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

The Big Data Newsvendor: Practical Insights from Machine Learning Analysis

Cynthia Rudin
Sloan School of Management, 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 present a version of the newsvendor problem where one has n observations of p features as well as past demand. We consider both "big data" ($p/n = O(1)$) as well as small data ($p/n = o(1)$). For small data, we provide a linear programming machine learning algorithm that yields an asymptotically optimal order quantity. We also derive a generalization bound based on algorithmic stability, which is an upper bound on the expected out-of-sample cost. For big data, we propose a regularized version of the algorithm to address the curse of dimensionality. A generalization bound is derived for this case as well, bounding the out-of-sample cost with a quantity that depends on n and the amount of regularization. We apply the algorithm to analyze the newsvendor cost of nurse staffing using data from the emergency room of a large teaching hospital and show that (i) incorporating appropriate features can reduce the out-of-sample cost by up to 23% relative to the featureless Sample Average Approximation approach, and (ii) regularization can automate feature-selection while controlling the out-of-sample cost. By an appropriate choice of the newsvendor underage and overage costs, our results also apply to quantile regression.

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

History: October 7, 2013

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 distribution of the demand. Realistically, one might instead have demand data from the past, including data about many features that are potentially associated with the demand. In this paper, we investigate the newsvendor problem when one has access to past demand observations as well as a 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. Relevant features could be derived from the seasonality (day, month, seasons), weather forecasts and various economic indicators. With the current interest in big data analytics, many organizations are starting to systematically collect such information, and this work 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. It is possible to have *big data*; that is, $p/n = O(1)$. We consider learning from these past data: given a new period a new set of features, the decision maker chooses an appropriate order quantity.

In the classical newsvendor problem, 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 direction 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. They derive a sample size bound; that is, a calculation of the minimal number of observations required in order for the estimated solution to be near-optimal with high probability. In this paper, we extend Levi et al. (2007), who did not consider features, by deriving a bound on the out-of-sample cost of the newsvendor problem for $p \neq 0$ (NB: we retrieve the bound of Levi et al. (2007) if p is set to zero). As far as we are aware, this paper is the first to derive insights about the data-driven newsvendor problem when features are involved.

Other perspectives on the data-driven newsvendor include Liyanage and Shanthikumar (2005), who showed that integrating estimation and optimization ("operational statistics") perform better than separating them, Huh et al. (2011) and Besbes and Muharremoglu (2013) who provide theoretical insights into the newsvendor problem with censored demand data, and Levi et al. (2011), who improve upon the bound of Levi et al. (2007) by incorporating more information about the 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 over distributions with the same mean and standard deviation. Perakis and Roels (2008) consider a minimax regret approach for the newsvendor with partial information about the demand distribution. None of the above mentioned works incorporate feature information.

Summary of contributions

(i) We introduce the newsvendor problem where 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 that can be used broadly for both iid as well as time-dependent data. This algorithm is based on the empirical risk minimization principle [for an in-depth discussion of this approach, see Vapnik (1998)].

(ii) We provide tight probabilistic bounds on the out-of-sample cost of the data-driven newsvendor with feature information. 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 is bounded by the in-sample cost and a complexity term that scales as $1/\sqrt{n}$. The first bound is useful for the low p setting, where regularization is not necessary, and the second bound handles the high dimensional setting. The first bound depends on p and reflects problems with the curse of dimensionality when p is too large and regularization is not used. Consequently, in order to reduce the out-of-sample cost, which is what the decision maker really cares about, s/he must carefully choose the features to use in order to prevent overfitting (having too many features and a large generalization error) with respect to the amount of training data when regularization is not used. This bound reflects that regularization becomes necessary in higher dimensions, motivating the second bound. For high p, our second bound scales inversely with the regularization parameter, and does not depend explicitly on p.

(iii) We demonstrate that our LP algorithm can be effective for nurse staffing in a hospital emergency room. In particular, we show that the nurse staffing costs in the emergency room of a large teaching hospital in the United Kingdom can be reduced by up to 23% compared to the featureless SAA approach by appropriately incorporating high-dimensional feature data. This setting is similar to He et al. (2012), who consider staffing in a hospital operating room (OR) with two features (number and type of cases) to predict the required OR time. Our investigation is different, however, because we are interested in the effect of high-dimensionality of the training data on the out-of-sample cost.

Before proceeding, we also mention that by a certain choice of the underage and overage costs, our feature-based newsvendor algorithm reduces to nonparametric quantile regression. The results in this paper thus also extend to nonparametric quantile regression [see Koenker (2005) for a general reference, Takeuchi et al. (2006) and Steinwart and Christmann (2011) for up-to-date results at the time of writing].

2. Problem setup

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 and does not depend on covariates, one can show the optimal decision is given by

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

The Data-Driven Newsvendor Problem In reality, the decision maker does not know the true distribution. Again assume the demand cannot be predicted from external covariates. 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 an 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 [Shapiro et al. (2009)].

The Big Data Newsvendor Problem The data-driven newsvendor problem is too simplistic to hold in many real situations; in reality, 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 Big Data Newsvendor (BDNV) problem is:

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

where $\hat{R}$ is called the *empirical risk* of function q with respect to dataset S_n, and this algorithm is an *empirical risk minimization* algorithm.

To solve (BDNV), one needs to specify the function class $\mathcal{Q}$. The size or "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 more or less complex depending on which transformations are included. We can solve (BDNV) via the following linear program:

NV Algorithm with Features

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

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

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

In what follows we will provide two probabilistic bounds on the estimated sample cost. Because we formulated the BDNV algorithm to fit into the framework of empirical risk minimization, we are able to use modern tools from machine learning. In particular, we will use *algorithmic stability* analysis which measures how much an algorithm's predictions change when one of the training examples is removed. If the algorithm is robust to the change in the sample, it tends to generalize better to new observations.

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 an upper bound on the true risk.

In what follows, the random demand is denoted by D, and is assumed to be bounded: $D \in \mathcal{D} := [0, \bar{D}]$. The 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 $S_n = \{(\mathbf{x}_i, d_i)\}_{i=1}^n$.

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 i.i.d. from an unknown distribution on $\mathcal{X} \times \mathcal{D}$:*

$$R_{true}(\hat{q}) \leq \hat{R}(\hat{q}; S_n) + (b \vee h)\bar{D}\left[\frac{2(b \vee h)}{b \wedge h}\frac{p}{n} + \left(\frac{4(b \vee h)}{b \wedge h}p + 1\right)\sqrt{\frac{\ln(1/\delta)}{2n}}\right].$$

This bound can be used for small p as it scales nicely as $1/\sqrt{n}$. However it suggests that the generalization error of the newsvendor cost could scale as $O(p/\sqrt{n})$, meaning that if the number of features is large relative to the number of observations, the previous bound would give a vacuous result. In the case of big data, i.e., when the ratio of the number of features to observations $p/n = O(1)$, we suggest instead solving the following regularized version of (NV-algo):

NV Algorithm with Regularization

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

where $\lambda > 0$ is the regularization parameter and $||\mathbf{q}||_2$ denotes the L_2-norm of the vector $\mathbf{q} = [q^1, \ldots, q^p]$. The problem is a quadratic program, which can be solved efficiently using widely available conic programming solvers.

We present the following generalization bound for (NV-reg).

THEOREM 2 **(Generalization Bound for (NV-reg))**. *Define* $X^2_{\max}$ *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 i.i.d. from an unknown distribution on* $\mathcal{X}\times\mathcal{D}$*:*

$$R_{true}(\hat{q}) \le \hat{R}(\hat{q};S_n) + (b\vee h)\left[\frac{(b\vee h)X^2_{\max}}{n\lambda} + \left(\frac{2(b\vee h)X^2_{\max}}{\lambda} + \bar{D}\right)\sqrt{\frac{\ln(1/\delta)}{2n}}\right]. \tag{4}$$

This bound does not depend explicitly on p, indicating that Algorithm (NV-reg) can gracefully 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 $\lambda \le O(1/p)$, for example $\lambda = O(1/p^2)$. Note additionally that λ should be chosen relative to $X^2_{\max}$.

Last, both bounds of Theorem 1 and 2 scale appropriately with δ, as $\mathcal{O}(\sqrt{\ln(1/\delta)})$.

The rest of this section is devoted to the proofs of the main theorems.

3.1. Proofs of Main Theorems

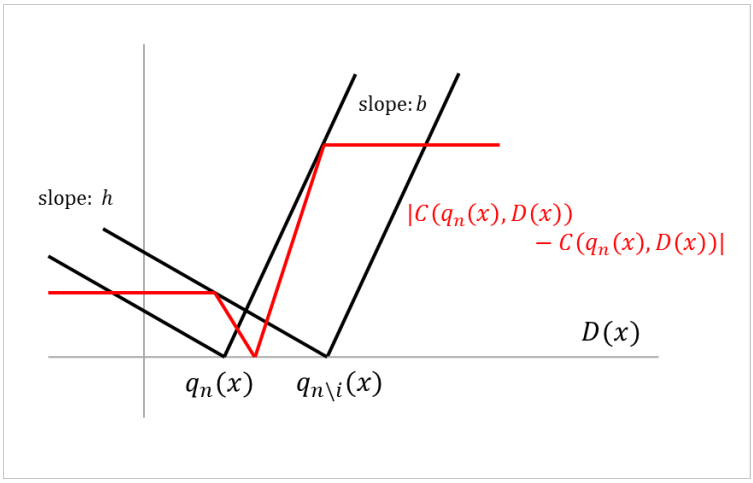

Figure 1 **A plot illustrating that the difference** $|C(q_n(\mathbf{x}),D(\mathbf{x})) - C(q_{n\setminus i}(\mathbf{x}),D(\mathbf{x}))|$ **is bounded.**

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).

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

$$\|\ell(A_{S_n}, \cdot) - \ell(A_{S_n^{\setminus i}}, \cdot)\|_\infty \leq \alpha. \tag{5}$$

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 i.i.d. data is symmetric and uniformly stable with respect to the newsvendor cost function $C(\cdot, \cdot)$ with stability parameter*

$$\alpha_n = \frac{\bar{D}(b \vee h)^2}{(b \wedge h)} \frac{p}{n}. \tag{6}$$

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

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:*

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

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 $q_n(\mathbf{x})=\bar{D}$, $|C(q_n(\mathbf{x}),d)|=\bar{D}h$. 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 i.i.d.. 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

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

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

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

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. 1). 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)):

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

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

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

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

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

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 $j = 1, \ldots, p$,

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

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

$$\begin{aligned}\sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)|q_n(\mathbf{x}) - q_{n\setminus i}(\mathbf{x})| &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)\left(\sum_{j=1}^{p}|q_n^j - q_{n\setminus i}^j||x_j|\right)\\ &\leq \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} (b\vee h)\cdot\sum_{j=1}^{p}\frac{|x_j|}{|\nu_j^*|}\cdot(\hat{R}(\mathbf{q}_{n\setminus i};S_n) - \hat{R}(\mathbf{q}_n;S_n))\\ &= \sup_{(\mathbf{x},D)\in\mathcal{X}\times\mathcal{D}} \frac{b\vee h}{n}\cdot\sum_{j=1}^{p}\frac{|x_j|}{|\nu_j^*|}\cdot|\hat{R}(\mathbf{q}_{n\setminus i};S_i)|. \qquad (8)\end{aligned}$$

We can further simplify the upper bound (8) 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 1, 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*

$$\alpha_n^r = \frac{(b\vee h)^2}{n}\frac{X_{\max}^2}{2\lambda}. \qquad (9)$$

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:

$$\forall\ y_1,y_2\in\mathcal{Y},\ \forall\ d\in\mathcal{D},\ |c(y_1,d) - c(y_2,d)| \leq \sigma|q_1 - q_2|, \qquad (10)$$

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 $\forall\ x\in\mathcal{X}$, $k(x,x)\leq\kappa^2<\infty$. Let ℓ be σ-admissible with respect to $\mathcal{F}$. The learning algorithm A defined by*

$$A_{S_n} = \arg\min_{g\in\mathcal{F}}\ \frac{1}{n}\sum_{i=1}^{n}\ell(g,z_i) + \lambda||g||_k^2 \qquad (11)$$

has uniform stability α_n wrt ℓ with

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

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 $C(\cdot; d)$,

$$\sup_{d\in\mathcal{D}} |C(q_1(\mathbf{x}), d) - C(q_2(\mathbf{x}), d)| \le (b \vee h)|q_1(\mathbf{x}) - q_2(\mathbf{x})|, \quad \forall q_1(\mathbf{x}), q_2(\mathbf{x}) \in \mathcal{Q} \tag{12}$$

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. □

We have thus far established the stability of the big-data newsvendor algorithm (NV-algo) for low- and high-dimensional problems, 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 **(Bousquet and Elisseeff (2002) Theorem 12 pp 507)**. *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:*

$$R_{true}(A, S_n) \le \hat{R}(A, S_n) + 2\alpha_n + (4n\alpha_n + M)\sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{13}$$

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 Theorems 4 and 6. □

4. Case study: nurse staffing in a hospital emergency room

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 if too many patients arrive, expensive agency nurses have to be called, incurring an underage cost, whereas if there are not many patients, regular nurses sit idle, incurring an overage cost. As nurse staffing contributes to a significant portion of hospital operations [see Green et al. (2013) and references therein], a sophisticated machine learning algorithm that can better predict the newsvendor critical minimum fractile and cost has potential for much impact.

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. We assume the hourly wage of an agency nurse is 2.5 times that of a regular nurse, that

No. of past days inc.	Total no. of features (p)	Average Cost (% of SAA)	Improvement stat. sig.?
0	2	96.17	No
2	26	93.90	No
4	50	93.40	No
6	74	91.29	Yes
8	98	86.66	Yes
10	122	85.67	Yes
12	146	84.82	Yes
14	170	83.58	Yes
16	194	81.40	Yes
18	218	77.74	Yes
20	242	77.06	Yes
22	266	76.58	Yes
24	290	76.78	Yes

Table 1 The out-of-sample cost (NV-algo) with increasing number of features (past demands) relative to featureless SAA approach on validation dataset.

is $b = 2.5/3.5$ and $h = 1/3.5$, yielding the target fractile $b/(b+h) = 2.5/3.5$. We use as features the day of the week, time of the day and m number of past demands. Our aim is to investigate whether incorporating features can improve upon the featureless SAA approach, and if so, the magnitude of improvement. We use $n = 12 \times 7 \times 16 = 1344$ observations as training data and compute the 3-period ahead out-of-sample newsvendor cost on $1344/2 = 672$ validation data on a rolling horizon basis (note we use the rule of thumb suggested by Friedman et al. (2009) in choosing the size of the validation data). In Table 1, we report the ratio of the average of the feature-based newsvendor cost to the average of the SAA newsvendor cost on the same validation dataset and indicate whether the improvement is statistically significant at the 5% level using the Wilcoxon rank-sum test.

In Fig. (2), we plot the empirical cumulative distribution function (ecdf) of the SAA out-of-sample costs versus the (NV-algo) on the validation dataset. Remarkably, not only does incorporating features improve upon the average out-of-sample cost, we find that the ecdf of (NV-algo) is stochatically dominated by that of SAA, i.e. the cost of (NV-algo) is smaller than SAA at every single percentile of the out-of-sample distribution, which is a very strong result. We also remark that the phenomenon of overfitting seems to kick in beyond $p = 266$, as evidenced by the increase in the average out-of-sample cost from $p = 266$ to $p = 290$.

5. 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 simple algorithms, one when the number of features is small and one when the number of features is large ("big data"), and characterized their expected out-of-sample cost performance analytically, leading to practical insights. Finally, we investigated nurse staffing in a

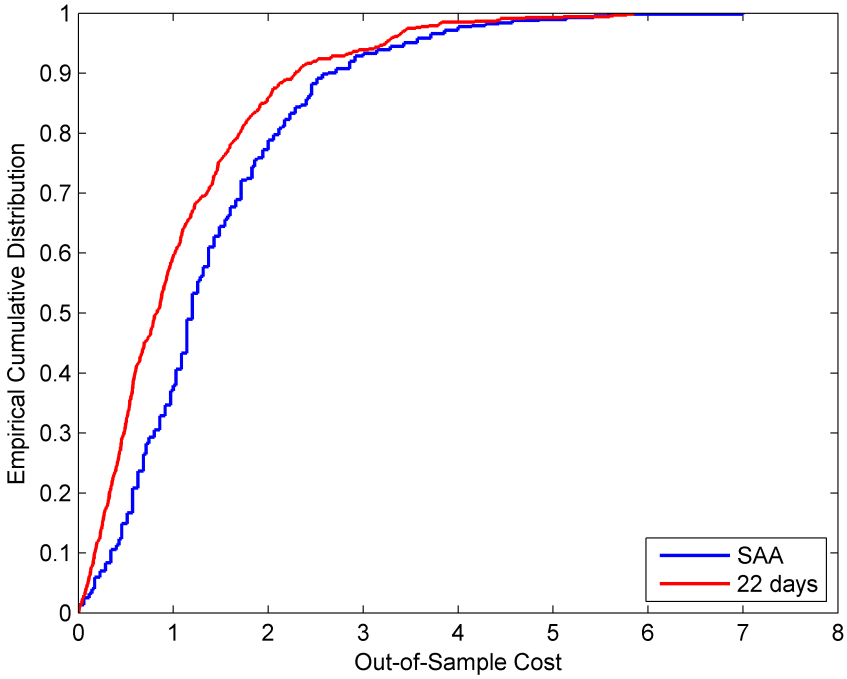

Figure 2 **The empirical cumulative distribution function of the cost of the featureless SAA approach and the feature-based (NV-algo) approach (which uses patient numbers from 22 previous days as features) on the validation dataset.**

hospital emergency room using a real dataset from a large teachng hospital in the United Kingdom and showed that incorporating appropriate features can reduce the average out-of-sample cost by up to 23% relative to the featureless SAA approach.

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 for providing the hospital emergency room data and feedback.

References

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.

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.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani. 2009. The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics* .

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

Green, Linda V, Sergei Savin, Nicos Savva. 2013. nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science* .

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.

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. 2011. The data-driven newsvendor problem: new bounds and insights. *Submitted to Operations Research, second revision was requested* .

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.

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.

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 .