we show that the nurse staffing cost in the emergency room of a large teaching hospital in the United Kingdom can be reduced by up to 46% compared to the featureless SAA approach (in terms of the median out-of-sample cost, statistically significant at the 1% level) by appropriately incorporating high-dimensional feature data. Our algorithms are also better than the separated estimation and optimization approach and Scarf's Minimax approach by 16% and 29% respectively (also in terms of the median out-of-sample cost, statistically significant at the 1% level).

Before proceeding, we also mention that just as the featureless newsvendor algorithm performs quantile estimation, the basic version of our feature-based newsvendor algorithm reduces to non-parametric quantile regression. The results in this paper thus also extend the literature on non-parametric quantile regression [see Koenker (2005) for a general reference on quantile regression, Takeuchi et al. (2006) and Steinwart and Christmann (2011) for up-to-date results at the time of writing].

2. Algorithms for 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

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

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 a realistic situation, the data-driven newsvendor problem is too simplistic to represent many real situations because one can collect data on exogenous information about the demand as well as the demand itself. In other words, the newsvendor has access to a richer information base from which s/he can make the present decision. We thus consider the newsvendor who has access to the historical data $S_n = [(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_n, d_n)]$, where $\mathbf{x}_i = [x_i^1, \ldots, x_i^p]$ represents *features* about the demand such as seasonality (day, month, season), weather and planning data for the local area. It is possible for the decision maker to have *big* data, where the number of features p is of a non-negligible size compared to the number of observations n, that is, $p/n = O(1)$. We assume the newsvendor observes the features $\mathbf{x}_{n+1}$ before making the next ordering decision.

The goal now is to compute an order quantity at the beginning of period $n+1$, after having observed the features $\mathbf{x}_{n+1}$. Thus the problem now becomes that of finding the optimal *function* $q(\cdot)$ that maps the observed features $\mathbf{x}_{n+1} \in \mathcal{X}$ to an order $q(\mathbf{x}_{n+1}) \in \mathbb{R}$. Then the data-driven newsvendor problem with features is:

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

where $\hat{R}$ is called the *empirical risk* of function q with respect to dataset S_n. This algorithm is based on the *empirical risk minimization principle* that has been widely adopted by the Machine Learning community for classification and regression problems [for an in-depth discussion of this principle, see Vapnik (1998)].

To solve (NV-features), 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

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

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. The choice of $\mathcal{Q}$ can then be made more or less complex depending on which transformations are included. We can solve (NV-features) via the following linear program:

2.4. NV Algorithm with Features

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

$$
\begin{aligned}
&\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad \frac{1}{n}\sum_{i=1}^{n}(bu_i + ho_i) \\
&s.t.\ \forall\ i=1,\ldots,n: \\
&\qquad\qquad u_i \geq d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j \\
&\qquad\qquad o_i \geq q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i \\
&\qquad\qquad u_i, o_i \geq 0
\end{aligned}
\tag{NV-algo}
$$

where the dummy variables u_i and o_i represent, respectively, underage and overage costs in period i.

Linear decision rules have also been studied in the stochastic programming literature [see for example, Ben-Tal et al. (2005), Chen et al. (2007) and Chen et al. (2008)]. There is a fundamental difference, however, because whereas our decision rule is a linear combination of primitive features, in stochastic programming the decision rule is a linear combination of underlying uncertainties.

In the case of big data, i.e., when the ratio of the number of features to observations p/n is $O(1)$, one could solve the LP (NV-algo) by selecting a subset of the most relevant features according to some criterion, for example via model selection criteria such as the Akaike Information Criterion [Akaike (1974)] or Bayesian Information Criteria [Schwarz (1978)]. Alternatively, one could automate the feature-selection by solving the following *regularized* version of (NV-algo):

2.5. NV Algorithm with Regularization

$$
\begin{aligned}
\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \quad \hat{R}(q(\cdot);S_n) + \lambda\|q\|_2^2 &= \frac{1}{n}\sum_{i=1}^{n}[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+] + \lambda\|\mathbf{q}\|_2 \\
&\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \quad \frac{1}{n}\sum_{i=1}^{n}(bu_i + ho_i) \\
&s.t.\ \forall\ i=1,\ldots,n: \\
&\qquad\qquad u_i \geq d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j \\
&\qquad\qquad o_i \geq q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i \\
&\qquad\qquad u_i, o_i \geq 0,
\end{aligned}
\tag{NV-reg}
$$

where $\lambda > 0$ is the regularization parameter and $\|\mathbf{q}\|_2$ denotes the L_2-norm of the vector $\mathbf{q} = [q^1,\ldots,q^p]$. This problem is a quadratic program (QP), which 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. That is, the regularization term changes from $\|\mathbf{q}\|_2$ to either $\|\mathbf{q}\|_0$ or $\|\mathbf{q}\|_1$. The resulting problem then becomes, respectively, a mixed-integer program (MIP) or an LP.

Further, 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 by the group lasso technique [Yuan and Lin (2006)], whereby a regularization term

$$\sum_{g=1}^{G} \|q_{\mathcal{I}_g}\|_2$$

is included, 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 over groups.

3. Generalization Bounds on the Out-of-Sample Cost

In what follows, we provide probabilistic bounds on the out-of-sample cost of the ordering decisions chosen by (NV-algo) and (NV-reg). As we formulated the algorithms to fit into the framework of empirical risk minimization, we are able to use modern tools from machine learning.

Let us define the *true risk* as 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,

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

We are interested in minimizing this cost, but we cannot measure it as the distribution is unknown. Recall that the empirical risk is the average cost over the training sample:

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

The empirical risk can be calculated using the data, and we would wish that a combination of the empirical risk and other calculable features lead to a bound on the true risk.

Statistical learning theory provides the foundation for creating bounds on the true risk, in terms of the empirical risk and a complexity ("generalization error") term. These bounds highlight the important quantities for learning, as they appear directly in the complexity term. Most often, statistical learning theory bounds are uniform bounds, meaning they are applicable for all decisions in some function class. One downside, however, is that uniform bounds are not algorithm-specific and thus do not consider the way in which the algorithm traverses the space of possible models. The bounds we provide in this section are thus not uniform bounds. Instead, we provide bounds based

on *algorithmic stability theory*, which does consider the solution created by the algorithm. To use algorithmic stability theory, we first show that the algorithm is "stable", meaning that we bound how much the algorithm's predictions change when one of the training examples is removed[1]. If the algorithm is robust to such a change, it tends to generalize better to new observations, and this is quantified by the bound.

In what follows, the random demand is denoted by D, and is assumed to be bounded: $D \in \mathcal{D} := [0, \bar{D}]$. The feature domain is also bounded, in particular, we assume all feature vectors live in a ball: $\|\mathbf{x}\|_2^2 \leq X_{\max}$ 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 B.

THEOREM 1 **(Generalization Bound for (NV-algo))**. *Let $\hat{q}$ be the model produced by Algorithm (NV-algo). Define $\bar{D}$ as the maximum value of the demand we are willing to consider. 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 $\mathcal{X} \times \mathcal{D}$:*

$$|R_{true}(\hat{q}) - \hat{R}(\hat{q}; S_n)| \leq \frac{2(b \vee h)^2 \bar{D}}{b \wedge h} \frac{p}{n} + \left(\frac{4(b \vee h)^2 \bar{D}}{b \wedge h} p + \bar{D} \right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{5}$$

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), which is the key result behind the sample-size bounds of Levi et al. (2007).

On the other hand, for large p, Theorem 1 is not very informative. In Sec. 2, we suggested regularizing the original algorithm instead. Below we present the generalization bound for (NV-reg), which serves as a theoretical justification for regularizing in the case of big data.

THEOREM 2 **(Generalization Bound for (NV-reg))**. *Define $X_{\max}^2$ as the largest possible value of $\|\mathbf{x}\|_2^2$ that we are willing to consider. Let $\hat{q}$ be the model produced by Algorithm (NV-reg). Define $\bar{D}$ as the maximum value of the demand we are willing to consider. 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 $\mathcal{X} \times \mathcal{D}$:*

$$|R_{true}(\hat{q}) - \hat{R}(\hat{q}; S_n)| \leq \frac{(b \vee h)^2}{X_{\max}^{-2}} \frac{1}{n\lambda} + \left(\frac{2(b \vee h)^2}{X_{\max}^{-2}} \frac{1}{\lambda} + \bar{D} \right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{6}$$

This bound does not depend explicitly on p, indicating that Algorithm (NV-reg) can handle problems with the curse of dimensionality through regularization. In fact p can implicitly enter the bound through the choice of the regularization parameter λ, which should necessarily be chosen depending on the ratio of features to dimensions. For large p, the bound indicates that it is sensible to choose $\lambda \leq O(1/p)$, for example $\lambda = O(1/p^2)$. Note additionally that λ should be chosen relative to $X_{\max}^2$.

Last but not the least, both bounds of Theorem 1 and 2 scale appropriately with δ, as $\mathcal{O}(\sqrt{\ln(1/\delta)})$.

To illustrate how the generalization error scales with the various parameters, we plot the generalization error of Theorem 1 as the number of observation n grows in Fig. 1. In Fig. 1 (a), we show the error versus n for a fixed accuracy level $1-\delta=0.9$ and varying p, and in (b), the error versus n for fixed $p=4$ and varying δ. We abbreviate illustrating the generalization error in Theorem 2 as the only difference is that the scaling of the error in λ is inverse to that of in p.

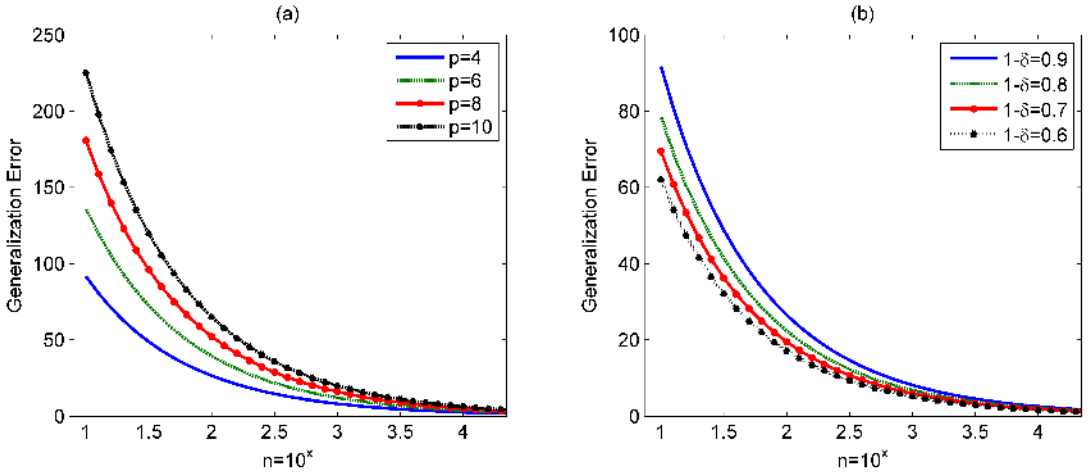

Figure 1 Plots of the generalization error in Theorem 1 as the number of observation n grows for (a) different p with fixed $1-\delta=0.9$ and (b) different δ with fixed $p=4$.

4. Extensions to the Big Data Newsvendor

As the feature-based newsvendor algorithms can incorporate complex data, it can be altered to accommodate extensions of the basic scenario in a straight-foward manner. We outline several realistic scenarios that may arise, and show how the algorithms can be altered to accommodate them.

4.1. 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.

4.2. 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

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \quad \sum_{i:\text{demand less than capacity}} [b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+] + \sum_{i:\text{demand equals capacity}} [b(q_{\max} - q(\mathbf{x}_i))^+] + \lambda\|\mathbf{q}\|_2.$$

4.3. Prior Knowledge on Signs of Coefficients

Often we have prior knowledge on whether a certain feature should have a positive or negative influence on the order quantity. In that case, we can constrain the coefficient to have the required sign. For instance, if feature j represents outdoor temperature, and $q(\mathbf{x}_{i'})$ is the quantity of lemonade to stock for a lemonade stand on day i, we might want to restrict $q^j \geq 0$.

4.4. 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 below:

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

Here we have included only the regularization term that encourages the jth coefficient from item $^{(1)}$ to be similar to that of item $^{(2)}$ for clarity, but other regularization terms can be included as well.

4.5. 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 would become:

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

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.

5. Comparison with Existing Data-Driven Methods

In this section, we highlight the differences between the learning algorithms of Sec. 2 to three other main data-driven methods known in the literature. We accompany the conceptual differences with examples where the existing methods may not work.

5.1. Comparison with SAA

To reiterate, the difference between the SAA approach and ours is in the data used for decision-making. In SAA, one assumes that only past demand observations are available, whereas we consider relevant features about the demand as well as the demand itself. If there is a strong relationship between the demand and some feature, the SAA approach would yield biased and inconsistent decisions, unlike (NV-algo). 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 $\{d_k | x_k = 0\} = d_k^0 + d_k^1$. 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 assume the following in order to arrive at the conclusions in this subsection.

ASSUMPTION 1. *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*

$$\sup_{0<y<1} y(1-y)\frac{|J_i(y)|}{f(F_i^{-1}(y))} \leq \gamma, \tag{7}$$

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

$$J_i(y) = \frac{-f_i'(F_i^{-1}(y))}{f_i(F_i^{-1}(y))} = -\frac{d}{dy} f_i(F_i^{-1}(y)). \tag{8}$$

Assumption 1 is satisfied by many standard distributions, and we list some in Table 1. The values of $f(F^{-1}(y))$ and $J(y)$ for the distributions in Table 1 can be found in Parzen (1979). The critical ratios for the uniform, exponential and logistic distributions follow immediately; for the normal distribution it is easier to compute the critical ratio by using the following equivalent formulation for the critical ratio:

$$\sup_{x\in \mathrm{dom}(D)} F(x)(1-F(x))\frac{|f'(x)|}{f(x)^2}. \tag{9}$$

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.

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 1 Some standard distributions that satisfy the requirement of Assumption 1. The standard normal cdf and pdf are denoted as $\Phi(\cdot)$ and $\phi(\cdot)$ respectively.

We defer all proofs of results in this subsection to Appendix A.

LEMMA 1 **(Optimal ordering decision of (NV-algo))**. *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-algo) 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 $x=1$.

PROPOSITION 1 **(Finite-sample bias and asymptotic optimality of (NV-algo))**. *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

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

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.

LEMMA 2 **(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*

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

PROPOSITION 2 **(Finite-sample bias and asymptotic (sub)-optimality of SAA)**. *With probability 1,*

$$\hat{q}_n^0 < \hat{q}_n^{SAA} < \hat{q}_n^0 + \hat{q}_n^1. \tag{10}$$

Moreover,

$$\left|\mathbb{E}[\hat{q}_n^{SAA}] - (F^{mix})^{-1}(r)\right| \leq O\left(\frac{\log n}{n}\right), \tag{11}$$

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

$$\left|\mathbb{E}[\hat{q}_n^{SAA} - \hat{q}_n^0]\right| = \left|(F^{mix})^{-1}(r) - F_0^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1)$$
$$\left|\mathbb{E}[\hat{q}_n^1 - \hat{q}_n^{SAA}]\right| = \left|F_1^{-1}(r) - (F^{mix})^{-1}(r)\right| + O\left(\frac{\log n}{n}\right) = O(1). \tag{12}$$

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,*

$$q_{opt}^0 < \lim_{n\to\infty} \hat{q}_n^{SAA} \overset{a.s.}{=} (F^{mix})^{-1}(r) < q_{opt}^1, \tag{13}$$

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

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

5.2. 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 key to this method is in the normality assumption of the residual error. 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-algo), 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 $n_0 = np_0$ are when $x=0$ and $n_1 = n - n_0$ are when $x=1$ (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 true newsvendor problem (1) under this assumption, then plug-in estimates of the conditional mean and variance of the demand to this solution. One can show, with straight-forward calculations, 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

$$q_{opt}(x) = \mu(x) + \sigma\Phi^{-1}(r). \tag{14}$$

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

$$\hat{q}_{sep}(x) = \hat{\mu}(x) + \hat{s}\Phi^{-1}(r). \tag{15}$$

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

LEMMA 3. *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 $n \to \infty$.*

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 4 **(Negative order quantity with model mis-specification).** *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.

5.3. 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 straight-forward 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

$$\hat{q}^1_{OS}(\alpha) = \alpha \bar{d}_n \tag{16}$$

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

$$\min_{\alpha \geq 0} \mathbb{E}_\theta[C(\hat{q}^1_{OS}(\alpha); D)]. \tag{17}$$

As $\alpha = \log((b+h)/b)$ is a feasible solution of (17), 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

$$\mathbb{E}_\theta[C(\hat{q}^1_{OS}(\alpha^*); D)] \leq \mathbb{E}_\theta[C(\hat{q}_{SEO}; D)], \tag{18}$$

where α^* is the optimal parameter in (17). 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

$$\hat{q}^2_{OS}(\alpha, \beta) = d_{\lceil \beta - 1 \rceil} + \alpha(d_{\lceil \beta \rceil} - d_{\lceil \beta - 1 \rceil}), \tag{19}$$

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

$$\min_{\alpha \geq 0, \beta \in \{1,\ldots,n\}} \mathbb{E}_\theta[C(\hat{q}^2_{OS}(\alpha, \beta); D)]. \tag{20}$$

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. 6, we will consider solving (NV-algo) and (NV-reg) both without and with OS-inspired features, to evaluate their practical benefit in terms of the out-of-sample cost.

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

6.1. Problem Description

In our numerical study, we consider nurse staffing in a hospital emergency room. Assuming a mandatory nurse-to-patient ratio, nurse staffing in an emergency room can be cast as a newsvendor problem in that the hospital would incur an underage cost if too many patients arrive and expensive agency nurses have to be called, and an overage cost if too many regular nurses are scheduled compared to the number of patients. As nurse staffing contributes to a significant portion of hospital operations [see Green et al. (2013) and references therein], a machine learning algorithm that can better predict staffing levels with demand uncertainty has potential for much impact.

In this section, we apply the two algorithms introduced in Sec. 2, (NV-algo) and (NV-reg), to the nurse staffing problem. Our data comes from the emergency room of a large UK teaching hospital from July 2008 to June 2009. The data include the total number of patients in the emergency room at 2-hour intervals. To get some sense of the data, we provide boxplots of the number of patients by day and by time periods in Fig. 2. We assume a nurse-to-patient ratio of 1 to 5, hence the demand is the total number of patients divided by 5. We do not require the staffing level to be an integer in our predictions, as multi-skilled workers could be used for part-time work. We also assume that the hourly wage of an agency nurse is 2.5 times that of a regular nurse, that is $b = 2.5/3.5$ and $h = 1/3.5$, resulting in a target fractile of $r = b/(b+h) = 2.5/3.5$. We consider two sets of features: the first set being the day of the week, time of the day and m number of days of past demands; the second set being the first set plus the sample average of past demands and the differences in the order statistics of past demands, which is inspired by the observation in Liyanage and Shanthikumar (2005) as described in Sec. 5.3. We use $n = 12 \times 7 \times 16 = 1344$ past data as training data and compute the critical staffing level 3 periods ahead. We then record the out-of-sample newsvendor cost of the predicted staffing level on $1344/2 = 672$ validation data on a rolling horizon basis [2].

For all numerical results, we used CVX, a package for specifying and solving convex programs [CVX Research (2012), Grant and Boyd (2008)] with the solver MOSEK.

6.2. Comparison of (NV-algo) and (NV-reg) with SAA

In Table 2 and 3, we report the ratio of the median of the out-of-sample cost of (NV-algo) and (NV-reg) to the median of the SAA newsvendor cost respectively on the same validation dataset. In parentheses we report the p-values from the Wilcoxon rank-sum test, to see whether the deviations from the SAA result are statistically significant. In Sec. 5.3, we saw that differences of order statistics of past demands can be used as features. We consider using the same algorithms, but

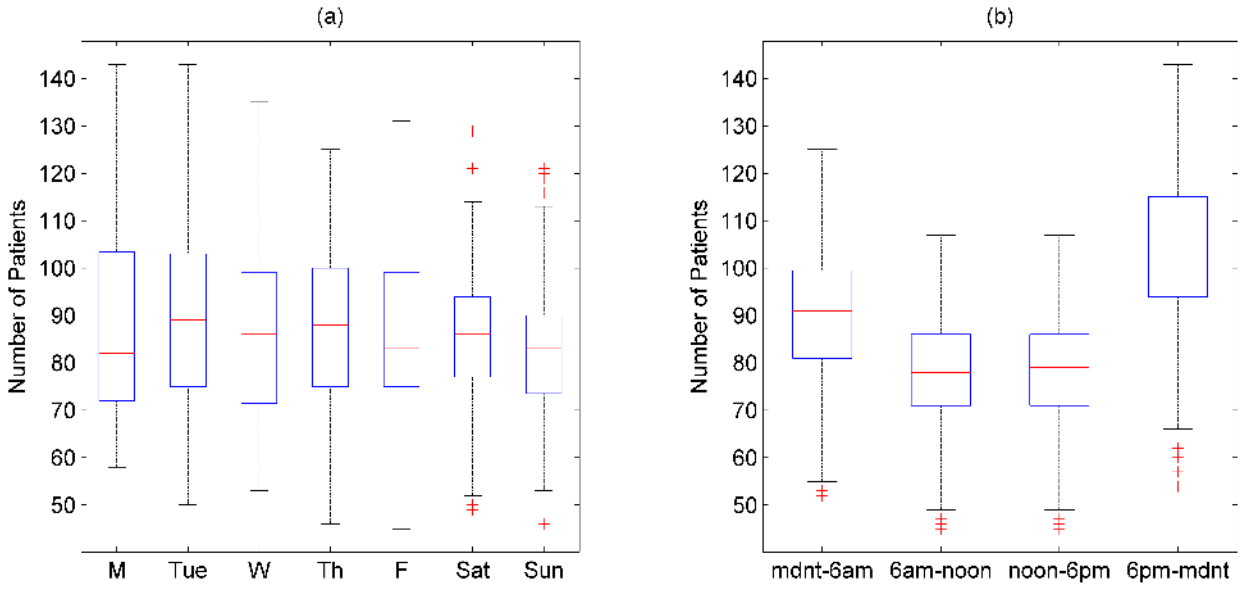

Figure 2 A boxplot of the number of patients in the emergency room (a) by day and (b) by time period.

with extra features inspired by OS, under the column marked "with OS". The best performances are highlighted in bold. We find that both our feature-based algorithms substantially reduce the out-of-sample costs with strong statistical significance (p-values are all less than 0.001). This shows the potential value of including features for medical staffing decisions.

6.3. Comparison of (NV-algo) and (NV-reg) with SEO and Minimax

In Table 4, we report the median of the out-of-sample cost of the SEO approach described in Sec. 5 and Scarf's Minimax approach [Scarf et al. (1958)]. The best performances of both methods, in bold, are also strongly statistically significantly better than the SAA method. However, both methods are worse than (NV-algo) results, which are also shown for comparative purposes, both without and with OS features. Again, this is not surprising given that (NV-algo) aims to more directly optimize the quantity that all methods are being evaluated on.

To see whether the best-performing cases of the six different methods considered in this section [(NV-algo) and (NV-reg) without and with OS features, SEO and Minimax] are statistically different from each other, we report the p-values from the Wilcoxon rank-sum test in Table 5, comparing the results pair-wise. We can see from this table that the four base-case results of solving (NV-algo) and (NV-reg) without and with OS features, are not statistically distinguishable from each other. These four results, however, are statistically different from the SEO and Minimax results at the 0.01 level.

6.4. Discussion of Predicted Staffing Decisions

Let us further investigate the staffing decision of Method 3: (NV-reg) with $\lambda = 5 \times 10^{-7}$ and no OS features. There is no loss of generality here because Table 5 ascertains that this result is statistically indistinguishable from the other best-performing results of (NV-algo) and (NV-reg).

In Fig. 3, we display the empirical cdf of the 672 out-of-sample costs of Method 3 and SAA. Remarkably, we find that the empirical cdf of the feature-based algorithm is stochastically dominated by that of SAA, i.e. the cost of the feature-based solutions are smaller than the SAA solution at every quantile of the out-of-sample distribution, which is a very strong result.

In Fig. 4 (a), we display the staffing levels predicted by Method 3 along with the actual required levels. The black dots indicate where the algorithm under-staffs with respect to the required level. Fig. 4 (b), we provide a scatter plot of the actual versus predicted staffing levels. It is apparent from both subplots that the algorithm over-predicts more than under-predicts, which reflects the asymmetry in the underage and overage costs due to agency nurses charging a higher hourly wage than regular nurses.

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 Algorithm (NV-reg). In Figs. 5 and 6, we show the conditional probability (frequency) of under- and over-prediction 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) weekdays 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 under-staffed then the period from midnight to noon, thus, given the choice of time to visit the emergency room, we would choose an early or a late morning, and (iii) the algorithm is most likely to over-staff by at least 20% of the required level on a Wednesday or a Sunday then any other day of the week (i.e. there is a middle-of-the-week effect), hence, given the flexibility, we would choose to visit the emergency room on a Wednesday or a Sunday.

6.5. Operational Recommendations

Based on the observations in this and previous sections, we derive the following guideline for effectively incorporating feature data for the newsvendor problem, which is not restricted to nurse staffing:

- Systematically record or obtain data on any information that may be associated with the demand.

• Invest in finding out what features are appropriate based on the information collected. Note nonlinear transformations of basic features (such as time of the day) can quickly enlarge the total number of features under consideration.

• Either perform some form of feature selection and use the small data newsvendor algorithm (NV-algo) if p/n is small (a good rule of thumb is 10%), otherwise automate feature-selection via the regularization-based big data algorithm (NV-reg).

• It is important to keep track of the performance of the algorithm over time, and revise information collection and feature selection. This way, one can incorporate domain expertise gained through implementation and experimentation as well as protect the system against any fundamental changes to the feature-demand dynamics.

7. 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 two tractable algorithms, one when the feature-observation ratio is small and one when the feature-observation ratio is large. For these algorithms, we derived tight generalization error bounds on the expected out-of-sample cost. We demonstrated how to modify the basic model to other realistic extensions, such as having censored demand data, having data on multiple, similar items, and introducing a new item with limited data. We further justified the feature-based approach by comparing it with other methods known in the literature. 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 several main benchmarks known in the literature.

No. of past days	without OS Features		with OS Features	
	Median Cost as % of SAA (p-value)	Total no. of Features (avg. chosen)	Median Cost as % of SAA (p-value)	Total no. of Features (avg. chosen)
0	61.90 (0.000)	12 (3.00)	61.90 (0.000)	12 (3.00)
1	63.07 (0.000)	15 (4.00)	62.73 (0.000)	27 (6.28)
2	61.90 (0.000)	27 (5.31)	67.17 (0.000)	51 (9.11)
3	61.84 (0.000)	39 (6.23)	57.24 (0.000)	75 (21.12)
4	57.53 (0.000)	51 (7.97)	57.45 (0.000)	99 (28.07)
5	**57.50 (0.000)**	**63 (8.25)**	56.63 (0.000)	123 (36.68)
6	58.29 (0.000)	75 (8.63)	58.79 (0.000)	147 (43.36)
7	57.85 (0.000)	87 (9.53)	58.53 (0.000)	171 (52.41)
8	57.63 (0.000)	99 (9.52)	**54.31 (0.000)**	**195 (56.51)**
9	57.64 (0.000)	111 (9.56)	58.19 (0.000)	219 (62.62)
10	57.70 (0.000)	123 (9.55)	61.23 (0.000)	243 (69.68)
11	57.64 (0.000)	135 (9.68)	60.05 (0.000)	267 (73.43)
12	57.64 (0.000)	147 (9.85)	66.13 (0.000)	291 (82.24)
13	57.88 (0.000)	159 (10.33)	59.22 (0.000)	315 (91.78)
14	57.67 (0.000)	171 (10.30)	60.54 (0.000)	339 (97.22)

Table 2 The median out-of-sample cost of (NV-algo) relative to SAA on the validation dataset. We use day of the week, time of the day and x number of days of past demand as features. The column marked "with OS" refer to results using differences of past demands as features, as inspired by OS. The best results without and with OS are highlighted in bold. In parentheses we report the p-values from the Wilcoxon rank-sum test to compare the result against SAA.

Regularization Param.	without OS Features		with OS Features	
	Median Cost as % of SAA (% of SAA)	Total no. of Features (avg. chosen)	Median Cost as % of SAA (% of SAA)	Total no. of Features (avg. chosen)
1×10^{-4}	67.21 (0.000)	171 (8.80)	67.09 (0.000)	339 (8.78)
5×10^{-5}	61.16 (0.000)	171 (11.31)	61.22 (0.000)	339 (11.28)
1×10^{-5}	57.52 (0.000)	171 (13.13)	57.52 (0.000)	339 (13.13)
5×10^{-6}	57.49 (0.000)	171 (13.58)	57.92 (0.000)	339 (13.96)
1×10^{-6}	56.83 (0.000)	171 (15.24)	57.05 (0.000)	339 (22.85)
$\mathbf{5 \times 10^{-7}}$	**55.04 (0.000)**	**171 (10.90)**	57.49 (0.000)	339 (35.57)
$\mathbf{1 \times 10^{-7}}$	57.39 (0.000)	171 (12.01)	**56.31 (0.000)**	**339 (107.53)**

Table 3 The median out-of-sample cost of (NV-reg) relative to SAA on the validation datset. We use day of the week, time of the day and 2 weeks of past demand as features. Without OS features, $p = 171$ and with OS features, $p = 339$. Note the average number of features chosen by the decision is calculated using the criteria that the decision element be at least 0.1% of the maximum element. The best results without and with OS are highlighted in bold. In parentheses we report the p-values from the Wilcoxon rank-sum test to compare the result against SAA.

Appendix A: For results in Sec. 5

A.1. Proofs of Main Theorems in Sec. 5

Proof. (Of Lemma 1) The feature-based algorithm (NV-algo) solves

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

	Median Cost (% of SAA)			
No. of past days	SEO	Minimax	(NV-algo) without OS	(NV-algo) with OS
1	**64.56 (0.000)**	93.15 (0.068)	63.07 (0.000)	62.73 (0.000)
2	65.15 (0.000)	94.45 (0.102)	61.90 (0.000)	67.17 (0.000)
3	65.55 (0.000)	96.74 (0.151)	61.84 (0.000)	57.24 (0.000)
4	65.69 (0.000)	94.51 (0.152)	57.53 (0.000)	57.45 (0.000)
5	67.89 (0.000)	90.87 (0.047)	**57.50 (0.000)**	56.63 (0.000)
6	67.78 (0.000)	90.52 (0.028)	58.29 (0.000)	58.79 (0.000)
7	71.93 (0.000)	89.26 (0.034)	57.85 (0.000)	58.53 (0.000)
8	71.40 (0.000)	86.28 (0.005)	57.63 (0.000)	**54.31 (0.000)**
9	70.68 (0.000)	81.66 (0.001)	57.64 (0.000)	58.19 (0.000)
10	71.95 (0.000)	79.83 (0.001)	57.70 (0.000)	61.23 (0.000)
11	72.57 (0.000)	81.72 (0.001)	57.64 (0.000)	60.05 (0.000)
12	72.39 (0.000)	**76.40 (0.000)**	57.64 (0.000)	66.13 (0.000)
13	72.74 (0.000)	76.95 (0.000)	57.88 (0.000)	59.22 (0.000)
14	73.58 (0.000)	76.93 (0.000)	57.67 (0.000)	60.54 (0.000)

Table 4 The median out-of-sample cost of SEO and Scarf's Minimax approaches relative to SAA on the validation dataset. We also report results from (NV-algo), without and with OS features, for comparison. We use day of the week, time of the day and x number of days of past demand as features. The best results are highlighted in bold. In parentheses we report the p-values from the Wilcoxon rank-sum test to compare the result against SAA.

Method	**1**	**2**	**3**	**4**	5	6
1. (NV-algo), 5 days	-	**N (0.076)**	**N (0.739)**	**N (0.186)**	Y (0.011)	Y (0.000)
2. (NV-algo), 8 days + OS	**N (0.076)**	-	**N (0.147)**	**N (0.649)**	Y (0.000)	Y (0.000)
3. (NV-reg), $\lambda = 5 \times 10^{-7}$	**N (0.739)**	**N (0.149)**	-	**N (0.337)**	Y (0.004)	Y (0.000)
4. (NV-reg), $\lambda = 1 \times 10^{-7}$ + OS	**N (0.187)**	**N (0.649)**	**N (0.337)**	-	Y (0.000)	Y (0.000)
5. SEO, 1 day	Y (0.011)	Y (0.000)	Y (0.004)	Y (0.000)	-	Y (0.008)
6. Minimax, 12 days	Y (0.000)	Y (0.000)	Y (0.000)	Y (0.000)	Y (0.008)	-

Table 5 Results from the Wilcoxon rank-sum test to compare the best-performing cases of the methods considered in this paper. Note the four results from (NV-algo) and (NV-reg) are not statistically distinguishable from each other but they are statistically different from SEO and Scarf's Minimax at the 1% level.

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

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

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). □

Proof. (Of Proposition 1) Under Assumption 1, 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

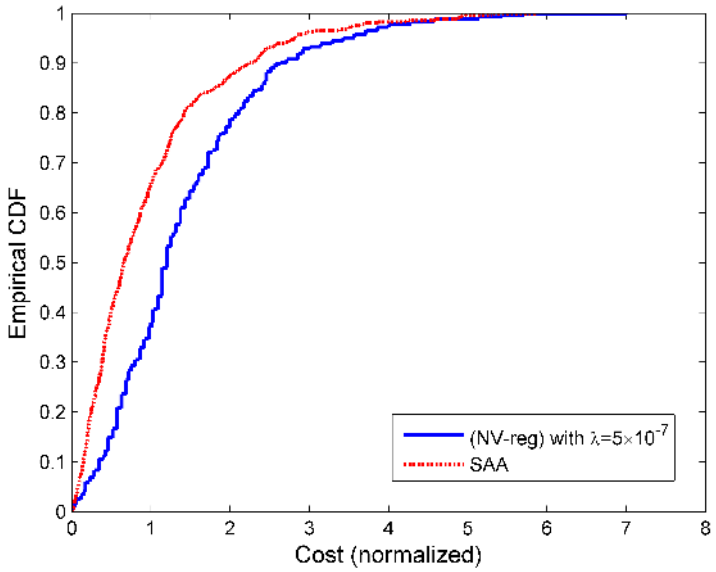

Figure 3 The empirical cdf of the best-performing case of (NV-reg) in dotted red ($\lambda = 5 \times 10^{-7}$, no OS features). The empirical cdf of the SAA solution is in solid blue. The SAA out-of-sample cost stochastically dominates that of (NV-reg).

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 $y = r$:

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

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. □

Proof. (Of Lemma 2) This is simply the SAA solution for the complete dataset. □

Proof. (Of Proposition 2) Proof of (10). 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.

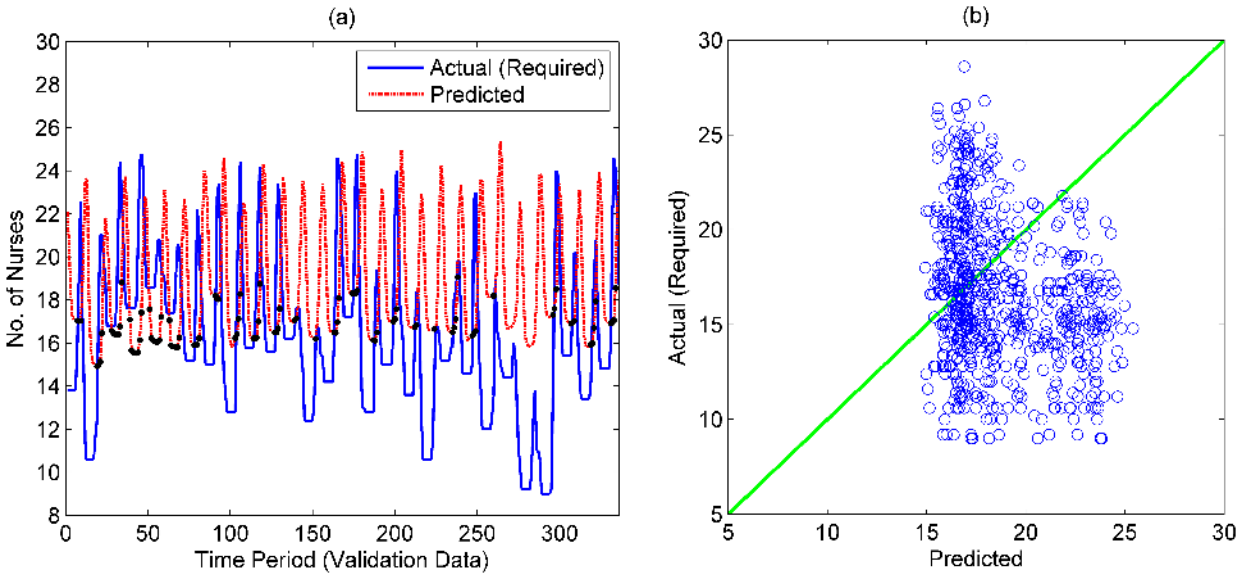

Figure 4 (a) A time-series plot of actual staffing demand (solid blue) versus staffing levels predicted by the best-performing case of (NV-reg) in dotted red ($\lambda = 5 \times 10^{-7}$, no OS features). The black circles indicate time periods where the staffing level has been under-predicted. (b) A scatter plot of actual staffing demand versus predicted by the best-performing case of (NV-reg). The green solid line shows a 1:1 relationship. It is apparent that the algorithm over-predicts more than under-predicts, which reflects the asymmetry in the underage and overage costs due to the more unfavorable consequences of being under-staffed.

Proof of (11) & (12). Proof of (11) parallels that of Proposition 1. Proof of (12) then follows from (11) and Proposition 1.

Proof of (13). 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 1. The statement then follows from 10. □

Proof. (Of Lemma 3) 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). □

Proof. (Of Lemma 4) 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

$$\hat{q}_{sep}(x) = d_0 + d_1 x + \frac{1}{\hat{\theta}} + \frac{1}{\hat{\theta}}\Phi^{-1}(r) = d_0 + d_1 x + \frac{1}{\hat{\theta}} - \frac{1}{\hat{\theta}}\Phi^{-1}(1-r), \tag{23}$$

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.

□

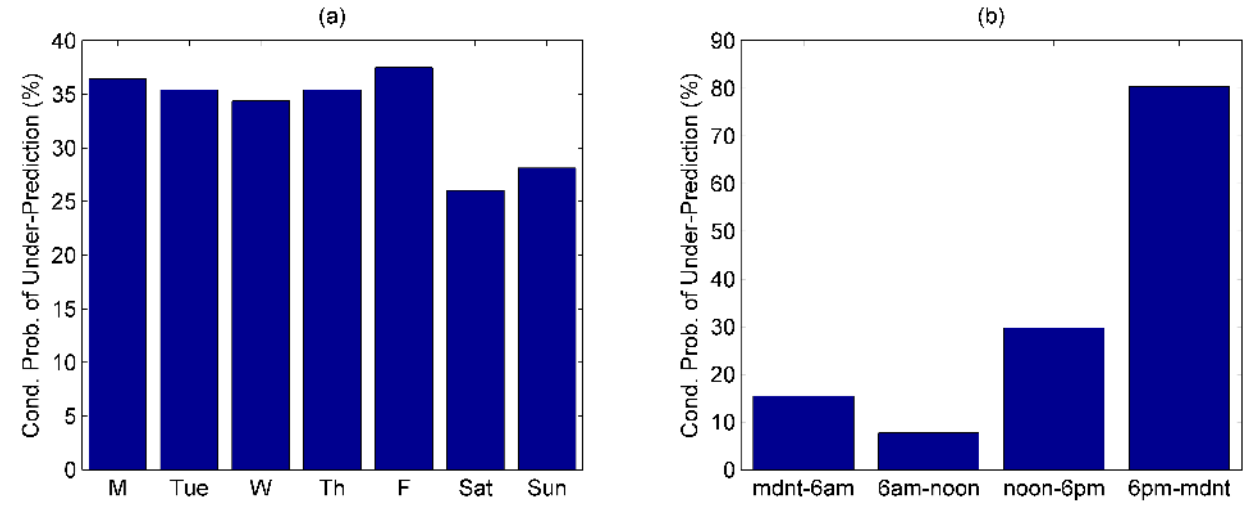

Figure 5 A plot of the conditional probabilities of under-prediction (a) by day and (b) by time period for the best-performing case of (NV-reg) ($\lambda = 5 \times 10^{-7}$, no OS features). The conditioning is done by the particular day or the time period, i.e. the probability of over-prediction given it is a Monday.

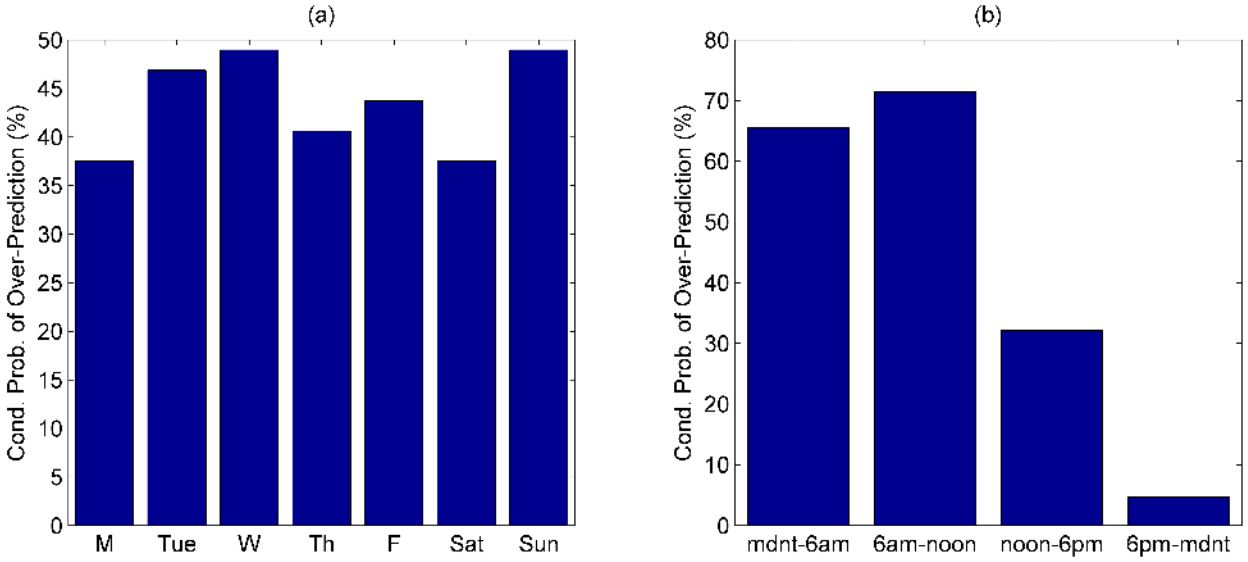

Figure 6 A plot of the conditional probabilities of over-prediction by at least 20% (a) by day and (b) by time period for the best-performing case of (NV-reg) ($\lambda = 5 \times 10^{-7}$, no OS features). The conditioning is done by the particular day or the time period, i.e. the probability of over-prediction given it is a Monday.

Appendix B: Proofs of Main Theorems in Sec. 3

We will use tools from algorithmic stability analysis to prove our results. Stability bounds were originally developed in the 1970's [Rogers and Wagner (1978), Devroye and Wagner (1979a) and Devroye and Wagner (1979b)], and was revitalized in the early 2000's Bousquet and Elisseeff (2002).

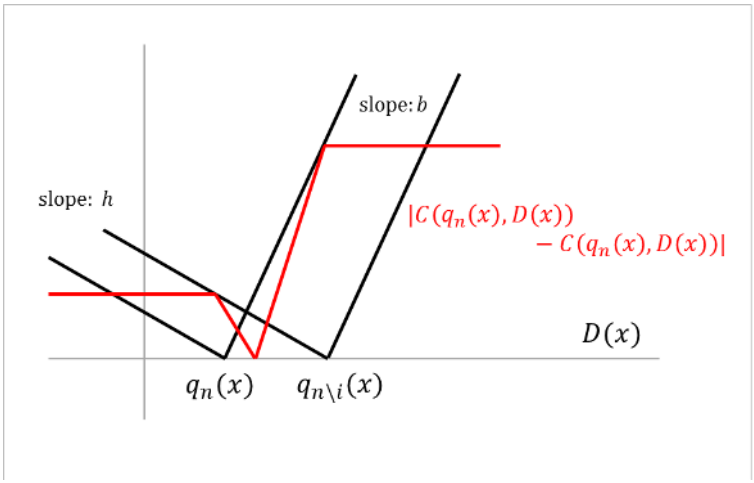

Figure 7 A plot illustrating that the difference $|C(q_n(\mathbf{x}), D(\mathbf{x})) - C(q_{n\setminus i}(\mathbf{x}), D(\mathbf{x}))|$ is bounded.

Denoting the training set by $S_n = \{z_1 = (\mathbf{x}_1, d_1), \ldots, z_n = (\mathbf{x}_n, d_n)\}$, we define the following modified training set:

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

which will be handy for the rest of the paper.

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

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

for some cost function c, which in our work will become the newsvendor cost C.

In what follows, we assume that all functions are measurable and all sets are countable. Also assume $\mathcal{Q}$ is a convex subset of a linear space. Our algorithm for the learning newsvendor problem turns out to have a very strong stability property, namely it is *uniformly stable*. In what follows we define this notion of stability and prove that the BDNV algorithm is uniformly stable in two different ways, in Theorem 3 and Theorem 4. The fact that the algorithm possesses these properties is interesting independently of other results. As we will discuss later, the proofs of Theorems 2 and 1 follow immediately from the stability properties.

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 $i \in \{1,\dots,n\}$,

$$\|\ell(A_{S_n},\cdot) - \ell(A_{S_n^{\setminus i}},\cdot)\|_\infty \leq \alpha. \tag{24}$$

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

The following will be the main result we need to prove Theorem 1.

Theorem 3 **(Uniform stability of (NV-algo))**. *The learning algorithm (NV-algo) with iid data is symmetric and uniformly stable with respect to the newsvendor cost function $C(\cdot,\cdot)$ with stability parameter*

$$\alpha_n = \frac{\bar{D}(b\vee h)^2}{(b\wedge h)}\frac{p}{n}. \tag{25}$$

We will use the following lemma in the proof of Theorem 3.

Lemma 5 **(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:*

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

Proof. (Of Lemma 5) 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 $q_n(\mathbf{x})=\bar{D}$, $|C(q_n(\mathbf{x}),d)| = \bar{D}h$. Conversely, if $d=\bar{D}$ and $q_n(\mathbf{x})=0$, $|C(q_n(\mathbf{x}),d)| = \bar{D}b$. Hence the upper bound is attained. □

Now for the proof of the theorem.

Proof. (Of Theorem 3) 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

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

where

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

is the solution to (NV-algo) for the set S_n without regularization, and

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

is the solution to (NV-algo) for the set $S_n^{\setminus i}$ without regularization. Note that:

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

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

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

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

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

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

(See Fig. 7). So we want to bound

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

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). Further, because $0 \in \partial\hat{R}(q_n; S_n)$ by the optimality of q_n, we have

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

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

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

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

$$|q_{n\setminus i}^j - q_n^j| \le \frac{\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n)}{|\nu_j^*|}.$$

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

$$\begin{aligned}\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)|q_n(\mathbf{x}) - q_{n\setminus i}(\mathbf{x})| &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)\left(\sum_{j=1}^p |q_n^j - q_{n\setminus i}^j||x_j|\right)\\ &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)\cdot\sum_{j=1}^p \frac{|x_j|}{|\nu_j^*|}\cdot(\hat{R}(\mathbf{q}_{n\setminus i}; S_n) - \hat{R}(\mathbf{q}_n; S_n))\\ &= \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{b\vee h}{n}\cdot\sum_{j=1}^p \frac{|x_j|}{|\nu_j^*|}\cdot|\hat{R}(\mathbf{q}_{n\setminus i}; S_i)|. \qquad (27)\end{aligned}$$

We can further simplify the upper bound (27) as follows. Recall that $\boldsymbol{\nu}^*$ is the subgradient of $\hat{R}(\cdot; S_n)$ at $\mathbf{q}_n$ that maximizes $\sum_{j=1}^p \nu_j(q_{n\setminus i}^j - q_n^j)$; 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:

$$\nu_j^* = \begin{cases} -bx_j & \text{if } q_{n\setminus i}^j - q_n^j \leq 0\\ hx_j & \text{if } q_{n\setminus i}^j - q_n^j \geq 0 \quad \forall\ j.\end{cases}$$

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 5, we get the desired result. □

We move on to the main result needed to prove Theorem 2.

THEOREM 4 **(Uniform stability of NV-reg)**. *The learning algorithm (NV-reg) is symmetric, and is uniformly stable with respect to the NV cost function C with stability parameter*

$$\alpha_n^r = \frac{(b\vee h)^2}{2X_{\max}^{-2}}\frac{1}{n\lambda}. \qquad (28)$$

Let us build some terminology for the proof of Theorem 4.

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:

$$\forall\ y_1, y_2 \in \mathcal{Y},\ \forall\ d \in \mathcal{D},\ |c(y_1,d) - c(y_2,d)| \le \sigma|q_1 - q_2|, \tag{29}$$

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 5 **(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*

$$A_{S_n} = \operatorname*{arg\,min}_{g\in\mathcal{F}}\ \frac{1}{n}\sum_{i=1}^{n}\ell(g, z_i) + \lambda\|g\|_k^2 \tag{30}$$

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 Theorem 4) By the Lipschitz property of $C(\cdot; d)$,

$$\sup_{d\in\mathcal{D}}\ |C(q_1(\mathbf{x}), d) - C(q_2(\mathbf{x}), d)| \le (b \vee h)|q_1(\mathbf{x}) - q_2(\mathbf{x})|,\quad \forall q_1(\mathbf{x}), q_2(\mathbf{x}) \in \mathcal{Q} \tag{31}$$

as before, hence $C : \mathcal{X}\times\mathcal{D} \to \mathbb{R}$ is $(b\vee h)$-admissible. Hence by Theorem 5 the algorithm (NV-reg) has uniform stability with parameter α_n^r as given. $\square$

We have thus far established the stability of the big-data newsvendor algorithms (NV-algo) and (NV-reg), which lead immediately to the risk bounds provided in Theorem 2 and Theorem 1 following the established theorems relating stability to generalization, as follows.

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

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

THEOREM 6. *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:*

$$|R_{true}(A, S_n) - \hat{R}(A, S_n)| \le 2\alpha_n + (4n\alpha_n + M)\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{32}$$

Proof. (Of Theorem 6) 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 Theorems 1 and 2.

Proof. (Of Theorem 1) The result follows from Theorems 3 and 6. □

Proof. (Of Theorem 2) By Lemma 5, $0 \leq \ell(A_S, z) \leq \bar{D}(b \vee h)$ for all $z \in \mathcal{Z}$ and all sets S. The result then follows from Theorems 4 and 6. □

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 Victor DeMiguel, Nicos Savva 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.

Ben-Tal, Aharon, Boaz Golany, Arkadi Nemirovski, Jean-Philippe Vial. 2005. Retailer-supplier flexible commitments contracts: a robust optimization approach. *Manufacturing & Service Operations Management* **7**(3) 248–271.

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

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.

Chen, Xin, Melvyn Sim, Peng Sun. 2007. A robust optimization perspective on stochastic programming. *Operations Research* **55**(6) 1058–1071.

Chen, Xin, Melvyn Sim, Peng Sun, Jiawei Zhang. 2008. A linear decision-based approximation approach to stochastic programming. *Operations Research* **56**(2) 344–357.

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.

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* .

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.

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.

Yuan, Ming, Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1) 49–67.