Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Big Data Newsvendor: Practical Insights from Machine Learning

Gah-Yi Ban, Cynthia Rudin

To cite this article:
Gah-Yi Ban, Cynthia Rudin (2019) The Big Data Newsvendor: Practical Insights from Machine Learning. Operations Research 67(1):90-108. https://doi.org/10.1287/opre.2018.1757

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 67, No. 1, January–February 2019, pp. 90–108
ISSN 0030-364X (print), ISSN 1526-5463 (online)

The Big Data Newsvendor: Practical Insights from Machine Learning

Gah-Yi Ban,[a] **Cynthia Rudin**[b]

[a] Management Science & Operations, London Business School, London NW1 4SA, United Kingdom; [b] Department of Computer Science, Department of Electrical and Computer Engineering, and Statistical Science, Duke University, Durham, North Carolina 27708

Contact: gban@london.edu, http://orcid.org/0000-0003-3229-1386 (G-YB); cynthia@cs.duke.edu, http://orcid.org/0000-0003-4283-2780 (CR)

Received: February 1, 2015
Revised: August 7, 2016; November 2, 2017
Accepted: March 2, 2018
Published Online in Articles in Advance: November 7, 2018

Subject Classifications: inventory/production; stochastic: programming; stochastic: statistics; estimation
Area of Review: Operations and Supply Chains

https://doi.org/10.1287/opre.2018.1757

Abstract. We investigate the data-driven newsvendor problem when one has n observations of p features related to the demand as well as historical demand data. Rather than a two-step process of first estimating a demand distribution then optimizing for the optimal order quantity, we propose solving the "big data" newsvendor problem via single-step machine-learning algorithms. Specifically, we propose algorithms based on the empirical risk minimization (ERM) principle, with and without regularization, and an algorithm based on kernel-weights optimization (KO). The ERM approaches, equivalent to high-dimensional quantile regression, can be solved by convex optimization problems and the KO approach by a sorting algorithm. We analytically justify the use of features by showing that their omission yields inconsistent decisions. We then derive finite-sample performance bounds on the out-of-sample costs of the feature-based algorithms, which quantify the effects of dimensionality and cost parameters. Our bounds, based on algorithmic stability theory, generalize known analyses for the newsvendor problem without feature information. Finally, we apply the feature-based algorithms for nurse staffing in a hospital emergency room using a data set from a large UK teaching hospital and find that (1) the best ERM and KO algorithms beat the best practice benchmark by 23% and 24%, respectively, in the out-of-sample cost, and (2) the best KO algorithm is faster than the best ERM algorithm by three orders of magnitude and the best practice benchmark by two orders of magnitude.

Funding: This research was supported by National Science Foundation Grant IIS-1053407 (C. Rudin) and the London Business School Research and Material Development Scheme (G.-Y. Ban).
Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2018.1757.

Keywords: big data • newsvendor • machine learning • sample average approximation • statistical learning theory • quantile regression

1. Introduction

We investigate the newsvendor problem when one has access to n past demand observations as well as a potentially large number, p, of *features* about the demand. By features, we mean exogenous variables (also known as covariates, attributes, and explanatory variables) that are predictors of the demand and are available to the decision maker (hereafter, "DM") before the ordering occurs. Although inventory models to date have typically been constructed with demand as the stochastic primitive, in the world of big data, the DM has access to a potentially large amount of relevant information, such as customer demographics, weather forecasts, seasonality (e.g., day of the week, month of the year, and season), and economic indicators (e.g., the consumer price index) as well as past demands to inform the DM's ordering decisions.

In this paper, we propose solving the "big data" newsvendor problem via distribution-free, one-step machine-learning algorithms that handle high-dimensional feature data and derive finite-sample performance bounds on their out-of-sample costs. The one-step algorithms contrast with the approach of solving the big data newsvendor problem via a two-step process of first estimating a (feature-dependent) demand distribution, then optimizing for the optimal order quantity. Such two-step processes can be problematic because demand model specification is difficult in high dimensions, and errors in the first step will amplify in the optimization step.

In this setting, the paper is structured to answer the following four questions: (Q1) How should the DM use a feature-demand data set to solve the newsvendor problem? (Q2) What is the value of incorporating features in newsvendor decision making in the first place? (Q3) What theoretical guarantees does the DM using such data have, and how do these scale with the various problem parameters? (Q4) How do newsvendor decisions based on the feature-demand data set compare with other benchmarks in practice?

We address (Q1) in Section 2, in which we propose one-step approaches to finding the optimal order quantity with a data set of both demand and related feature observations. One approach is based on the machine-learning principle of empirical risk minimization (ERM) (Vapnik 1998), with and without regularization, and the other, which we call kernel-weights optimization (KO), is inspired by the Nadaraya–Watson kernel regression method (Nadaraya 1964, Watson 1964). The ERM approach, equivalent to high-dimensional quantile regression, is a linear programming (LP) algorithm without regularization and a mixed-integer program (MIP), an LP, or a quadratic program (QP) with ℓ_0, ℓ_1 and ℓ_2 regularizations, respectively. Under the KO approach, the optimal data-driven order quantity can be found by a simple sorting algorithm.

We justify the use of features by answering (Q2) in Section 3, in which we analytically quantify the value of features by comparing against decisions made without any features (the "sample average approximation (SAA)" method). This is necessary as most data-driven inventory works to date do not consider features. We consider two demand models for the comparison: a two-population model and the linear demand model. For both models, we show that the no-feature SAA decision does not converge to the true optimal decision whereas the feature-based ERM decision does. In other words, the SAA decision can have a constant bias (i.e., $O(1)$ error) regardless of how many observations n the DM has whereas any finite-sample bias of the feature-based decision shrinks to zero as n tends to infinity. Accordingly, the SAA decision can have a higher newsvendor cost than the ERM decision. We quantify the additional cost incurred by biased decisions in Theorem 4.

We address (Q3) in Section 4. In practice, the DM may never make the truly optimal decision with a finite amount of data even if the DM uses as much relevant feature information as possible. To understand the trade-offs of not having the full distributional information of the demand, we derive theoretical performance bounds for the DM who uses the algorithms proposed in Section 2. Our bounds characterize how the in-sample cost of the in-sample decision (which the DM can calculate as opposed to the expected cost of the in-sample decision) deviates from the true expected cost of the true optimal decision in terms of the various parameters of the problem, in particular in terms of p and n. The bounds show how the out-of-sample cost of the in-sample decision (the "generalization error") deviates from its in-sample cost by a complexity term that scales gracefully as $1/\sqrt{n}$ for n and as $\sqrt{\log(1/\delta)}$ for δ, where $1-\delta$ is the probabilistic accuracy of our bound under the minimal assumption that the demand data are independent and identically distributed (iid) in the high-dimensional feature space. The bounds also show how the finite-sample bias from the true optimal decision scales as $n^{-1/(2+p/2)}\sqrt{\log n}$ under the additional assumption of the linear demand model.

From a practical perspective, our bounds explicitly show the trade-offs between the generalization error (which measures the degree of in-sample overfitting) and the finite-sample bias and how they depend on the size of the data set, the newsvendor cost parameters, and any free parameters (controls) in the algorithms. Although some past papers in operations management have incorporated specific features in inventory models, to the best of our knowledge, none have analyzed the out-of-sample performance (cost) with high-dimensional data. Our work also contrasts with past works in quantile regression by using the algorithm-specific stability theory of Bousquet and Elisseeff (2002) to analyze the generalization error as opposed to the Vapnik–Chervonenkis (VC) theory (Vapnik 1998), which results in bounds with tight constants dependent only on the newsvendor cost parameters as opposed to ones dependent on uniform complexity measures, such as covering numbers, VC dimension, and Rademacher averages, which are often difficult to compute and interpret in practice. Detailed discussions of the literature can be found in Section 1.1.

We address (Q4) in Section 5, in which we evaluate our algorithms against other known benchmarks through an extensive empirical investigation. Specifically, we apply our algorithms and other methods to a nurse-staffing problem in a hospital emergency room. The best result using the ERM approach was with ℓ_1 regularization with a cost improvement of 23% (a saving of about £44,200 per annum (p.a.)) relative to the best practice benchmark (sample average approximation clustered by day of the week), and the best result using the KO approach had a cost improvement of 24% (a saving of about £46,600 p.a.) relative to the same benchmark. Both results were statistically significant at the 5% level. The best KO method, solved using the sorting procedure described in Section 2, was also very computationally efficient, taking just 0.05 seconds to compute the optimal staffing level for the next period, which is three orders of magnitude faster than the best ERM method and two orders of magnitude faster than the best practice benchmark and other benchmarks.

Finally, in Section 6, we conclude with a discussion of the practical takeaways and generalizable insights as well as limitations of our investigation and thoughts on future directions for research.

1.1. Literature Review

Our work contributes to the following areas of investigation in operations management and machine learning.

(1) *Data-driven inventory models.* Models in newsvendor/inventory management have long been constructed

with the available (or, rather, the lack of) data in mind with the stochastic nature of the demand modeled with various assumptions. The earliest papers (Arrow et al. 1958, Scarf 1959b) make the assumption that the demand distribution is fully known, and this has been relaxed in recent years. Overall, there have been three main approaches to modeling demand uncertainty in the literature. In the Bayesian approach, the demand distribution is assumed to be known up to unknown parameter(s), which are dynamically learned starting with prior assumptions (Scarf 1959a, Azoury 1985, Lovejoy 1990). In the minimax approach, the decision maker opts for the best robust decision among all demand distributions within a specified uncertainty set (Scarf et al. 1958, Gallego and Moon 1993, Chen et al. 2007, Perakis and Roels 2008). Finally, in the data-driven approach, the DM has access to samples of demand observations drawn from an unknown distribution. In this setting, Burnetas and Smith (2000), Kunnumkal and Topaloglu (2008), and Huh and Rusmevichientong (2009) propose stochastic gradient algorithms; Godfrey and Powell (2001) and Powell et al. (2004) consider the adaptive value estimation method; Levi et al. (2007) provide a sampling-based method based on solving a shadow problem to solve for the optimal ordering quantities; and Levi et al. (2015) improve upon the bounds of Levi et al. (2007) for the single-period, featureless newsvendor case.

Within this line of work, the distinguishing aspect of our paper is in the incorporation and analysis of a potentially large number of features directly in an inventory model. As far as we are aware, the works of Liyanage and Shanthikumar (2005), Hannah et al. (2010), and See and Sim (2010) are the only other preceding methodological papers in operations management to incorporate some form of feature information in the decision model. We compare them with our approaches in detail in Section 2.4. Most importantly, we derive performance bounds that quantify the effect of the feature dimension on the out-of-sample cost, which has no precedence in this line of literature.

Our setup and subsequent analysis also differ from the parametric modeling approaches of Feldman (1978), Lovejoy (1992), Song and Zipkin (1993), Gallego and Özer (2001), Lu et al. (2006), and Iida and Zipkin (2006), in which the demand is modeled as Markov-modulated processes with known state-dependent distributions, in which the states capture various exogenous information, such as economic indicators or advanced demand information. The key difference is that we make minimal assumptions about the underlying demand distribution (iid for the most part and, later, for more specific finite-sample bias analysis, the linear demand model with unknown error distribution) whereas in all of the aforementioned works, the demand is modeled parametrically with known, state-dependent distributions (e.g., normal or Poisson, in which the mean is a function of the state) with the state evolving as a Markov process.

(2) *Performance-bound analysis of data-driven inventory decisions.* Although there is no precedent for feature-dependent performance bounds in the inventory theory literature, Levi et al. (2007) and Levi et al. (2015) provide such bounds without features. Levi et al. (2007) studies the single-period and dynamic inventory problems with zero setup cost from a nonparametric perspective and provides sample average approximation–type algorithms to solve them. They then provide probabilistic bounds on the minimal number of iid demand observations that are needed for the algorithms to be near optimal. Levi et al. (2015) improves upon the bound for the single-period case. Our performance bounds are generalizations of the bounds of Levi et al. (2007) and Levi et al. (2015) to incorporate features in the DM's data set. We demonstrate how our bounds can retrieve the bounds of Levi et al. (2007) when $p = 1$ in Online Appendix D; a similar result can be shown for Levi et al. (2015) as well.

(3) *High-dimensional quantile regression.* Because of the equivalence of the newsvendor cost function with the loss function in quantile regression, our work can be classified as a study in high-dimensional quantile regression as well, albeit with the twist that we analyze (analytically and empirically) the cost of the estimated quantile as opposed to the estimated quantile itself. We make two contributions to this literature. First, the KO method is, to the best of our knowledge, a new nonparametric quantile regression method. Second, our out-of-sample performance analyses of both the ERM methods and the KO method, which uses algorithmic stability theory from machine learning (Bousquet and Elisseeff 2002), are new. Finally, we extend the results of Chaudhuri (1991) to the high-dimensional setting to derive bounds on the biases of the newsvendor algorithms under consideration. For references on quantile regression, we refer the readers to Koenker (2005) for a textbook reference and Takeuchi et al. (2006), Chernozhukov and Hansen (2008), Chernozhukov et al. (2010), and Belloni and Chernozhukov (2011) and references therein for more recent works on high dimensionality.

2. Solving the Newsvendor Problem 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

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

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 back-ordering and holding costs. If the demand distribution, F, is known, one can show the optimal decision is given by the $b/(b+h)$ quantile, that is,

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

2.2. The Data-Driven Newsvendor Problem

In practice, the decision maker does not know the true distribution. If one has access to historical demand observations $\mathbf{d}(n) = [d_1, \ldots, d_n]$ but no other information, then a 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 ˆ notation to emphasize quantities estimated from data. This approach is called the sample average approximation approach in stochastic optimization (for further details on the SAA approach in stochastic optimization, 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 cumulative distribution function (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 practice, the demand depends on many observable *features* (equivalently, independent/explanatory variables, attributes, or characteristics), such as seasonality (day, month, season), weather, location, and economic indicators, which are available prior to making the order. In other words, the real newsvendor problem is to optimize the *conditional* expected cost function:

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

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

The decision maker intent on finding an optimal order quantity in this new setting has three issues to address. The first issue is knowing on what features the demand depends, which prescribes what data to collect. As this is application-specific, we assume that the decision maker has already collected appropriate historical data $S_n = [(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_n, d_n)]$. The data may be low dimensional, where the number of features p is small compared with the number of observations n, or high dimensional, where the number of features is large compared with n (analysis of identifying the low- and high-dimensional regimes is in Section 4). The second issue is how to solve the problem (5) in an efficient manner given the feature-demand data set. In this section, we propose two approaches to solving (5): the ERM and KO approaches. Both approaches are direct in that the decision maker solves for the (in-sample) optimal order quantity in a single step. As such, our proposed algorithms are customized for the feature-based newsvendor problem and are distinct from SAA (which are independent of features) and the separated estimation and optimization (SEO) approach, against which we compare in Section 5 along with other known benchmarks. The final concern is what performance guarantee is possible prior to observing the demand in the next period. We address this in Section 4.

Before we begin, we clarify that the setting of interest is one in which the DM observes the features $\mathbf{x}_{n+1}$ before making the ordering decision at time $n+1$.

2.3.1. Empirical Risk Minimization Algorithms.

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

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

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

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

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

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

$$D \approx f(\mathbf{0}) + \partial f(\mathbf{0})'\mathbf{x} + \mathbf{x}'[D^2 f(\mathbf{0})]\mathbf{x} + \ldots + \varepsilon$$
$$= f(\mathbf{0}) + \sum_{i=1}^{p} \partial f_i(\mathbf{0}) x_i + \sum_{i=1}^{p} \sum_{j=1}^{p} [D^2 f(\mathbf{0})]_{ij} x_i x_j + \ldots + \varepsilon, \tag{6}$$

which means that the demand function of a basic feature vector $\mathbf{x}$ can be approximated by a linear demand model with a much larger feature space. For example, the second-order Taylor approximation of the demand model can be considered to be a linear demand model with the $(p+p^2)$ features mentioned earlier: $[x_1, \ldots, x_p, x_1^2, \ldots, x_p^2, x_1x_2, \ldots, x_{p-1}x_p]$. Regardless of the motivation for the transformations of basic features, we can choose them to be arbitrarily complex; hence, our choice of decision functions that depend linearly on the feature vector is not particularly restrictive. The choice of $\mathcal{Q}$ can be made more or less complex depending on which transformations are included. For a discussion on how piecewise linear decisions can be considered by transforming the features, see Section 2.4.2. This comes at the cost of increasing the number of features, perhaps dramatically so, thereby increasing the likelihood of overfitting if there are not enough data. We, thus, propose ERM with regularization for large p (discussed further in Section 4). Although we consider linear decision functions for the rest of the paper, we show how one can consider nonlinear functional spaces for $\mathcal{Q}$ in Online Appendix A via mapping the original features onto a higher dimensional reproducing kernel Hilbert space.

When p is relatively small, the DM can solve (NV-ERM) via the following linear program:

ERM Algorithm 1

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \hat{R}(q(\cdot); S_n)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+\right]$$
$$\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \frac{1}{n}\sum_{i=1}^{n}(bu_i + ho_i)$$
$$\text{s.t. } \forall\, i = 1, \ldots, n:$$
$$u_i \ge d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j$$
$$o_i \ge q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i$$
$$u_i, o_i \ge 0, \tag{NV-ERM1}$$

where the dummy variables u_i and o_i represent, respectively, underage and overage costs in period i. This is an LP with a $p+2n$-dimensional decision vector and $4n$ constraints. We see in Section 4 that although (NV-ERM1) yields decisions that are algorithmically stable, the performance guarantee relative to the true optimal decision is loose when p is large (relative to n). Thus, in the case of high-dimensional data, one can solve the LP (NV-ERM1) by selecting a subset of the most relevant features according to some feature-selection criterion, for example, via cross-validation or via model selection criteria, such as the Akaike information criterion (Akaike 1974) or Bayesian information criteria (Schwarz 1978). Alternatively, one can automate feature selection by solving the following *regularized* version of (NV-ERM1):

ERM Algorithm 2 (with Regularization)

$$\min_{q:q(\mathbf{x})=\sum_{j=1}^{p} q^j x^j} \hat{R}(q(\cdot); S_n) + \lambda \|q\|_2^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[b(d_i - q(\mathbf{x}_i))^+ + h(q(\mathbf{x}_i) - d_i)^+\right] + \lambda\|\mathbf{q}\|_k^2$$
$$\equiv \min_{\mathbf{q}=[q^1,\ldots,q^p]} \frac{1}{n}\sum_{i=1}^{n}(bu_i + ho_i)$$
$$\text{s.t. } \forall\, i = 1, \ldots, n:$$
$$u_i \ge d_i - q^1 - \sum_{j=2}^{p} q^j x_i^j$$
$$o_i \ge q^1 + \sum_{j=2}^{p} q^j x_i^j - d_i$$
$$u_i, o_i \ge 0, \tag{NV-ERM2}$$

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

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

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

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

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

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

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

where $K_w(\cdot)$ is a kernel function with bandwidth w. Typical examples of the kernel function include the uniform kernel

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

where $I(\cdot)$ is the indicator function, and the Gaussian kernel

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

with $K_w(\cdot) := K(\cdot / w)/w$. Now, for an order quantity q, the feature-dependent newsvendor expected cost after observing features $\mathbf{x}_{n+1}$ is given by

$$\mathbb{E}[C(q; D) \mid \mathbf{x}_{n+1}], \tag{7}$$

which depends (implicitly) on the demand distribution at $\mathbf{x}_{n+1}$. Thus, if we consider the newsvendor cost to be the dependent variable, we can estimate (7) by the Nadaraya–Watson estimator

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

This gives rise to a new approach to the feature data–driven newsvendor, which we call the kernel optimization method.

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

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

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

$$\hat{q}_n^\kappa = \hat{q}_n^\kappa(\mathbf{x}_{n+1}) = \inf\left\{q : \frac{\sum_{i=1}^n \kappa_i \mathbb{I}(d_i \le q)}{\sum_{i=1}^n \kappa_i} \ge \frac{b}{b+h}\right\}, \tag{8}$$

where, for simplicity, we introduce $\kappa_i = K_w(\mathbf{x}_{n+1} - \mathbf{x}_i)$. In other words, we can find $\hat{q}_n^\kappa$ by ranking the past demand in increasing order and choosing the smallest value at which the inequality in (8) is satisfied.

Notice that the left hand side (lhs) of the inequality in (8) is similar to the empirical cdf of the demand except that each past demand observation d_i is reweighted by the distance of its corresponding feature $\mathbf{x}_i$ to the current feature $\mathbf{x}_{n+1}$.

2.4. Comparison with Other Inventory Papers that Incorporate Features

In the preceding sections, we identified three systematic approaches to incorporating features for the newsvendor problem. However, incorporating exogenous information in inventory decision making is not entirely new. In what follows, we compare and contrast the algorithms introduced thus far to past works that incorporate exogenous information in inventory decision making.

2.4.1. Comparison with Liyanage and Shanthikumar (2005). Our first 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, that is, $D \sim \exp(1/\theta)$, and that the decision maker has access to $d_1, \dots, d_n$ observations of past data. Then, with straightforward calculations, one can show first estimating then optimizing ("separated estimation and optimization") leads to the decision

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

where $\bar{d}_n$ is the sample average of the demand. Now consider instead the decision

$$\hat{q}_{OS}^1(\alpha) = \alpha \bar{d}_n \tag{9}$$

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

$$\min_{\alpha \ge 0} \mathbb{E}_\theta[C(\hat{q}_{OS}^1(\alpha); D)]. \tag{10}$$

As $\alpha = \log((b+h)/b)$ is a feasible solution of (10), 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

$$\mathbb{E}_\theta[C(\hat{q}^1_{OS}(\alpha^*); D)] \le \mathbb{E}_\theta[C(\hat{q}_{SEO}; D)], \tag{11}$$

where α^* is the optimal parameter in (10). 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

$$\hat{q}^2_{OS}(\alpha, \beta) = d_{[\beta-1]} + \alpha(d_{[\beta]} - d_{[\beta-1]}), \tag{12}$$

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

$$\min_{\alpha \ge 0, \beta \in \{1, \ldots, n\}} \mathbb{E}_\theta[C(\hat{q}^2_{OS}(\alpha, \beta); D)]. \tag{13}$$

As this 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 valuable. In Section 5, we consider solving (NV-ERM1) and (NV-ERM2) both without and with OS-inspired features to evaluate their practical benefit in terms of the out-of-sample cost.

2.4.2. Comparison with See and Sim (2010). See and Sim (2010) investigate the multiperiod inventory management problem in the presence of features such as "market outlook, oil prices, trend, seasonality, cyclic variation" that a product demand can depend on. Specifically, they model the demand at time t as

$$d_t(\tilde{z}) = d^0_t + \sum_{k=1}^{N_t} d^k_t \tilde{z}_k,$$

where $\tilde{z} = [\tilde{z}_1, \ldots, \tilde{z}_{N_t}]$ represents random features and d^k_t, $k = 0, \ldots, N_t$ are the coefficients. They then make a number of assumptions on the random features (zero mean; positive definite covariance matrix; bounded support set, which is second-order conic representable; etc.), all of which are assumed to be known to the DM, and solve an approximation of the robust problem by considering linear decision rules (linear as a function of the features) as well as piecewise linear decision rules. See and Sim (2010) demonstrate that piecewise linear decision rules have the best performance.

Although See and Sim (2010) certainly consider the presence of features in their problem setup, their work is distinctly different from ours on a number of fronts. First of all, See and Sim (2010) is about *robust* decision making as opposed to *data-driven* decision making, and as such, their theoretical results, although interesting, do not pertain to the data-driven questions we explore in this paper. Second, See and Sim (2010) assume that the DM has access to a number of key statistics regarding the random features known to the DM, and performance bounds for their approximate decisions are derived as a function of these statistics. In contrast, we do not make any assumptions about the data-generating process other than iid, and as such, our performance analyses are independent of statistics that are presumed known. Third, See and Sim (2010) do not study the effect of high dimensionality (the "big" in big data), which is the central theme of this paper.

Finally, See and Sim (2010) make an interesting observation that considering decisions that are piecewise linear functions of the underlying features perform better than linear decision rules. However, in this paper, we consider only linear decision rules because nonlinear decisions can be transformed into linear decision rules with the addition of new features. We explained how to enlarge the feature space when the true decision is an analytic function of the features in Section 2.3.1. Let us now illustrate how to enlarge the feature space to allow for piecewise linear decisions. For simplicity, consider a single feature x. We can then construct new features $\tilde{x}_1 = \mathbb{I}(x \le c_1)$, $\tilde{x}_2 = \beta\mathbb{I}(c_1 < x \le c_2)$, and $\tilde{x}_3 = \mathbb{I}(x > c_3)$ based on the "basic" feature x and some prespecified constants c_1 and c_2. Then the corresponding linear decision rule

$$\mathbf{q} = q^0 + q^1\tilde{x}_1 + q^2\tilde{x}_2 + q^3\tilde{x}_3$$

has the same piecewise linear structure as the decisions considered in See and Sim (2010). By incorporating a large number of breakage points, any piecewise linear decision can be approximated by linear decision rules arbitrarily well although perhaps with the addition of a large number of "new" features corresponding to each breakage point. Thus, we solely focus on the feature-based newsvendor problem with decisions that are linear functions of the feature vector when the feature dimension may be large.

2.4.3. Comparison with Hannah et al. (2010).

Hannah et al. (2010) consider the single-period stochastic optimization problem

$$\min_{x\in\mathcal{X}} \mathbb{E}_Z[F(x,Z)\,|\,S=s]$$

where $x\in\mathcal{X}$ is the decision variable, Z is a state-dependent random variable, and $S=s$ is the current state of the world. They propose solving this problem by the weighted empirical stochastic optimization problem

$$\min_{x\in\mathcal{X}} \sum_{i=1}^{n} w_n(s,S_i)F(x,Z(S_i)),$$

where $w_n(s,S_i)$ are the weights given to the past data $(S_i,Z(S_i))_{i=1}^n$ determined by the Nadaraya–Watson-based kernel estimator as in the (KO) method or by a complex Dirichlet process mixture model.

This is similar to our feature-based newsvendor setup; however, there are a few key differences. First, the model studied by Hannah et al. (2010) requires discrete state variables whereas we consider both discrete and continuous feature variables. Second, Hannah et al. (2010) do not study the effect of high dimensionality (the "big" in big data), which is the central theme of this paper (their numerical example, which also considers the newsvendor problem, has only two states). Finally, Hannah et al. (2010) report numerical studies of computing the in-sample decisions whereas we provide both theoretical performance guarantees and extensive empirical computation of out-of-sample performance of the in-sample decisions.

3. Value of Feature Information

In this section, we quantify the value of incorporating features in newsvendor decision making by comparing the (NV-ERM1) decision against the SAA decision, which is made with only past demand data. We consider two demand models: a two-population demand model and the linear demand model. In both cases, the SAA decision yields inconsistent decisions (i.e., the in-sample decision does not converge to the true optimal even when given an infinite amount of iid demand data) whereas the feature-based decision is consistent. We further quantify the implication on the expected cost, which is necessarily larger for decisions that are further away from the true optimal. All proofs can be found in Online Appendix B.

3.1. Motivating Example I: Two-Population Model

Consider the following demand model:

$$D = D_0(1-x) + D_1 x, \tag{14}$$

where D_0 and D_1 are nonnegative continuous random variables such that the corresponding critical newsvendor fractiles q_0^* and q_1^* follow $q_0^* < q_1^*$, and $x\in\{0,1\}$ is a binary feature (e.g., 0 for weekday and 1 for weekend or 0 for male and 1 for female). Let p_0 be the proportion of time $x=0$. We have n historical observations: $[(x_1,d_1),\ldots,(x_n,d_n)]$, of which $n_0=np_0$ are when $x=0$ and $n_1=n-n_0$ are when $x=1$ (assume rounding effects are negligible). Note the observations d_k can be decomposed into $\{d_k\,|\,x_k=0\}=d_k^0$ and $\{d_k\,|\,x_k=1\}=d_k^1$. Also let $r=b/(b+h)$ to simplify notations. Let F_0 and F_1 denote the cumulative distribution functions, 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.

We also assume the following.

Condition A. *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*

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

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

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

Condition A is satisfied by many standard distributions, such as uniform, exponential, logistic, normal, and log normal, for γ between zero and ~1.24. The critical ratios for the uniform, exponential, and logistic distributions can be computed straightforwardly; for the normal distribution, it is easier to compute the critical ratio by using the following equivalent formulation for the critical ratio:

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

In Table 1, we display some standard distributions that satisfy the requirement of Condition A. For more details see Parzen (1979).

In the following, we derive the optimal in-sample solution given by (NV-ERM1).

Table 1. Some Standard Distributions that Satisfy the Requirement of Condition A

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	$\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\lvert\Phi^{-1}(y)\rvert^2\right\}$	$\Phi^{-1}(y)$	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

Note. The standard normal cdf and pdf are denoted as $\Phi(\cdot)$ and $\phi(\cdot)$, respectively.

Lemma 1 (Optimal Ordering Decision of (NV-ERM1)). *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-ERM1) is given by*

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

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

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 $x = 1$.

Theorem 1 (Finite-Sample Bias and Asymptotic Optimality of (NV-ERM1)). *We can show*

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

that is, the finite-sample decision of the feature-based decision is biased by at most $O(\log n_i/n_i)$, $i = 1, 2$, and

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

that is, 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*

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

Theorem 2 (Finite-Sample Bias and Asymptotic (Sub)-optimality of SAA). *The finite-sample bias of the SAA decision is given by*

$$|\mathbb{E}[\hat{q}_n^{SAA}] - (F^{mix})^{-1}(r)| \leq O\left(\frac{\log n}{n}\right), \tag{15}$$

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

$$|\mathbb{E}[\hat{q}_n^{SAA} - \hat{q}_n^0]| = |(F^{mix})^{-1}(r) - F_0^{-1}(r)| + O\left(\frac{\log n}{n}\right) = O(1)$$
$$|\mathbb{E}[\hat{q}_n^1 - \hat{q}_n^{SAA}]| = |F_1^{-1}(r) - (F^{mix})^{-1}(r)| + O\left(\frac{\log n}{n}\right) = O(1). \tag{16}$$

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,

$$q_0^* < \lim_{n\to\infty} \hat{q}_n^{SAA} \overset{a.s.}{=} (F^{mix})^{-1}(r) < q_1^*, \tag{17}$$

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

As a final point, we remark that these observations are analogous in spirit to the bias and inconsistency of regression coefficients when there are, in econometric parlance, correlated omitted variables in the model.

3.2. Motivating Example II: Linear Demand Model

Suppose the demand is given by the following linear model:

$$D|(\mathbf{X} = \mathbf{x}) = \beta^\top \mathbf{x} + \varepsilon, \tag{18}$$

where $\varepsilon \sim F_\varepsilon$ is independent of the (random) feature vector $\mathbf{X}$, is continuous with probability density function $f_\varepsilon(\cdot)$, which is bounded away from zero on the ordering domain $[\underline{D}, \bar{D}]$; and has zero mean, and $\mathbf{X}_1 = 1$ almost surely to allow for a constant location term. In other words, the demand D depends linearly on the random features $\mathbf{X} : \mathcal{X} \to \mathbb{R}^p$ with some error. This is a widely used and useful demand model that, apart from the fact that it can arbitrarily approximate nonlinear models as outlined in (6), also subsumes times series models and the Martingale model of forecast evolution (MMFE) of Graves et al. (1986) and Heath and Jackson (1994). For example, the autoregressive model of degree p, $AR(p)$ is modeled by

$$D_t = \alpha_0 + \alpha_1 D_{t-1} + \ldots + \alpha_p D_{t-p} + \varepsilon_t,$$

where ε_t is a noise term with zero mean; this is clearly a linear demand model with features $D_{t-p}, \ldots, D_{t-1}$. Also, the additive MMFE model for the demand at time t is given recursively by

$$D_t = D_{t-1} + \varepsilon_{t-1,t},$$

where $\varepsilon_{t-1,t}$ is a mean zero normal random variable that captures forecast update at time $t-1$ for demand at time t. Expanding the recursion, we get, at time 1,

$$D_t = D_0 + \sum_{i=0}^{t-1} \varepsilon_{i,i+1},$$

where D_0 is known, and so the demand follows a linear model with $\varepsilon_{i,i+1}$, $i = 0, \ldots, t-1$ as features.

A DM without the feature information only has access to past demand data: $\mathcal{D} = \{d_1, \ldots, d_n\}$; and a DM who has both past feature and demand data has the information: $\mathcal{D}_{\mathbf{x}} = \{(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_n, d_n)\}$. Let $\mathbf{x}_{n+1}$ denote the feature at time $n+1$, which is available to the DM. Then, the optimal order quantity is given by

$$q^*(\mathbf{x}_{n+1}, \mathbf{z}_{n+1}) = Q_\varepsilon\left(\frac{b}{b+h}\right) + \beta^\top \mathbf{x}_{n+1} \tag{19}$$

where

$$Q_\varepsilon\left(\frac{b}{b+h}\right) = \inf\left\{y : F_\varepsilon(y) \geq \frac{b}{b+h}\right\}$$

is the $b/(b+1)$ quantile of the distribution of ε.

The SAA solution is given by

$$\hat{q}^{SAA}(\mathbf{x}_{n+1}) = \inf\left\{y : \hat{F}_n^0(y) \geq \frac{b}{b+h}\right\} + \bar{d}_n \tag{20}$$

where $\bar{d}_n$ is the sample average of the demand data and $\hat{F}_n^0$ is the empirical distribution of $\{d_i - \bar{d}_n\}_{i=1}^n$. Note the SAA decision is not dependent on any of the features as the DM does not have access to any feature data. Thus, the SAA decision is to order the same critical fractile quantity for the entire population at time $n+1$ regardless of what the population or the particular point in time may be. As a concrete example, this is like a national newspaper vendor who stocks the same number of newspapers at all shops, disregarding location features pertaining to the shop (e.g., customer demographics in the area) as well as relevant temporal features (e.g., holiday or not, weekday versus weekend, major political or sporting events) or features that pertain to both (e.g., historical demand for the shop).

The DM with features, however, orders the quantity

$$\hat{q}^{DM2}(\mathbf{x}_{n+1}) = \inf\left\{y : \hat{F}_n^2(y) \geq \frac{b}{b+h}\right\} + \sum_{k=2}^{p} \hat{q}^k x_{n+1}^k$$

where $\hat{F}_n^2$ is the empirical distribution of $\{d_i - \sum_{k=2}^p \hat{q}^k x_i^k\}$ and $\hat{q}^k$, $k = 2, \ldots, p$ are the solution coefficients to (NV-ERM1). Unlike the SAA decision, DM2's decision does depend on all relevant features. Continuing on with the national newspaper example, this corresponds to orders being different across stores as well as in time, taking into account such information as past sales and customer demographics at each store as well as temporal effects, such as holidays/weekends and major public events.

Theorem 3 (Using No Features Leads to Inconsistent Decisions). *Under the linear demand model of* (18), *given features* $\mathbf{X} = \tilde{\mathbf{x}}$,

$$\hat{q}_n^{SAA}(\tilde{\mathbf{x}}) \xrightarrow{a.s.} Q_\varepsilon\left(\frac{b}{b+h}\right) + \mathbb{E}_{\mathbf{X}}[\mathbb{E}_\varepsilon[D|\mathbf{X}]]$$
$$= Q_\varepsilon\left(\frac{b}{b+h}\right) + \beta^\top \mathbb{E}[\mathbf{X}],$$

and

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

as n tends to infinity.

Considering the same national newspaper example as before, we see that the SAA decision converges to the critical fractile of the dispersion ε plus the population-temporal average demand $\mathbb{E}_{\mathbf{X}}[\mathbb{E}_\varepsilon[D\,|\,\mathbf{X}]]$ whereas DM2's decision converges to the correct one, $q^*(\tilde{\mathbf{x}})$.

The results of Theorems 1–3 indicate that the no-feature SAA decisions are inconsistent; that is, even with an infinite amount of demand data, the SAA decisions converge to quantities different from the true optimal. The natural question "is a decision that is further away from the true optimal necessarily worse in terms of the expected cost?" then follows. In other words, does the loss in the expected cost increase when the effect of the feature information increases? The answer is in the affirmative, which we detail as follows.

Theorem 4 (Expected Cost Difference). *Let $q^* = q^*(\tilde{\mathbf{x}})$ denote the true optimal newsvendor decision given feature $\tilde{\mathbf{x}}$ and $\hat{q}$, some other decision not equal to q^*. Then the difference of the expected costs of the two decisions is given by*

$$\mathbb{E}C(\hat{q}; D) - \mathbb{E}C(q^*; D) = (b+h)\mathbb{E}[|\hat{q} - D|\,\mathbb{I}\{(\hat{q} \wedge q^*) \leq D \leq (\hat{q} \vee q^*)\}]. \tag{21}$$

Theorem 4, thus, provides an exact formula for the expected cost difference of the suboptimal decision $\hat{q}$ from the true optimal decision q^*. We observe that the expected cost difference scales as the expectation of $|\hat{q} - D|$ over the interval between the two decisions, $\hat{q}$ and q^*. Thus, what matters is the size of this interval and how the demand is distributed over it—the more concentrated the distribution over this interval, the larger the difference. Although the exact quantity can only be computed with the knowledge of the demand distribution and the true optimal decision, we nevertheless arrive at the universal insight that the expected cost difference increases as $\hat{q}$ deviates further away from q^*.

In particular, Theorem 4 implies that for the two-population model (17), the more distinct the two population demands D_0 and D_1 in their critical fractiles, the worse the expected cost of the no-feature SAA decision to the true optimal solution. Likewise, for the linear demand model (18), the more idiosyncratic the feature information $\beta^\top \tilde{\mathbf{x}}$ over the average $\beta^\top \bar{\mathbf{X}}$, the worse the expected cost of the no-feature SAA decision in comparison with the true optimal decision.

Although Theorem 4, together with Theorems 1–3, justifies the collection of features, a shortcoming is that the DM needs to know the demand distribution to quantify the gain in the expected cost resulting from a suboptimal in-sample decision, which, of course, the DM does not know. In the next section, we characterize, with high probability bounds, the expected cost of the DM's in-sample decision using the information at hand.

4. Bounds on the Out-of-Sample Cost Using In-Sample Information

In this section, we provide theoretical guarantees on the out-of-sample cost of the ordering decisions chosen by (NV-ERM1), (NV-ERM2), and (NV-KO). Although (EC.4) from Theorem 4 states the exact expected cost difference between a proposed decision and the optimal decision, in practice one cannot compute the expectations because the demand distribution is unknown. The goal of this section is, thus, to provide performance bounds that are computable with the data at hand. We see that the performance bound splits into two components: one that pertains to the bias arising from having a finite amount of data, commonly referred to as the *finite-sample bias*, and one that pertains to the variance of the in-sample decision, known as the generalization error.

The term "generalization" refers to the generalizability of the in-sample decision to out-of-sample data, and is a measure of the degree of overfitting as decisions with larger generalization errors are associated with greater overfitting. Decisions that overfit can be misleading; for instance, if there is a large degree of freedom in the choice of the in-sample decision, then it may be possible to have zero in-sample cost, leading the DM to think perfect ordering is possible. An astute DM who is aware of the perils of overfitting, however, would be cautious to make conclusions on the in-sample data alone, and the generalization error provides a description of how overfitting arises for the DM's decision.

The performance bounds derived in this section describe the mechanisms at play. They inform how the finite-sample bias and the generalization error scale with respect to the size of the data set at hand (i.e., with respect to p and n), the problem parameters (b and h), and any decision parameters (e.g., regularization parameter λ in (NV-ERM2)). As we see, there is a tension between finite-sample bias and generalization error, which can be controlled by parameters such as the regularization parameter λ in (NV-ERM2) or the bandwidth parameter w in (NV-KO).

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

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

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

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

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

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

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

which leaves one observation out.

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

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

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

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

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

We derive performance bounds on (NV-ERM1)–(NV-KO) under the following assumptions.

4.1. Assumptions for Theorems 5–7

Assumption 1. *The feature vector $\mathbf{X}$ is normalized ($\mathbf{X}_1 = 1$ almost surely, $\mathbf{X}_{[2:p]}$ has mean zero and standard deviation one), and it lives in a closed unit ball: $\|\mathbf{X}\|_2 \leq X_{\max}\sqrt{p}$.*

Assumption 2. *The demand follows the linear model (18), where the distribution of ε, f_ε, is bounded away from zero on the domain $[\underline{D}, \bar{D}]$ (otherwise unspecified).*

Assumption 3. *All decision functions (policies) described are measurable, and $\mathcal{Q}$ is a convex subset of a linear space.*

Assumption 1 is for the feature vector $\mathbf{X}$; we note the normalization assumption is to simplify the exposition, and the results do not require that the DM knows the true mean or standard deviations of $\mathbf{X}$, only the size bound $X_{\max}$, the existence of which is realistic and not prohibitive. Assumption 2 details assumptions on the demand model, which is assumed to be linear for tractability, but we do not assume any distributional knowledge beyond its total range. Finally, Assumption 3 is a necessary requirement for sensible optimization over a function class when the demand is linear.

First, we state the performance bound on (NV-ERM1).

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

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

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

Theorem 5 is a statement about how close the in-sample cost of the in-sample decision, $\hat{R}_{in}(\hat{q}; S_n)$, is to the expected cost of the true optimal decision, $R_{true}(q^*)$, in terms of quantities the DM knows.

The first term on the right-hand side upper bound is the bound on the generalization error, which is the difference between the training error and test error for the in-sample decision. For fixed cost parameters b and h, we find that the generalization error scales as $O(p/\sqrt{n})$. Thus, if the number of relevant features in the population model is small and not growing relative to the number of observations, then in-sample decisions generalize well to out-of-sample data; in other words, overfitting should not be an issue. However, if $p/\sqrt{n}$ is large, or growing (which happens when new observations are associated with new features), then overfitting will be an issue, and Theorem 5 suggests that (NV-ERM1) may not be a good algorithm in such a scenario. The dependence on the upper bound on the demand, $\bar{D}$ is necessary so that the bound is not scale invariant. In other words, if the risks on the left-hand side of the inequality changed units (e.g., from dollars per kilo of demand to dollars per ton), it would not make sense for the right-hand side of the inequality to stay the same. Finally, we note that the bound on the generalization error is tight in the sense that it comes from showing that the probability of large deviation of $|\hat{R}_{true}(\hat{q}) - \hat{R}_{in}(\hat{q}; S_n)|$ decays exponentially fast in the number of observations n, which we establish through a property known as *uniform stability* of a learning algorithm and because the constants in the bound are the smallest possible. For further details, we refer the reader to the proof in Online Appendix C and Bousquet and Elisseeff (2002). These are the best finite-sample bounds we know of for this problem. This is because they are not uniform bounds, which require a complexity measure for the entire decision space (e.g., covering numbers, VC dimension, or Rademacher complexity), rather algorithm-specific bounds that consider how the algorithm searches the decision space.

The second term on the right-hand side upper bound is due to the finite-sample bias, $\mathbb{E}|q^* - \hat{q}|$. The only way a DM can reduce the finite-sample bias is by collecting more observations. The rate $n^{-1/(2+p/2)}\sqrt{\log n}$ is optimal and cannot be improved upon without further assumptions on the demand model and/or the data-generating process. For details, we refer the reader to the proof of Theorem 5 in Online Appendix C.

When $p = 1$, we are in the setup in which the demand does not depend on any exogenous features (recalling that $\mathbf{X}_1 = 1$ is the intercept term). This setting was studied by Levi et al. (2007), and our results are consistent with them in that their sampling bound, up to a constant factor, can be obtained from our bound. The details of this can be found in Online Appendix D.

We now state the performance bound on (NV-ERM2).

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

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

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

The performance bound of Theorem 6 has three components. The first term is a bound on the generalization error, which is of $O(p/(\sqrt{n}\,\lambda))$. Thus, the amount of overfitting can be directly controlled by the amount of regularization imposed on the problem; the larger the λ, the smaller the generalization error. Choosing $\lambda = O(1/p^2)$ retrieves the same error rate as for (NV-ERM1), so $\lambda = O(1/p^2)$ is a good starting point for choosing λ. The bound, thus, provides a sense of the "right" scale for lambda, which is useful when you have to search for its best value in practice.

The second term, which does not appear in Theorem 5, is the bias of the in-sample decision resulting from regularization—in other words the bias resulting from having perturbed the optimization problem away from the true problem of interest. This term is larger for larger λ, and so there is an inherent trade-off between the generalization error and the regularization bias. Ultimately, however, regularization gives the DM an extra degree of control while being agnostic to which feature is important a priori and, in practice, would work with the optimal value of λ that balances the generalization error-regularization bias trade-off on a validation data set (see Section 5).

The third and final term is the finite-sample bias. We note that although the regularization bias can be controlled by λ, the finite-sample bias can only be controlled by collecting more data.

Finally, we have the following result for (NV-KO).

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

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

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

As with Theorem 6, the performance bound on (KO) has three components: a bound on the generalization error, the bias resulting from optimizing with a scalar decision when the true decision is a function, and the finite-sample bias term, which is the same as in Theorems 5 and 6. The generalization error is of $O(1/r_w(p)\sqrt{n})$, so it can be controlled by reducing $r_w(p)$ by increasing the kernel bandwidth w. Setting $w = O(\sqrt{p})$ gives an error that is of $O(1/\sqrt{n})$, which is as good as having the demand not depend on any features. When w is set to an arbitrarily large number, $r_w(p) = 1$, so the error rate $O(1/\sqrt{n})$ cannot be improved upon. It is not surprising that the generalization error can be made small with large w because this corresponds to smoother comparisons of the feature vectors from the past to the one in period $n+1$. However, as with (NV-ERM2), increased w increases the second term; thus, in practice, w needs to be optimized over a reasonable range of values. Finally, the finite-sample bias term plays the same role as the corresponding terms in Theorems 5 and 6.

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

In this section, we compare the three algorithms introduced in Section 2, (NV-ERM1), (NV-ERM2) and (NV-KO), against the main data-driven benchmarks known in the literature and practice through an extensive empirical investigation. Although some analytical comparisons are possible under assumptions about the true demand model, the ultimate test of data-driven methods must be on real data sets. Furthermore, we report other practical observations, such as the computational time required and trends in the optimal staffing solution.

In particular, we apply the three learning algorithms to find the optimal staffing levels of nurses for a hospital emergency room. As most hospitals in the developed world either impose or recommend a minimum nurse-to-patient ratio, we can approximate the nurse-staffing problem as a newsvendor problem if we assume the hospital incurs a linear underage cost if too many patients arrive and expensive agency nurses have to be called and a linear overage cost if too many regular nurses are scheduled compared with the number of patients. As nurse staffing contributes to a significant portion of hospital operations (see, e.g., Green et al. 2013) and as many hospitals are starting to harness the value of data, the nurse-staffing problem is a natural setting to test the data-driven models introduced in this paper.

Our data comes from the emergency room of a large teaching hospital in the United Kingdom from July 2008 to June 2009. The data set includes the total number of patients in the emergency room at two-hour intervals. We provide box plots of the number of patients by day and by time periods in Figure 1. We assumed a nurse-to-patient ratio of one to five; hence, the demand is the total number of patients divided by five. We do not require the staffing level to be an integer in our predictions as multiskilled workers could be used for part-time work. We also assumed 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 $r = b/(b+h) = 2.5/3.5$. Although the exact agency nurse rate differs by location, experience, and agency, our assumption is a modest estimate (Donnelly and Mulhern 2012).

We considered 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 and 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 analysis in Liyanage and Shanthikumar (2005) as described in Section 2.4.1. We refer to these features as *operational statistics* features. We used $n = 1{,}344$ past demand observations (16 weeks) as training data and computed the critical staffing level 3 periods ahead. We then recorded the out-of-sample newsvendor cost of the predicted staffing level on $1{,}344/2 = 672$ validation data on a rolling horizon basis, following the-rule-of thumb in Friedman et al. (2009) for choosing the size of the validation data set. Any parameter that needs calibration was calibrated on the validation data set. We then applied the algorithms to a test set of 672 unseen observations.

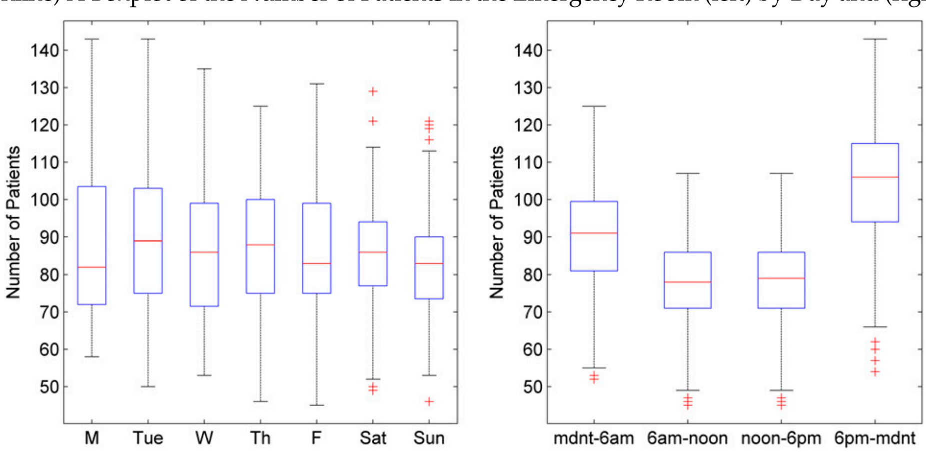
All computations were carried out on MATLAB2013a with the solver MOSEK and CVX, a package for specifying and solving convex programs (Grant and Boyd 2008, 2013) on a Dell Precision T7600 workstation with two Intel Xeon E5-2643 processors, each of which has four cores and 32.0 GB of RAM.

5.1. Methods Considered

We investigate the following methods in detail:

1. SAA by day of the week: take a sample average of the training data set by day of the week (because our training data set consists of 16 weeks of demand, there are $1344/16 = 84$ observations for each day of the week). We note that this is reflective of nurse staffing done in practice.
2. Cluster + SAA: we take the vector of features, then first classify them into $k = 2, \ldots, 12$ clusters before applying SAA. This is an intuitive and alternative method to use the feature data. For clustering, we use the k-means clustering algorithm.
3. Solve (NV-KO) with the Gaussian kernel with the day of the week and time of the day features and an increasing number of days of past demands (for up to two weeks) with and without operational statistics features. These constitute a total of two algorithms.
4. Solve (NV-ERM1) with the day of the week and time of the day features and an increasing number of days of past demands (for up to two weeks) with and without operational statistics features, which are explained in Section 2.4.1. These constitute a total of two algorithms.
5. Solve (NV-ERM2) with the day of the week and time of the day features and two weeks of past

Figure 1. (Color online) A Boxplot of the Number of Patients in the Emergency Room (left) by Day and (right) by Time Period



demands with and without OS features for a range of regularization parameters. We investigate both ℓ_1 and ℓ_2 regularizations. These constitute a total of four algorithms.

6. Separated estimation and optimization: a commonsense approach to incorporating features in 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 quantile using the assumption of normality for the demand (optimization). We use day of the week and time of the day features and an increasing number of days of past demands (for up to two weeks). We consider two cases: one without and one with OS features.

7. Separated estimation and optimization approach with ℓ_1 or ℓ_2 regularization: we apply OLS regression with ℓ_1 or ℓ_2 or with no regularization to first estimate a demand model, then choose the optimal quantile under the assumption of normally distributed demand. For the demand estimation step, we use day of the week and time of the day features and two weeks of past demands and consider using and not using OS features.

8. We also consider Scarf's Minimax approach (Scarf et al. 1958).

In Table 2, we summarize the abbreviations used to describe the 16 different methods considered in this paper.

5.2. Discussion of Results

In Table 3, we report the out-of-sample performance of the 16 methods considered. We report the calibrated parameter (if any), the mean, and the 95% confidence interval for the out-of-sample staffing cost in normalized units (parameters are calibrated by in-sample calculations, which can be found in Online Appendix E, Tables EC.1–EC.7). In the last column, we report the annual cost savings of the method relative to SAA-day with which there is a statistically significant net cost saving, assuming a regular nurse salary of £25,000 (which is the Band 4 nurse salary for the National Health Service in the United Kingdom in 2014) and standard working hours of 37.5 hours per week. Cost savings in U.S. dollars (USD) are also reported, assuming an exchange rate of £1:USD 1.6.

The best result was obtained by the KO method with OS features with bandwidth $w = 1.62$, which yields a cost improvement of 24% (a saving of £46,555 p.a.) relative to the best practice benchmark ("SAA-day") with statistical significance at the 5% level. The next best results were obtained by the ERM method with ℓ_1 regularization and the KO method without OS features, which have average annual cost improvements of 23% (£44,219 p.a.) and 21% (£39,915 p.a.), respectively.

The computational costs of the feature-based methods are very different, however. The KO method is *three* orders of magnitude faster than the ERM-based methods. For instance, it takes just 0.0494 seconds to find the next optimal staffing level using the KO method with OS features, which is the best in terms of the out-of-sample cost, whereas it takes 114 seconds for the ERM method with ℓ_1 regularization, which is the second-best performing method. The KO method is also faster than SAA-day, SEO methods, and Scarf by two orders of magnitude.

5.3. Optimal Staffing Decisions

Let us further investigate the staffing decision of the best method: KO-OS with $w = 1.62$. In Figure 2(a), we display the staffing levels predicted by KO-OS with $w = 1.62$ along with the actual required levels. For comparison, we also provide the staffing levels predicted by the second-best method, NVreg1-OS with $\lambda = 1 \times 10^{-7}$ in Figure 2(b). A striking observation is that both KO-OS and NVreg1-OS methods anticipate

Table 2. A Summary of the Methods Considered

Abbreviation	Description	OS features?	Regularization?	Free parameter
1a. SAA-day	SAA by day of the week	No	None	None
1b. Cluster + SAA	First cluster then SAA	No	No. of clusters	None
2a. Ker-0	solve (NV-KO) with Gaussian kernel	No	None	Bandwidth
2b. Ker-OS	"	Yes	None	"
3a. NV-0	Solve (NV-ERM1)	No	None	No. of days of past demand
3b. NV-OS	"	Yes	None	"
4a. NVreg1	Solve (NV-ERM2)	No	Yes, ℓ_1	Regularization parameter
4b. NVreg1-OS	"	Yes	Yes, ℓ_1	"
5a. NVreg2	"	No	Yes, ℓ_2	"
5b. NVreg2-OS	"	Yes	Yes, ℓ_2	"
6a. SEO-0	OLS regression + NV optimization	No	None	No. of days of past demand
6b. SEO-OS	"	No	None	"
7a. SEOreg1	Lasso regression + NV optimization	No	Yes, ℓ_1	Regularization parameter
7b. SEOreg1-OS	"	Yes	Yes, ℓ_1	"
8a. SEOreg2	Ridge regression + NV optimization	No	Yes, ℓ_2	"
8b. SEOreg2-OS	"	Yes	Yes, ℓ_2	"
9. Scarf	Minimax optimization	No	None	No. of days of past demand

Table 3. A Summary of Results

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

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

periods of high demand fairly well as evidenced by the matching of the peaks in the predicted and actual staffing levels. The two methods are otherwise quite different in the prediction; in particular, the KO-OS method balances both overstaffing and understaffing whereas the NVreg1-OS method seems to systematically overpredict the staffing level.

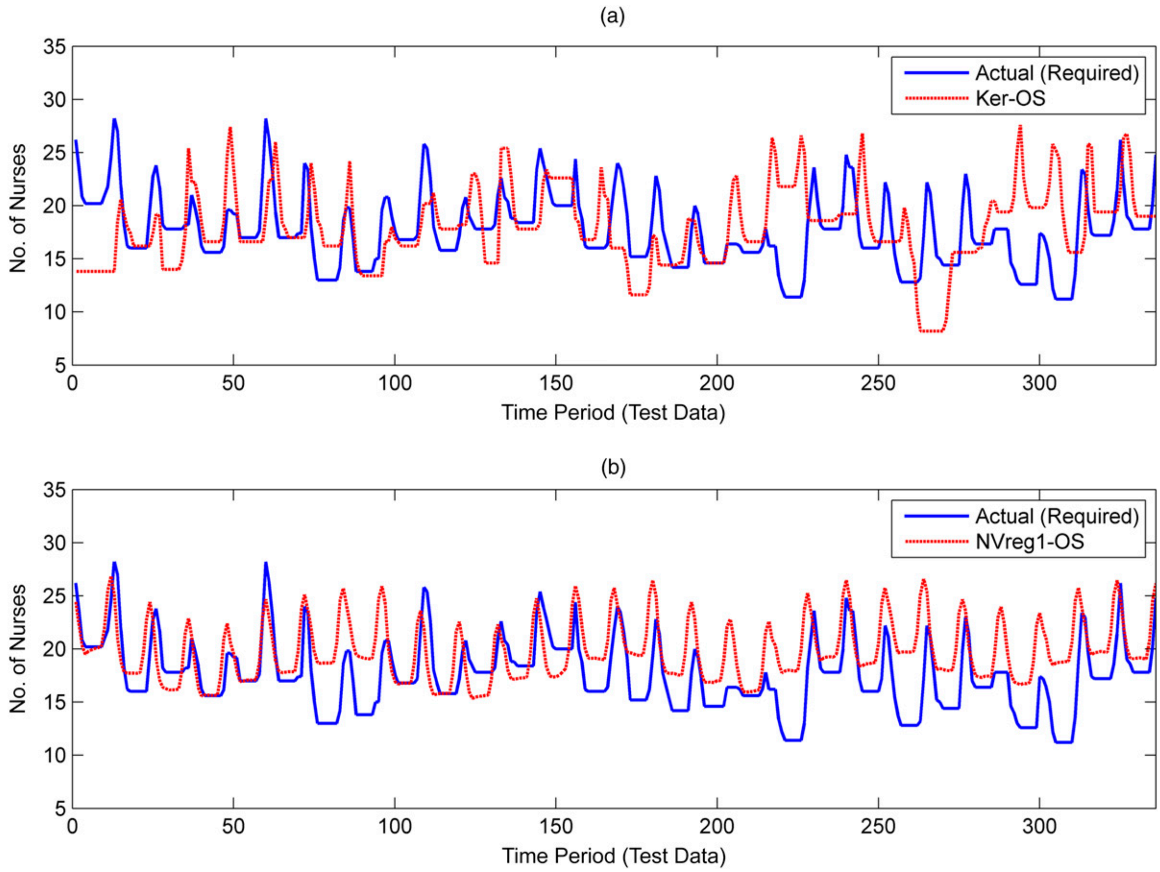
Let us now suppose the hospital indeed implements our algorithm for its nurse-staffing decisions. We wish to gain some insight into the predictions made by the algorithm. In particular, we would like to know when the hospital is overstaffed or understaffed, assuming the hospital chooses to implement the best possible method, provided by KO-OS with $w = 1.62$. In Figures 3 and 4, we show the conditional probability (frequency) of understaffing and overstaffing 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: (1) Midweek days are more likely to be understaffed than weekends; thus, given the choice to visit the emergency room on a weekday or weekend, we would choose a weekend. (2) The period from noon to midnight is substantially more likely to be overstaffed then the period from midnight to noon; thus, given the choice of time to visit the emergency room, we would choose visiting in the afternoon. (3) The algorithm is most likely to overstaff by at least 50% of the required level on a Monday then any other day of the week; hence, given the flexibility, we would choose to visit the emergency room on a Monday.

6. Conclusion

We investigated the newsvendor problem when the DM has historical data on both demands and p features that are related to the demand. We have analyzed this problem using recent techniques from machine learning (algorithmic stability theory) as well as theoretical statistics (theory of quantile estimators). Rather than reiterate the contributions detailed in the Introduction, we summarize some practical insights that may not be obvious upon first reading (especially to readers unfamiliar with machine learning) and discuss potential directions for future research.

Some practical insights from this work are as follows. (1) There is not a single approach to solving the "big data" newsvendor problem. In this paper, we proposed three approaches (ERM with and without regularization and KO) to using the feature-demand data set and compared against a number of other potential approaches to solving the problem with or without features. This is not to say these approaches are exhaustive; we are optimistic that new methods can be developed. (2) A "big data"–driven decision maker always needs to be weary of overfitting when decisions can perform well in-sample but not so out-of-sample, hence, leading to not only decisions that perform badly, but also decisions that are misleading. We have shown that overfitting can be controlled by a priori feature selection or regularization, which gives the DM extra control to bias the decision in favor of improved out-of-sample generalization. (3) What

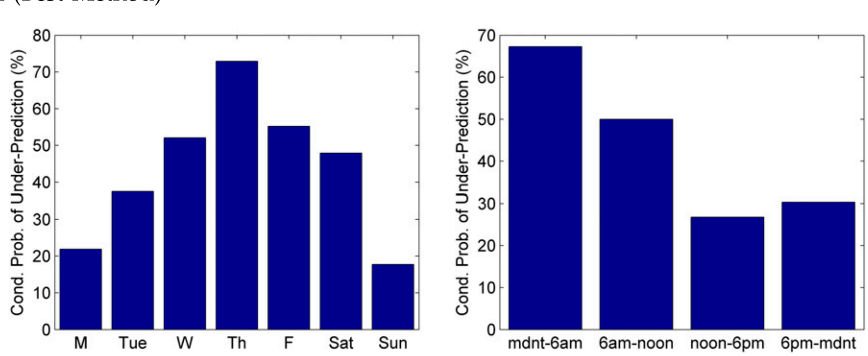
Figure 2. (Color online) Actual Required Number of Nurses Versus Values Predicted by the Two Best-Performing Algorithms



Notes. (a) A time-series plot of actual staffing demand (solid line) vs. staffing levels predicted by KO-OS with $w = 1.62$ (best method) on test data (dotted line). (b) A time-series plot of actual staffing demand (solid line) vs. staffing levels predicted by NVreg1-OS with $\lambda = 1 \times 10^{-7}$ (second best method) on test data (dotted line).

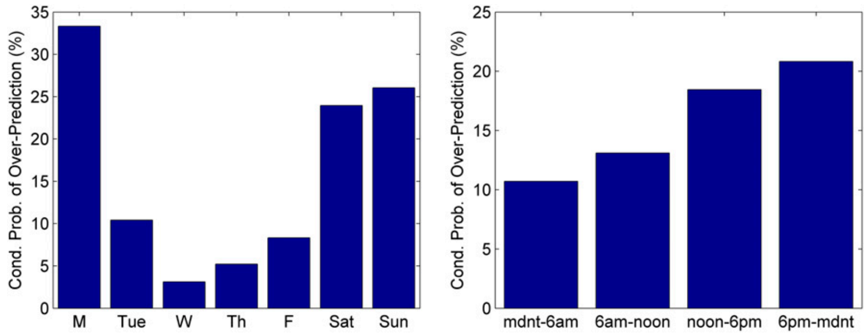
affects the true performance of the in-sample decision are the generalization error (a.k.a. overfitting error), bias from regularization, and finite-sample bias simply from having a finite amount of data. The DM can control the generalization error and bias from regularization through the regularization parameter, and in practice, needs to optimize over a range of parameter values on a separate validation data set. Finite-sample bias cannot be controlled except by collecting more data. (4) Although the KO method dominates all others in terms of the out-of-sample performance in our case study, this need not be the case for a different data set. We believe the most important point of our case study is in demonstrating how to carry out a careful data-driven investigation, not in the case-specific conclusion that the KO method performed

Figure 3. (Color online) A Plot of the Conditional Probabilities of Understaffing (left) by Day and (right) by Time Period for KO-OS with $w = 1.62$ (Best Method)



Note. The conditioning is done by the particular day or the time period, for example, the probability of understaffing given it is a Monday.

Figure 4. (Color online) A Plot of the Conditional Probabilities of Overstaffing by at Least 50% (left) by Day and (right) by Time Period for KO-OS with $w = 1.62$ (Best Method)



Note. The conditioning is done by the particular day or the time period, for example, the probability of overstaffing given it is a Monday.

the best. We note, however, the relative speed of the KO method is generalizable.

There are many directions for follow-up work, and we discuss a few. First, investigating how the dynamic inventory management problem can be solved with a demand-feature data set remains an open question. Second, as mentioned in point (1) of the previous paragraph, the methods considered in this paper are not exhaustive; new methods can be developed, especially if different assumptions are made on the data-generating process than we have assumed here. It would be interesting to see, for instance, if Markov-modulated demand processes can be adapted to handle a large number of feature information. Finally, extending our theoretical results for a more general classes of stochastic optimization problems remains an open task.

Acknowledgments

The authors thank Nicos Savva and Stefan Scholtes for providing the data for the case study. The authors also thank Paul Zipkin, Chung-Piaw Teo, the anonymous referees, and seminar attendees at a number of institutions and conferences for helpful suggestions.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19(6):716–723.

Arrow KJ, KarlinS, Scarf H (1958) *Studies in the Mathematical Theory of Inventory and Production*, Vol. 1 (Stanford University Press, Stanford, CA).

Azoury KS (1985) Bayes solution to dynamic inventory models under unknown demand distribution. *Management Sci.* 31(9):1150–1160.

Belloni A, Chernozhukov V (2011) ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* 39(1):82–130.

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

Burnetas AN, Smith CE (2000) Adaptive ordering and pricing for perishable products. *Oper. Res.* 48(3):436–443.

Chaudhuri P (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.* 19(2):760–777.

Chen X, Sim M, Sun P (2007) A robust optimization perspective on stochastic programming. *Oper. Res.* 55(6):1058–1071.

Chernozhukov V, Hansen C (2008) Instrumental variable quantile regression: A robust inference approach. *J. Econom.* 142(1):379–398.

Chernozhukov V, Fernández-Val I, Galichon A (2010) Quantile and probability curves without crossing. *Econometrica* 78(3):1093–1125.

Csörgö M (1983) *Quantile Processes with Statistical Applications* (SIAM, Philadelphia).

Devroye L, Wagner T (1979a) Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inform. Theory* 25(2):202–207.

Devroye L, Wagner T (1979b) Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory* 25(5):601–604.

Donnelly L, Mulhern M (2012) NHS pays £1,600 a day for nurses as agency use soars. *The Telegraph* (July 14), http://www.telegraph.co.uk/news/9400079/NHS-_pays-_1600-_a-_day-_for-_nurses-_as-_agency-_use-_soars.html.

Durrett R (2010) *Probability: Theory and Examples*, 4th edition (Cambridge University Press, New York).

Feldman RM (1978) A continuous review (s, s) inventory system in a random environment. *J. Appl. Probab.* 15(3):654–659.

Friedman J, Hastie T, Tibshirani R (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., (Springer Science+Business Media, New York).

Gallego G, Moon I (1993) The distribution free newsboy problem: Review and extensions. *J. Oper. Res. Soc.* 44(8):825–834.

Gallego G, Özer Ö (2001) Integrating replenishment decisions with advance demand information. *Management Sci.* 47(10):1344–1360.

Godfrey GA, Powell WB (2001) An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Sci.* 47(8):1101–1112.

Grant M, Boyd S. CVX: Matlab software for disciplined convex programming, version 2.0 beta. Accessed September 2013, http://cvxr.com/cvx.

Grant M, Boyd S (2008) Graph implementations for nonsmooth convex programs. Blondel V, Boyd S, Kimura H, eds. *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences (Springer, London), 95–110.

Graves SC, Meal HC, Dasu S, Qui Y (1986) Two-stage production planning in a dynamic environment. *Multi-Stage Production Planning and Inventory Control* (Springer, Berlin, Heidelberg), 9–43.

Green LV, Savin S, Savva N (2013) Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Sci.* 59(10):2237–2256.

Hannah L, Powell W, Blei DM (2010) Nonparametric density estimation for stochastic optimization with an observable state variable. Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Advances in Neural Information Processing Systems (NIPS 2010)* (Curran Associates, Red Hook, NJ), 820–828.

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts *i*th application to safety stock analysis in production/distribution systems. *IIE Trans.* 26(3):17–30.

Hofmann T, Schölkopf B, Smola AJ (2008) Kernel methods in machine learning. *Ann. Statist.* 36(3):1171–1220.

HuhWT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Math. Oper. Res.* 34(1):103–123.

Iida T, Zipkin PH (2006) Approximate solutions of a dynamic forecast-inventory model. *Manufacturing Service Oper. Management* 8(4):407–425.

Koenker R (2005) *Quantile Regression* (Cambridge University Press, New York).

Kunnumkal S, Topaloglu H (2008) Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Oper. Res.* 56(3):646–664.

Levi R, Perakis G, Uichanco J (2015) The data-driven newsvendor problem: New bounds and insights. *Oper. Res.* 63(6):1294–1306.

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* 32(4):821–839.

Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Oper. Res. Lett.* 33(4):341–348.

Lovejoy WS (1990) Myopic policies for some inventory models with uncertain demand distributions. *Management Sci.* 36(6):724–738.

Lovejoy WS (1992) Stopped myopic policies in some inventory models with generalized demand processes. *Management Sci.* 38(5):688–707.

Lu X, Song J-S, Regan A (2006) Inventory planning with forecast updates: Approximate solutions and cost error bounds. *Oper. Res.* 54(6):1079–1097.

Manton JH, Amblard P-O (2015) A primer on reproducing kernel Hilbert spaces. *Found. Trends Signal Process.* 8(1–2):1–126.

Nadaraya EA (1964) On estimating regression. *Theory Probab. Appl.* 9(1):141–142.

Parzen E (1979) Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* 74(365):105–121.

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Oper. Res.* 56(1):188–203.

Powell W, Ruszczyński A, Topaloglu H (2004) Learning algorithms for separable approximations of discrete stochastic optimization problems. *Math. Oper. Res.* 29(4):814–836.

Rockafellar RT (1997) *Convex Analysis* (Princeton University Press, Princeton, NJ).

Rogers WH, Wagner TJ (1978) A finite sample distribution-free performance bound for local discrimination rules. *Ann. Statist.* 6(3):506–514.

Scarf H (1959a) Bayes solutions of the statistical inventory problem. *Ann. Math. Statist.* 30(2):490–508.

Scarf H (1959b) The optimality of (s,S) policies in the dynamic inventory problem. *Mathematical Methods in the Social Science*, Arrow KJ, Karlin S, Suppes P, eds. (Stanford University Press, Stanford, CA).

Scarf H, Arrow KJ, Karlin S (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, Vol. 10 (Stanford University Press, Stanford CA) 201–209.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.* 6(2):461–464.

See C-T, Sim M (2010) Robust approximation to multiperiod inventory management. *Oper. Res.* 58(3):583–594.

Shapiro A, Dentcheva D, Ruszczyński AP (2009) *Lectures on Stochastic Programming: Modeling and Theory*, Vol. 9 (SIAM, Philadelphia).

Song J-S, Zipkin P (1993) Inventory control in a fluctuating demand environment. *Oper. Res.* 41(2):351–370.

Takeuchi I, Le QV, Sears TD, Smola AJ (2006) Nonparametric quantile estimation. *J. Mach. Learn. Res.* 7:(July): 1231–1264.

Vapnik VN (1998) *Statistical Learning Theory* (Wiley, New York).

Watson GS (1964) Smooth regression analysis. *Sankhyā: Indian J. Statist. Ser. A.* 26(4):359–372.

Gah-Yi Ban (formerly Gah-Yi Vahn) is an assistant professor of management science and operations at London Business School. Her research is in big data analytics, specifically decision making with complex, high dimensional, and/or highly uncertain data with applications to operations management and finance.

Cynthia Rudin is an associate professor of computer science, electrical and computer engineering, and statistics at Duke University. She is the recipient of the 2013 and 2016 INFORMS Innovative Applications in Analytics Awards, and an NSF CAREER award.