POLYPUBLIE

Bilevel optimization for feature selection in the data-driven newsvendor problem

Breno Serrano, Stefan Minner, Maximilian Schiffer, Thibaut Vidal

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2024
Type	Article de revue / Article
Version	Version officielle de l'éditeur / Published version
URL de PolyPublie PolyPublie URL	https://publications.polymtl.ca/57338/
Évaluation par les pairs Peer reviewed	Révisé par les pairs / Refereed
Conditions d'utilisation Terms of use	Creative Commons Attribution-Utilisation non commerciale-Pas d'oeuvre dérivée 4.0 International / Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND)
Référence Citation	Serrano, B., Minner, S., Schiffer, M., & Vidal, T. (2024). Bilevel optimization for feature selection in the data-driven newsvendor problem. European Journal of Operational Research, 315(2), 703-714. https://doi.org/10.1016/j.ejor.2024.01.025

Publication officielle
Official publication

Titre de la revue Journal title	European Journal of Operational Research (vol. 315, no. 2)
Maison d'édition Publisher	Elsevier BV
URL officielle Official URL	https://doi.org/10.1016/j.ejor.2024.01.025
Mention légale Legal notice	Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/)

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal
https://publications.polymtl.ca

European Journal of Operational Research 315 (2024) 703–714

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Analytics, Computational Intelligence and Information Management

Bilevel optimization for feature selection in the data-driven newsvendor problem

Breno Serrano [a,*], Stefan Minner [a,b], Maximilian Schiffer [a,b], Thibaut Vidal [c,d]

[a] School of Management, Technical University of Munich, Germany
[b] Munich Data Science Institute, Technical University of Munich, Germany
[c] Department of Mathematical and Industrial Engineering, Polytechnique Montréal, Canada
[d] Department of Computer Science, Pontifical Catholic University of Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Feature selection
Bilevel optimization
Newsvendor
Mixed integer programming

ABSTRACT

We study the feature-based newsvendor problem, in which a decision-maker has access to historical data consisting of demand observations and exogenous features. In this setting, we investigate feature selection, aiming to derive sparse, explainable models with improved out-of-sample performance. Up to now, state-of-the-art methods utilize regularization, which penalizes the number of selected features or the norm of the solution vector. As an alternative, we introduce a novel bilevel programming formulation. The upper-level problem selects a subset of features that minimizes an estimate of the out-of-sample cost of ordering decisions based on a held-out validation set. The lower-level problem learns the optimal coefficients of the decision function on a training set, using only the features selected by the upper-level. We present a mixed integer linear program reformulation for the bilevel program, which can be solved to optimality with standard optimization solvers. Our computational experiments show that the method accurately recovers ground-truth features already for instances with a sample size of a few hundred observations. In contrast, regularization-based techniques often fail at feature recovery or require thousands of observations to obtain similar accuracy. Regarding out-of-sample generalization, we achieve improved or comparable cost performance.

1. Introduction

The newsvendor problem and its variants have served as fundamental building blocks for models in inventory and supply chain management. In the classical newsvendor problem, a decision-maker optimizes the inventory of a perishable product that has a stochastic demand with a known distribution. However, having complete knowledge of the demand distribution is a strong assumption that does not hold in practice: often, the only information available is a limited set of historical data. Against this background, data-driven approaches became popular and strive to use past demand data to inform the newsvendor's ordering decisions.

In this context, we study the feature-based newsvendor problem (cf. Ban and Rudin 2019, Beutel and Minner 2012) in which the decision-maker has access not only to historical demand observations but also to a set of feature variables – often referred to as contextual information or covariates – that may provide partial information about future realizations of the uncertain demand. For example, consider a retail company that sells products in stores at different locations and with different assortments, e.g., newspaper stands, a chain of restaurants, or fashion retail stores. Then, features could include information about store location, product characteristics, weather forecasts, e.g., temperature and expected precipitation, day of the week, seasonal trends, social media events, competitor activity, epidemic outbreaks, supplier promotions, regulatory changes, and information about holidays or special events close to the store location. In a broader sense, feature variables can include any information that is available to the decision-maker before or at the time instant when ordering decisions are made, e.g., past sales of related products to capture correlated demands, or forecasts based on human expert knowledge. In the context of a car retailer, Tian and Zhang (2023) applied a feature-based newsvendor model where features included historical sales data, textual online reviews, search traffic data, and macroeconomic indicator data. Moreover, the newsvendor model has applications beyond inventory management, such as finding the optimal staffing level of nurses for hospital emergency rooms (Ban & Rudin, 2019), where the uncertain

* Corresponding author.
E-mail addresses: breno.serrano@tum.de (B. Serrano), stefan.minner@tum.de (S. Minner), schiffer@tum.de (M. Schiffer), thibaut.vidal@polymtl.ca (T. Vidal).

https://doi.org/10.1016/j.ejor.2024.01.025
Received 28 June 2023; Accepted 18 January 2024
Available online 24 January 2024
0377-2217/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

demand corresponds to the number of arriving patients and features include the day of the week and time of the day, among others.

Companies nowadays have large amounts of data that are used to train machine learning models with the aim of improving operational decisions. In practice, such models often suffer from overfitting to the training data, or lack explainability, which is crucial, e.g., when dealing with high-stakes decisions. In this setting, selecting a subset of the available features can lead to sparser, more explainable models with improved out-of-sample performance. Against this background, we investigate the challenge of feature selection (cf. Kuhn and Johnson 2019, Molina et al. 2002): given a data set with a possibly large set of feature variables, we aim to learn a linear decision function for the feature-based newsvendor that can generalize to out-of-sample data, utilizing only relevant features.

The goal of this paper is to propose an approach to feature selection based on bilevel optimization. Accordingly, we learn a linear decision function for the feature-based newsvendor based on a training data set, where the set of features available for training is restricted to a given subset of all features. The problem of learning a linear decision function corresponds to the lower-level problem of our bilevel optimization formulation, while the problem of selecting a subset of available features corresponds to the upper-level problem. In the upper-level problem, we search for a subset of features that minimizes the out-of-sample cost measured on a held-out data set, which we denote as the validation data set. In the remainder of this section, we first review related literature before we detail our contribution and describe the organization of this paper.

1.1. Related works

Our work relates to the fields of data-driven optimization for the newsvendor problem, and more broadly to prescriptive analytics, machine learning, and bilevel programming. We briefly review the most related papers in the following.

Newsvendor problem. Research on the newsvendor problem often assumed a decision-maker with full knowledge about the demand distribution, and considered various settings, e.g., with different objectives or utility functions (Chen et al., 2007; Wang & Webster, 2009), pricing policies (Petruzzi & Dada, 1999), and multi-product or multi-period settings (Kogan & Lou, 2003; Lau & Lau, 1996). For general surveys on newsvendor models and extensions, we refer the interested reader to Khouja (1999), Qin et al. (2011) and Choi (2012). In practice, the decision-maker often has only a finite set of demand observations and cannot estimate the true underlying distribution, which motivated works on the distribution-free newsvendor problem. In this context, the seminal work of Scarf (1958) derived the optimal order quantity that maximizes profit against the worst-case demand distribution, assuming that only the mean and variance of demand are known. For a review on the distribution-free newsvendor and extensions thereof, we refer to Gallego and Moon (1993), Moon and Gallego (1994), and Yue et al. (2006). Later works on this problem variant assumed additional information about the demand distribution, such as percentiles (Gallego et al., 2001), symmetry, and unimodality (Perakis & Roels, 2008).

In contrast to working with moments or distributional parameters, data-driven approaches build directly upon a sample of available data that reflects realizations of the underlying uncertainty. In this context, a common solution approach is sample average approximation (SAA) (cf. Kleywegt et al. 2002, Shapiro 2003). Levi et al. (2007) applied SAA for the single-period featureless newsvendor problem and established upper bounds on the number of samples required to achieve a specified relative error. In this course, Levi et al. (2015), Cheung and Simchi-Levi (2019), and Besbes and Mouchtaki (2023) further improved upon previous SAA bounds. Besbes and Muharremoglu (2013), Sachs and Minner (2014), and Ban (2020) studied the impact of demand censoring, i.e., a problem variant in which only *sales* observations are available but excess demand is not recorded. They derived upper and lower bounds on the difference between the cost achieved by a policy and the optimal cost with knowledge of the demand distribution. Adopting a robust optimization perspective, Bertsimas and Thiele (2005) proposed a data-driven approach that can be reformulated as a linear program (LP) and trades off higher profits for a decrease in the downside risk. Robust optimization approaches were also investigated by Bertsimas and Thiele (2006) and See and Sim (2010) for a multi-period inventory problem. Finally, many authors applied data-driven distributionally robust approaches for dealing with uncertainty in the context of multi-item newsvendor problems (see, e.g., Ben-Tal et al. 2013, Bertsimas et al. 2018, Hanasusanto et al. 2015, Wang et al. 2016).

Despite numerous extensions to the newsvendor problem, most data-driven approaches consider only demand data but no feature variables to be available. However, ignoring the presence of features can lead to inconsistent decisions as shown in Ban and Rudin (2019). In the following, we review papers that also consider the presence of features in the context of data-driven optimization.

Data-driven optimization. Beyond the newsvendor problem, some recent works have studied the integration of estimation and optimization. In particular, Bertsimas and Kallus (2020) proposed a framework for feature-based stochastic optimization problems based on a weighted SAA approach, in which the weights are generated by machine learning methods, such as k-nearest neighbors regression, local linear regression, classification and regression trees, or random forests. Elmachtoub and Grigas (2021) focused on problems with a linear objective and used features to learn a prediction model for the stochastic cost vector. They proposed a modified loss function that directly leverages the structure of the optimization problem instead of minimizing a standard prediction error, such as the least squares loss. Despite this modification, their approach still handles prediction and optimization as separate tasks and does not integrate them into a one-step process. Mandi et al. (2020) further adapted the approach from Elmachtoub and Grigas (2021) to solve some hard combinatorial problems, e.g., by proposing tailored warm-starting techniques.

In the context of the feature-based newsvendor, Beutel and Minner (2012) integrated estimation and optimization by learning a decision function that predicts ordering decisions directly from features, opposed to first estimating the demand and then optimizing the inventory level. The proposed model formulation is an LP that solves an Empirical Risk Minimization (ERM) problem over a training data set. Oroojlooyjadid et al. (2020) and Zhang and Gao (2017) applied neural networks to the newsvendor problem, proposing specific loss functions that consider the impact of inventory shortage and holding costs. Huber et al. (2019) provided an empirical evaluation of different data-driven approaches for the feature-based newsvendor and compared their performance against *model-based* approaches, which model the uncertainty through a demand distribution assumption. Their experiments on real-world data showed that data-driven approaches outperform their model-based counterparts in most cases. Further, Mandl and Minner (2023) studied a multi-period commodity procurement problem under price uncertainty and proposed a data-driven model to derive optimal purchase policies based on economic indicators.

Regarding feature selection, Ban and Rudin (2019) extended the model of Beutel and Minner (2012) by including a regularization term to the objective function, which penalizes the complexity of the solution, thereby favoring the selection of fewer features. However, feature selection is not the main focus of Ban and Rudin (2019), and an open challenge remains regarding the specification of the regularization parameter, for which heuristics are often employed. In this work, we avoid regularization by formalizing the task of feature selection as a bilevel optimization problem for which we provide a tractable single-level reformulation.

Bilevel optimization in machine learning. Bilevel optimization has been applied in the field of machine learning for hyperparameter optimization (Bennett et al., 2006, 2008; Franceschi et al., 2018; Mackay

et al., 2019) and feature selection (Agor & Özaltın, 2019). In particular, Bennett et al. (2006, 2008) proposed a bilevel program for optimizing the hyperparameters of a support vector regression model. They reformulated the model into a single-level nonlinear program and employed off-the-shelf solvers based on Sequential Quadratic Programming (Fletcher & Leyffer, 2002). Franceschi et al. (2018) also proposed a bilevel programming approach for hyperparameter optimization, highlighting connections to meta-learning, and solved it with a gradient-based method.

Only Agor and Özaltın (2019) addressed feature selection as a bilevel optimization problem in the context of classification models, e.g., Lasso-based logistic regression and support vector machines. However, their bilevel formulations do not apply to our problem setting, since the feature-based newsvendor combines aspects from supervised learning, i.e., regression, and data-driven optimization. Moreover, the solution method of Agor and Özaltın (2019) consists of a tailored genetic algorithm, which does not provide solution-quality guarantees. In contrast, our methodology is based on mixed integer linear programming (MILP) and allows to optimally solve the proposed bilevel programming formulations.

Mixed integer optimization for feature selection. The problem of feature selection, also referred to as the best subset selection problem, has been extensively studied in machine learning research. For an introduction to different methods, we refer to Molina et al. (2002), Guyon and Elisseeff (2003), and Kuhn and Johnson (2019). Recent works proposed mixed integer programming (MIP) formulations for feature selection in the context of multiple linear regression (Takano & Miyashiro, 2020), classification with support vector machines (Maldonado et al., 2014), and cluster analysis (Benati & García, 2014), among others. In particular, Maldonado et al. (2014) proposed a MILP formulation that simultaneously learns the classifier and selects relevant features by limiting the number of selected features using a budget constraint.

For regression models, Miyashiro and Takano (2015) proposed a mixed integer second-order cone programming formulation for feature selection with respect to various statistical criteria. Gómez and Prokopyev (2021) showed that a formulation based on mixed integer fractional programming has a stronger convex relaxation than (Miyashiro & Takano, 2015). Bertsimas et al. (2016) considered a MIP formulation for linear regression under a cardinality constraint on the subset of selected features, and proposed discrete first-order algorithms and tailored warm-starting techniques. Park and Klabjan (2020) solved a mixed integer quadratically constrained program for feature selection with respect to criteria such as the mean squared error and the mean absolute error. Kimura and Waki (2018) proposed a mixed integer nonlinear program and a tailored branch-and-bound algorithm for minimizing the Akaike information criterion (AIC) criterion. However, many statistical criteria designed for linear regression are not meaningful in the context of feature-based stochastic optimization, as we are interested in the impact of feature selection in the downstream newsvendor problem. Therefore, we adopt a cross-validation criterion that evaluates the out-of-sample cost of a selection of features on a held-out validation data set, which requires only mild assumptions in contrast to other information criteria (cf. Takano and Miyashiro 2020).

Related work by Takano and Miyashiro (2020) proposed a bilevel optimization formulation for feature selection based on a cross-validation criterion. They focused on multiple linear regression and proposed a single-level reformulation based on mixed integer quadratic programming (MIQP). Due to its computational complexity, they cannot optimally solve the MIQP in most cases. A key difference to our work is that they rely on regularization in the lower-level problem and require a regularization parameter value to be given a priori. In the experiments, they employ a grid search method to tune this parameter and leave it for future research to devise a formulation that simultaneously selects features and the optimal regularization value. As we discuss later, the problem of hyperparameter tuning itself can be cast as a bilevel optimization problem for which currently no tractable reformulation exists. In contrast, we propose a bilevel program for feature selection which does not require regularization and has a MILP single-level reformulation.

1.2. Contribution

We close the research gaps outlined above by proposing a novel bilevel optimization model that directly incorporates feature selection into solving the data-driven newsvendor problem. Specifically, our contribution is fourfold. First, we introduce a bilevel program designed for feature selection, which we denote the Bilevel Feature Selection (BFS) model. In contrast to regularization-based methods, which penalize the norm of the solution vector, BFS captures the more intuitive notion of selecting a subset of features that minimizes an estimate of the out-of-sample cost, measured on a held-out validation set. We reformulate the bilevel program into a single-level optimization problem, which we solve to optimality with off-the-shelf optimization solvers. Second, we extend the BFS model to accommodate cross-validation strategies, which further improves its solution quality. Third, to illustrate the drawback of regularization-based methods for feature selection, we present a bilevel program, which we refer to as Bilevel Hyperparameter Optimization (BHO), that searches for the optimal hyperparameter for the regularized ERM model (cf. Ban and Rudin 2019). BHO formally describes the optimization model that established hyperparameter optimization methods implicitly solve by means of heuristics, such as grid search, random search, or Bayesian optimization. Fourth, we conduct extensive numerical experiments, using synthetic instances with correlated features. We compare the proposed BFS models against regularization-based methods in terms of out-of-sample performance and ground-truth feature recovery. We further compare the methods' behavior under demand misspecification, assuming a nonlinear demand model. Our results show that the proposed BFS approach consistently achieves higher accuracy in feature recovery. In most cases, we also observe an improvement in out-of-sample cost performance, i.e., a decrease in test cost.

1.3. Organization

The remainder of this paper is structured as follows. In Section 2, we review the model formulations for the classical newsvendor and the feature-based newsvendor problem. Section 3 presents the BHO and the BFS models, and consecutively extends the BFS to cross-validation. Section 4 describes our experimental design, and Section 5 presents the results comparing the proposed method against state-of-the-art techniques based on regularization. Section 6 concludes this paper and gives an outlook on future research.

2. Fundamentals

In the classical newsvendor problem, a risk-neutral decision-maker sets the order quantity of a product before observing its uncertain demand. Here, the objective is to minimize the expected cost:

$$\min_{q\geq 0} \mathbb{E}\left[C(q;d)\right], \tag{1}$$

where q is the order quantity, $d \sim D$ is the random variable representing the uncertain demand,

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

is the cost of ordering q units and observing demand d, based on the per unit shortage cost b for lost profits and unit holding cost h, corresponding to the procurement cost of unsold products discounted by their unit salvage value. If the demand distribution is known, then the optimal decision q^* is given at the $b/(b+h)$ quantile of its cumulative distribution function.

In practice, the demand distribution is often not known. We consider the feature-based newsvendor problem, in which the decision-maker has access to historical demand data and contextual information given by a set of feature variables $\mathbf{x} \in \mathbb{R}^{m+1}$ (cf. Beutel and Minner 2012). Here, the uncertain demand d and feature variables $\mathbf{x}$ follow an (unknown) joint probability distribution $(\mathbf{x}, d) \sim \mathcal{X} \times \mathcal{D} = \mathcal{Z}$. The decision-maker's objective is to minimize the expected cost conditioned on the observed features:

$$\min_{q \geq 0} \mathbb{E}\left[C(q(\mathbf{x}); d(\mathbf{x}))|\mathbf{x}\right]. \tag{3}$$

One approach to solve the feature-based newsvendor is to separate the estimation and optimization problems, i.e., one first estimates the conditional demand distribution from historical data and then optimizes the order quantity based on new feature observations. One drawback of this approach is that the first step's estimation problem does not account for the asymmetry in the newsvendor cost function, related to under- and over-predicting demand. To address this issue, Beutel and Minner (2012) proposed to integrate estimation and optimization into a one-step process, by introducing a linear decision function that maps feature observations directly to ordering decisions. To learn the optimal coefficients of the decision function, one minimizes the empirical cost over a data set with demand and feature observations.

Let $\left\{(\mathbf{x}_i, d_i)\right\}_{i \in S}$ be a data set indexed by $S = \{1, \ldots, n\}$, where $\mathbf{x}_i$ is an $(m+1)$-dimensional feature vector and d_i is a scalar demand observation. Let $J = \{0, \ldots, m\}$ denote the set of feature indices. We assume that $x_i^0 = 1$ represents the feature-independent intercept term, for all $i \in S$. In this setting, Beutel and Minner (2012) consider a linear decision function of the form:

$$q(\mathbf{x}) = \beta^0 + \sum_{j=1}^{m} \beta^j x^j = \boldsymbol{\beta}^\top \mathbf{x}, \tag{4}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$ is the parameter vector, whose values are learned by minimizing the empirical cost on data set S. Upon observing new feature values, the decision-maker can then directly decide upon the order quantity instead of first estimating the uncertain demand.

Since the learned decision function may overfit to the in-sample data set S, it is common practice in machine learning to evaluate the out-of-sample generalization on a separate test data set S_{test}. To avoid overfitting and improve the out-of-sample generalization, Ban and Rudin (2019) proposed an extension of Beutel and Minner (2012) by integrating a regularization term into the loss function. Accordingly, the objective comprises a trade-off between minimizing the empirical in-sample cost and the regularization term, with a constant hyperparameter balancing these two terms:

$$\text{(ERM-}\ell_p\text{)} \qquad \min_{\boldsymbol{\beta}} \ \frac{1}{|S|} \sum_{i \in S} C(q_i\,; d_i) + \lambda \left\|\boldsymbol{\beta}\right\|_p \tag{5}$$

$$\text{s.t.} \quad q_i = \boldsymbol{\beta}^\top \mathbf{x}_i \qquad \forall i \in S, \tag{6}$$

where $\lambda \geq 0$ is the regularization hyperparameter and $\|\boldsymbol{\beta}\|_p$ is the ℓ_p-norm of the vector $\boldsymbol{\beta}$. Depending on the choice of p in the regularization, the resulting model may be a MILP, an LP, or a second-order cone program, for ℓ_0, ℓ_1, and ℓ_2-norm regularization, respectively. Effectively, regularization enables feature selection by penalizing the complexity of the solution, thereby favoring sparse solution vectors.

3. Methodology

We start this section presenting the BHO model, which incorporates hyperparameter fitting in (ERM-ℓ_p). Then, we introduce the BFS model as an alternative bilevel program that avoids regularization.

3.1. Bilevel hyperparameter optimization (BHO)

In Section 2, we assumed the hyperparameter λ as introduced in (ERM-ℓ_p) to be given. However, identifying λ constitutes a challenge in itself as a respective misspecification can significantly reduce cost performance. To parametrize λ correctly, one may utilize existing techniques for hyperparameter optimization, which partition the original data set S into a training set T and a validation set V. On the training set, one learns the model parameters for a fixed hyperparameter value. Using the validation set, one can then assess the cost of the trained model for a variety of hyperparameter values, to finally choose the value λ^* that leads to a minimum cost on the validation set. Next, we present the BHO formulation, which models the search for the optimal hyperparameter λ^* as a bilevel optimization problem.

We introduce variables u_i to model the inventory shortage and variables o_i to model the surplus inventory at the end of period $i \in T \cup V$. In the following bilevel programming formulation, the upper-level (UL) problem searches for an optimal regularization value $\lambda^* \geq 0$ that minimizes cost on the validation set V. In turn, the lower-level (LL) problem solves the feature-based newsvendor, as stated in (ERM-ℓ_p), on the training set T:

$$\text{(BHO-}\ell_p\text{ UL)} \qquad C^*_{\text{BHO}} = \min \ \frac{1}{|V|} \sum_{i \in V} \left(bu_i + ho_i\right) \tag{7}$$

$$\text{s.t.} \quad u_i \geq d_i - \boldsymbol{\beta}^\top \mathbf{x}_i \qquad \forall i \in V \tag{8}$$

$$o_i \geq \boldsymbol{\beta}^\top \mathbf{x}_i - d_i \qquad \forall i \in V \tag{9}$$

$$u_i \geq 0,\ o_i \geq 0 \qquad \forall i \in V \tag{10}$$

$$\lambda \geq 0 \tag{11}$$

$$\boldsymbol{\beta} \in \Omega_p(\lambda), \tag{12}$$

where $\Omega_p(\lambda)$ is the set of optimal solutions $\boldsymbol{\beta}$ to the lower-level problem, parameterized by λ:

$$\text{(BHO-}\ell_p\text{ LL)} \qquad \Omega_p(\lambda) := \arg\min \ \frac{1}{|T|} \sum_{i \in T} \left(bu_i + ho_i\right) + \lambda \left\|\boldsymbol{\beta}\right\|_p^2 \tag{13}$$

$$\text{s.t.} \quad u_i \geq d_i - \boldsymbol{\beta}^\top \mathbf{x}_i \qquad \forall i \in T \tag{14}$$

$$o_i \geq \boldsymbol{\beta}^\top \mathbf{x}_i - d_i \qquad \forall i \in T \tag{15}$$

$$\boldsymbol{\beta} \in \mathbb{R}^{m+1} \tag{16}$$

$$u_i \geq 0,\ o_i \geq 0 \qquad \forall i \in T \tag{17}$$

The upper-level objective (7) minimizes the newsvendor cost on the validation set V and the lower-level objective (13) minimizes the regularized newsvendor cost on the training set T. Constraints (8) and (14) define the inventory shortage for period $i \in V$ and $i \in T$, respectively, given the decision function parameterized by $\boldsymbol{\beta}$. Constraints (9) and (15) define the surplus inventory at period $i \in V$ and $i \in T$. Constraints (10), (11), (12), (16), and (17) define the variable domains.

So far, we define the BHO formulation in (7)–(17) for a general ℓ_p-norm, which leads to a different model for different p. In the following, we illustrate some properties of BHO under the special case of the ℓ_0-norm regularization, which minimizes the number of non-zero elements in the $\boldsymbol{\beta}$ vector. In this case, we introduce the binary variable z^j to indicate whether coefficient β^j is non-zero. The lower-level problem can then be formulated as a MIP:

$$\text{(BHO-}\ell_0\text{ LL)} \qquad \Omega_0(\lambda) := \arg\min \ \frac{1}{|T|} \sum_{i \in T} \left(bu_i + ho_i\right) + \lambda \sum_{j \in J} z^j \tag{18}$$

$$\text{s.t.} \quad (14)\text{–}(17)$$

$$\beta^j = 0 \text{ if } z^j = 0 \qquad \forall j \in J \tag{19}$$

$$z^j \in \{0, 1\} \qquad \forall j \in J, \tag{20}$$

where Constraints (14)–(17) define the shortage and surplus inventory and Constraints (19) enforce that $\beta^j = 0$ if the corresponding feature is not selected.

The BHO formulation (7)–(17) generalizes many common methods for hyperparameter optimization. To avoid the high computational effort of solving the BHO model to optimality, existing methods relax the assumption that λ can take any value in $\mathbb{R}_{\geq 0}$, and consider a finite support set $\Lambda \subseteq \mathbb{R}_{\geq 0}$ instead (Bergstra & Bengio, 2012; Bergstra et al., 2013). For example, suppose the values in Λ are equally spaced along a grid, i.e., a line segment, then the resulting model corresponds to the well-known *grid search* method. If the values in Λ are randomly selected in a closed region, then the formulation describes the *random search* method. Other approaches, e.g., based on Bayesian optimization, would perform an adaptive search, iteratively selecting a value λ for the upper-level variable and then optimizing the lower-level problem. The iterative selection of new values for λ depends on the validation performance of previously selected points. In essence, current methods for hyperparameter optimization, such as the examples described above, are heuristics that avoid solving the BHO model to optimality.

3.2. Bilevel feature selection (BFS)

To remedy the drawback of BHO, we introduce a bilevel programming formulation specifically designed for feature selection. Instead of penalizing the number of selected features, we propose a more intuitive model, in which the upper-level problem selects a subset of features that minimize the empirical cost on a validation set. We then reformulate the resulting model into a single-level problem, which is computationally more tractable, and finally compare the proposed BFS and BHO models.

Consider our original data set S, which we partition into a training set T and a validation set V. We introduce binary variables z^j, for $j \in J$, to indicate whether feature j is marked as relevant ($z^j = 1$) or not ($z^j = 0$). In the upper-level, BFS selects a subset of features that minimizes the empirical cost on the validation set V. The lower-level problem then learns the optimal coefficients of the decision function in the training set T by solving the ERM model using only the features selected in the upper-level. We formulate the resulting upper-level problem as follows:

$$\text{(BFS UL)} \qquad C^*_{\text{BFS}} = \min \quad \frac{1}{|V|}\sum_{i\in V}\left(bu_i + ho_i\right) \tag{21}$$

$$\text{s.t.} \quad (8)\text{–}(10)$$

$$z^j \in \{0,1\} \qquad \forall j \in J \tag{22}$$

$$\boldsymbol{\beta} \in \Pi_0(\mathbf{z}), \tag{23}$$

where $\Pi_0(\mathbf{z})$ is the set of optimal solutions $\boldsymbol{\beta}$ to the lower-level problem:

$$\text{(BFS LL)} \qquad \Pi_0(\mathbf{z}) := \arg\min \quad \frac{1}{|T|}\sum_{i\in T}\left(bu_i + ho_i\right) \tag{24}$$

$$\text{s.t.} \quad (14)\text{–}(17)$$

$$\beta^j = 0 \text{ if } z^j = 0 \qquad \forall j \in J \tag{25}$$

The upper and lower-level objectives (21) and (24) minimize the newsvendor cost on the validation set V and training set T, respectively. Constraints (8)–(10) and (14)–(17) define the shortage and surplus inventory. Constraints (22)–(23) define the variable domains and Constraints (25) ensure that $\beta^j = 0$ if feature j is not selected.

We reformulate Model (21)–(25) by substituting the lower-level problem by its Karush-Kuhn–Tucker (KKT) conditions (cf. Cao and Chen 2006, Fontaine and Minner 2014). We introduce the dual variables μ_i, and γ_i corresponding to constraints (14) and (15) of the lower-level problem. The equivalent single-level (SL) optimization problem can then be expressed by using indicator constraints:

$$\text{(BFS SL)} \qquad \min \quad \frac{1}{|V|}\sum_{i\in V}\left(bu_i + ho_i\right) \tag{26}$$

$$\text{s.t.} \quad (8)\text{–}(10),\ (14)\text{–}(17),\ (22),\ (25)$$

$$\frac{1}{|T|}\sum_{i\in T}(bu_i + ho_i) \leq \sum_{i\in T}(\gamma_i - \mu_i)d_i \tag{27}$$

$$\mu_i + \frac{b}{|T|} \geq 0 \qquad \forall i \in T \tag{28}$$

$$\gamma_i + \frac{h}{|T|} \geq 0 \qquad \forall i \in T \tag{29}$$

$$\sum_{i\in T}(\mu_i - \gamma_i)x_i^j = 0 \text{ if } z^j = 1 \qquad \forall j \in J \tag{30}$$

$$\mu_i \leq 0,\ \gamma_i \leq 0 \qquad \forall i \in T \tag{31}$$

As before, Constraints (8)–(10) and (14)–(17) define the shortage and surplus inventory. Constraints (22) and (25) model the selection of features. Constraint (27) represents the optimality condition of the lower-level problem, by comparing its primal objective value with the corresponding dual objective value. Constraints (28) and (29) are the dual constraints of the lower-level problem associated with primal variables u_i and o_i for $i \in T$. Constraints (30) are the dual constraints related to the primal variables β^j for $j \in J$, and Constraints (31) define the domain of the dual variables. The single-level reformulation has $2n + 2|T| + 2|J|$ variables and $2n + 2|T| + 2|J| + 1$ constraints.

The BFS model shares some similarities with the BHO model. Both models have the same upper-level objective and the lower-level objectives differ only in the presence of the regularization term. We provide an overview of the main properties of both models in Table 1. The main advantage of the BFS model regarding tractability is due to the existence of binary variables being limited to the upper-level problem. Consequently, we can reformulate the BFS model into a MILP and leverage existing exact methods to find optimal solutions at a limited scale or heuristic methods to efficiently solve larger instances.

Moreover, the following results show that the optimal cost of the BFS model is a lower bound to the optimal cost of the BHO model when adopting ℓ_0-norm regularization.

Lemma 1. *Given a fixed selection of features* $\mathbf{z}$ *for both BHO and BFS, i.e.,* $\mathbf{z}_{BHO} = \mathbf{z}_{BFS} = \mathbf{z}'$ *(assuming that* $\mathbf{z}'$ *is feasible for both problems), solving the remaining problems for the rest of the decision variables yields optimal solutions* $\boldsymbol{\beta}_{BHO}\big|_{\mathbf{z}=\mathbf{z}'} = \boldsymbol{\beta}_{BFS}\big|_{\mathbf{z}=\mathbf{z}'}$, *with costs* $C_{BHO}\big|_{\mathbf{z}=\mathbf{z}'} = C_{BFS}\big|_{\mathbf{z}=\mathbf{z}'}$.

Proof. By fixing $\mathbf{z}_{\text{BHO}} = \mathbf{z}'$, the regularization term in the lower-level objective becomes constant and λ does not influence the optimal solution. Therefore, we can ignore regularization and the lower-level problem of the BHO becomes equal to the lower-level problem of the BFS, leading to $\boldsymbol{\beta}_{\text{BHO}}\big|_{\mathbf{z}=\mathbf{z}'} = \boldsymbol{\beta}_{\text{BFS}}\big|_{\mathbf{z}=\mathbf{z}'}$ as the optimal solution. Since the upper-level objectives are equal in both models, the optimal costs will be equal: $C_{\text{BHO}}\big|_{\mathbf{z}=\mathbf{z}'} = C_{\text{BFS}}\big|_{\mathbf{z}=\mathbf{z}'}$. □

Proposition 1. *The optimal cost of the BFS model is a lower bound for the optimal cost of the BHO model with* ℓ_0*-norm regularization:* $C^*_{BHO} \geq C^*_{BFS}$.

Proof. Let $\mathbf{z}_{\text{BHO}} = \mathbf{z}^*$ be the optimal selection of features according to BHO with cost C^*_{BHO}. Suppose that $\mathbf{z}_{\text{BFS}} = \mathbf{z}'$ is a solution to BFS, such that $C^*_{\text{BHO}} < C_{\text{BFS}}\big|_{\mathbf{z}=\mathbf{z}'}$. We can always improve the cost of BFS by setting $\mathbf{z}_{\text{BFS}} = \mathbf{z}^*$ in the upper-level problem. Because of Lemma 1, this will result in a new solution with cost $C_{\text{BFS}}\big|_{\mathbf{z}=\mathbf{z}^*} = C^*_{\text{BHO}} \geq C^*_{\text{BFS}}$. □

3.3. Bilevel feature selection with cross-validation (BFS-CV)

Cross-validation strategies often improve the generalization ability of machine learning models and prevent overfitting by using data re-sampling methods. Accordingly, we extend the BFS model to cross-validation instead of simple hold-out validation. We consider K training-validation splits of the data and search for the set of features that minimize the average cost over all K validation sets. For each $k \in [K] = \{1, \ldots, K\}$, we consider a subset of observations $S_k \subseteq S$ sampled from the original data set S. Analogously to BFS, we partition the set S_k into a training set T_k and a validation set V_k. We introduce variables u_{ik} and o_{ik} to model the inventory shortage and surplus,

Table 1
Comparison between BFS and BHO with ℓ_0-norm regularization.

Formulation		BHO (ℓ_0-norm reg.)	BFS
Upper-level	Objective	minimize validation cost	minimize validation cost
	Variables	$\lambda \in \mathbb{R}_{\geq 0}$	$\mathbf{z} \in \{0,1\}^{m+1}$
Lower-level	Objective	minimize training cost + regularization	minimize training cost
	Variables	$\boldsymbol{\beta} \in \mathbb{R}^{m+1}$, $\mathbf{z} \in \{0,1\}^{m+1}$	$\boldsymbol{\beta} \in \mathbb{R}^{m+1}$

respectively, at the end of period $i \in T_k \cup V_k$ for each training-validation split $k \in [K]$. We then learn the model parameters $\boldsymbol{\beta}_k \in \mathbb{R}^{m+1}$ using the corresponding training set T_k, and select features by minimizing the average validation cost over all validation sets V_k for $k \in [K]$. The resulting problem is a bilevel program with K lower-level problems:

$$\text{(BFS-CV UL)} \quad \min \ \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\left(bu_{ik}+ho_{ik}\right) \tag{32}$$

$$\text{s.t.} \quad u_{ik} \geq d_i - \boldsymbol{\beta}_k^\top \mathbf{x}_i \qquad \forall k \in [K], \forall i \in V_k \tag{33}$$

$$o_{ik} \geq \boldsymbol{\beta}_k^\top \mathbf{x}_i - d_i \qquad \forall k \in [K], \forall i \in V_k \tag{34}$$

$$u_{ik} \geq 0,\ o_{ik} \geq 0 \qquad \forall k \in [K], \forall i \in V_k \tag{35}$$

$$z^j \in \{0,1\} \qquad \forall j \in J \tag{36}$$

$$\boldsymbol{\beta}_k \in \Pi_k(\mathbf{z}) \qquad \forall k \in [K], \tag{37}$$

where $\Pi_k(\mathbf{z})$ is the set of optimal solutions corresponding to the kth lower-level problem:

$$\text{(BFS-CV LL)} \quad \Pi_k(\mathbf{z}) := \arg\min \ \frac{1}{|T_k|}\sum_{i\in T_k}\left(bu_{ik}+ho_{ik}\right) \tag{38}$$

$$\text{s.t.} \quad u_{ik} \geq d_i - \boldsymbol{\beta}_k^\top \mathbf{x}_i \qquad \forall i \in T_k \tag{39}$$

$$o_{ik} \geq \boldsymbol{\beta}_k^\top \mathbf{x}_i - d_i \qquad \forall i \in T_k \tag{40}$$

$$u_{ik} \geq 0,\ o_{ik} \geq 0 \qquad \forall i \in T_k \tag{41}$$

$$\beta_k^j = 0 \text{ if } z^j = 0 \qquad \forall j \in J \tag{42}$$

$$\boldsymbol{\beta}_k \in \mathbb{R}^{m+1} \tag{43}$$

Constraints (33)–(35) and (39)–(41) define the shortage and surplus inventory for period $i \in V_k$ and $i \in T_k$, respectively for each split $k \in [K]$. Constraints (36), (37), and (43) define the variable domains and Constraints (42) ensure that $\beta_k^j = 0$ if feature j is not selected for the training-validation split k.

The above model can accommodate different cross-validation strategies, such as K-fold, random permutations (Shuffle & Split), or Leave-P-Out cross-validation (see, e.g., Arlot and Celisse 2010, Hastie et al. 2009). Each particular choice of cross-validation strategy corresponds to a different approach for constructing the subsets S_k and partitioning the data into T_k and V_k. Moreover, the special case with $K = 1$ corresponds to the previously introduced BFS model.

Analogously to BFS, Bilevel Feature Selection with cross-validation (BFS-CV) can be reformulated into a single-level MILP:

$$\text{(BFS-CV SL)} \quad \min \ \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\left(bu_{ik}+ho_{ik}\right) \tag{44}$$

$$\text{s.t.} \quad \text{(33)–(36), (39)–(43)}$$

$$\frac{1}{|T_k|}\sum_{i\in T_k}(bu_{ik}+ho_{ik}) \leq \sum_{i\in T_k}(\gamma_{ik}-\mu_{ik})d_i \qquad \forall k \in [K] \tag{45}$$

$$\mu_{ik} + \frac{b}{|T_k|} \geq 0 \qquad \forall k \in [K], \forall i \in T_k \tag{46}$$

$$\gamma_{ik} + \frac{h}{|T_k|} \geq 0 \qquad \forall k \in [K], \forall i \in T_k \tag{47}$$

$$\sum_{i\in T_k}(\mu_{ik}-\gamma_{ik})x_i^j = 0 \text{ if } z^j = 1 \qquad \forall k \in [K], \forall j \in J \tag{48}$$

$$\mu_{ik} \leq 0,\ \gamma_{ik} \leq 0 \qquad \forall k \in [K], \forall i \in T_k \tag{49}$$

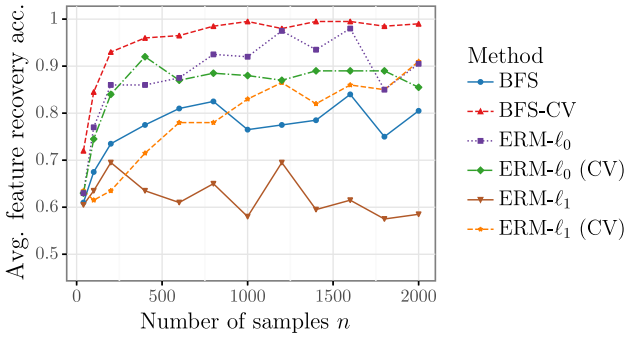


Fig. 1. Impact of instance size on the accuracy of feature recovery.

Constraints (45) represents the optimality condition of the lower-level problem, for each training-validation split $k \in [K]$. Constraints (46) and (47) are the dual constraints of the lower-level problem associated with primal variables u_{ik} and o_{ik} for $i \in T_k$ for $k \in [K]$. Constraints (48) are the dual constraints related to the primal variables $\boldsymbol{\beta}^k$ for $k \in [K]$, and Constraints (49) define the domain of the dual variables. The single-level reformulation has $K(4|T| + 2|V| + |J|) + |J|$ variables and $K(4|T| + 2|V| + 2|J| + 1)$ constraints.

4. Experimental design

To benchmark our approaches, we perform extensive computational experiments on synthetic instances. The goals of our computational experiments are fourfold.

(i) We evaluate the performance of a MILP approach for the proposed BFS and BFS-CV models in terms of feature selection and generalization to out-of-sample data;
(ii) We compare the performance of our approach against existing regularization-based methods;
(iii) For each method, we compare the performance of hold-out validation against cross-validation;
(iv) We analyze the effect of different instance parameters on each method's performance.

We implemented all methods in C++, using CPLEX 20.1 to solve the respective MILP formulations. All experiments were performed on machines with Intel Core i7-6700 CPU at 3.40 GHz, with 16 GB of RAM, and Ubuntu 16.04.6 LTS operating system, under a time limit of 900 s. We provide the source code and data at https://github.com/tumBAIS/Feature-Selection-Newsvendor.

4.1. Instances

We adapt the experimental setup from Zhu et al. (2012) and consider a linear demand model with additive noise:

$$d_{\text{linear}}(\mathbf{x}) = 5 + \boldsymbol{\beta}^\top \mathbf{x} + \epsilon, \tag{50}$$

where $\boldsymbol{\beta} = (0, 2, -2, -1, 1, 0, \ldots, 0)^\top / \sqrt{10}$ is an $(m+1)$-dimensional vector representing the ground-truth coefficients. The feature variables $\mathbf{x} \in$

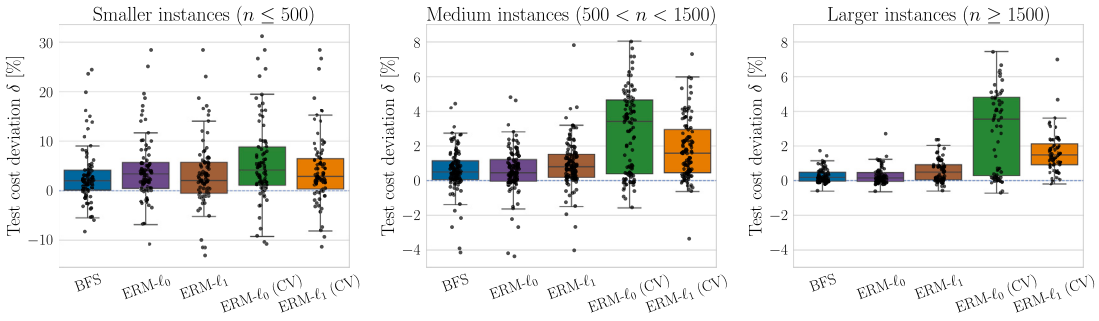


Fig. 2. Impact of instance size on the test cost performance of different methods, relative to BFS-CV.

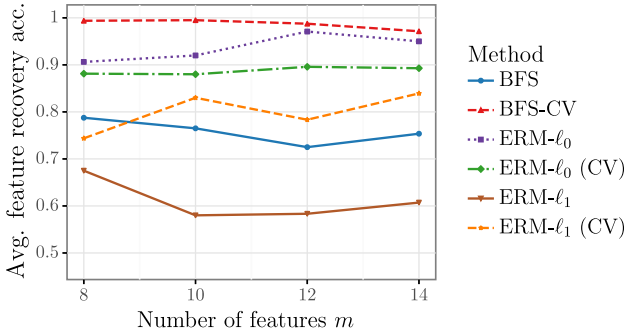


Fig. 3. Impact of number of features m on the accuracy of feature recovery.

$\mathbb{R}^{m+1}$ are drawn from a multivariate Gaussian distribution with mean zero and covariance matrix with entries $\sigma_{ij} = 0.5^{|i-j|}$, for $(i,j) \in J^2$. The noise term follows a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. We set negative demand values to zero.

We generate instances varying the number of samples n from 40 to 2000 and the number of features m from 8 to 14. We deliberately limit the instance sizes in our experiments so that we can find globally optimal or near-optimal solutions within reasonable computation time. For each configuration, we generate 20 instances to account for variability in the distributions. Additionally, we generate a separate test set with 1000 observations associated with each instance, following the same distributions.

Furthermore, we analyze the impact of demand misspecification, i.e., when the demand is not linearly related to the features. We investigate the following nonlinear demand model (cf. Zhu et al. 2012):

$$d_{\text{nonlinear}}(\mathbf{x}) = 10 + \sin\big(2(\boldsymbol{\beta}^\top \mathbf{x})\big) + 2\exp\big(-16(\boldsymbol{\beta}^\top \mathbf{x})^2\big) + \varphi(\boldsymbol{\beta}^\top \mathbf{x})\epsilon, \tag{51}$$

where $\varphi(\boldsymbol{\beta}^\top \mathbf{x}) = 1$ for a homoscedastic case and $\varphi(\boldsymbol{\beta}^\top \mathbf{x}) = \exp(\boldsymbol{\beta}^\top \mathbf{x})$ for a heteroscedastic case.

We solve the proposed BFS and BFS-CV models and compare the results against regularization-based methods from the literature. Since our motivation for feature selection is to provide more explainable decisions, we focus on methods based on linear decision functions, which are intrinsically more explainable. We consider the ERM model of Ban and Rudin (2019) with ℓ_0 and ℓ_1-norm regularization and use grid search with 50 break-points to calibrate the regularization parameter. We run each considered method once using a hold-out validation set and once with Shuffle & Split cross-validation (CV). For hold-out validation, we use half of the samples in each instance as a training set and the other half as a validation set, following the setting in Zhu et al. (2012). For cross-validation, we perform $K = 50$ re-sampling iterations, where we sample a subset of size $|S_k| = \min\{200, n\}$, for each $k \in [K]$.

4.2. Performance metrics

We assess the ability of each method to recover the ground-truth feature vector, adopting the accuracy measure:

$$\alpha = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}(\hat{z}_j = z_j^*), \tag{52}$$

where $\mathbb{1}(\cdot)$ is the indicator function, $\hat{\mathbf{z}} = [\hat{z}_1, \ldots, \hat{z}_m]$ is the estimated binary feature vector, and $\mathbf{z}^* = [z_1^*, \ldots, z_m^*]$ is the ground-truth vector, defined as $z_j^* = 0$ if $\beta_j^* = 0$, i.e., feature j is non-informative, otherwise $z_j^* = 1$. Our definition of accuracy is analogous to the one commonly adopted for binary classification, where $z_j^* \in \{0, 1\}$ represents the class assigned to feature j.

Additionally, we analyze the cost performance of applying the learned decision functions to out-of-sample data. We therefore evaluate the out-of-sample cost on a separate test data set with 1000 observations. We report the test cost values of each method $\mathcal{M}$ in terms of its percentage deviation relative to the test cost achieved by BFS-CV on the same configuration:

$$\delta(\mathcal{M}) = 100 \times \frac{C^{test}_{\mathcal{M}} - C^{test}_{\text{BFS-CV}}}{C^{test}_{\text{BFS-CV}}}, \tag{53}$$

where $C^{test}_{\mathcal{M}}$ is the average newsvendor cost of method $\mathcal{M}$ calculated on the test set. Deviation values greater than zero indicate that BFS-CV improves upon method $\mathcal{M}$ regarding test cost performance, while values below zero indicate that method $\mathcal{M}$ achieves lower test cost than BFS-CV.

5. Results

First, we present results concerning instances generated by the linear demand model, and then discuss results on instances with nonlinear demand.

5.1. Linear demand model

We analyze how instance size, number of features, noise level, shortage cost, and holding cost affect the performance of each method. Unless otherwise stated, we use a setting with $n = 1000$ samples, $m = 10$ features, a shortage cost of $b = 2$, and a holding cost of $h = 1$ as reference configuration. For the noise term, we consider $\sigma_\epsilon = 1$ as reference configuration for the Gaussian distribution. We provide results regarding computation times in Appendix A.

Instance size. Fig. 1 reports the feature recovery accuracy of the different methods for a varying number of samples $n \in [40, 2000]$, averaged over 20 randomly generated instances. In general, BFS-CV achieves the highest accuracy among all methods and faster convergence for increasing n, with accuracy values above 0.9 already for instances with 200 samples. In contrast, existing methods often yield accuracy values below 0.9 and fail to recover the ground-truth features accurately, even for instances with a larger number of samples. We confirm these results

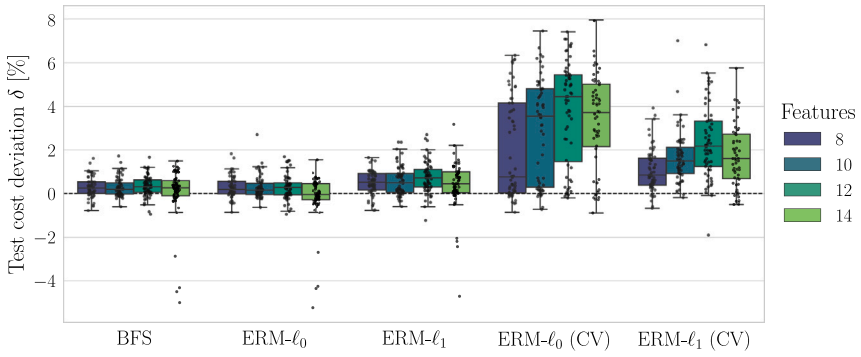


Fig. 4. Impact of number of features m on the test cost performance of different methods, relative to BFS-CV.

at 5% significance level by pairwise Wilcoxon signed-rank tests, and refer to Appendix B for details on the respective p-values.

Besides feature recovery, we evaluate the out-of-sample cost performance of each method. Fig. 2 shows the distribution of percentage deviations, where a positive deviation indicates that BFS-CV is superior to the respective other method. We split the results in three different plots based on the sample size n, classifying each instance as small, medium, or large. BFS-CV outperforms the other methods in most cases, as the lower quartiles are always above or close to zero. For smaller instances, test cost deviations can be as high as 30% in the best case. As we increase n, all methods present improving test cost performance and the variance in the test cost distribution decreases. Yet, BFS-CV is still superior to the other methods in the wide majority of cases. For large instances, all methods present mostly positive test cost deviations, with values ranging from −1% to 8%.

Number of features. Fig. 3 shows average feature recovery accuracy values, where we now fix the number of samples to $n = 1000$ and vary the number of features $m \in \{8, 10, 12, 14\}$. For all methods, the number of features does not strongly affect the accuracy performance. Notably, BFS-CV achieves average accuracy values consistently above 0.95, being superior to the other methods, as confirmed by pairwise Wilcoxon tests.

Regarding test cost performance, Fig. 4 shows the distribution of test cost deviations. We focus on large instances ($n \geq 1500$) in this analysis, which have lower variance, so that we can isolate the impact of m on the test cost. For BFS, ERM-ℓ_0, and ERM-ℓ_1, the number of features has no strong influence on the test cost deviations. In contrast, the performance of ERM-ℓ_0 (CV) and ERM-ℓ_1 (CV) shows increasing deviation values for increasing m. In the majority of cases, BFS-CV outperforms the other considered methods.

Noise level. We vary the coefficient of variation $c_v = \sigma_\epsilon/\mu \in [0.2, 1]$ and report the average feature recovery accuracy for each method in Fig. 5. As we increase the level of noise, it becomes harder to recover the informative features and the accuracy of all methods deteriorates. For $c_v \in [0.2, 0.6]$, BFS-CV achieves the highest accuracy among the considered methods. Outside this range, BFS-CV is outperformed by ERM-ℓ_0 (CV) for $c_v = 0.1$ and by ERM-ℓ_0 for $c_v \geq 0.8$, respectively. Still, BFS-CV generally attains comparatively high accuracy, being superior to most other methods.

Fig. 6 illustrates the impact of different noise levels on test cost performance, considering large instances ($n \geq 1500$). Methods BFS, ERM-ℓ_0, and ERM-ℓ_1 mostly outperform BFS-CV for $c_v \geq 0.4$. In such cases, test cost deviations range from −4% to 4%, indicating that BFS-CV achieves comparable results even when its performance is inferior to other methods. Methods ERM-ℓ_0 (CV) and ERM-ℓ_1 (CV) perform comparatively worse, with mostly positive deviations and larger variance in the distribution.

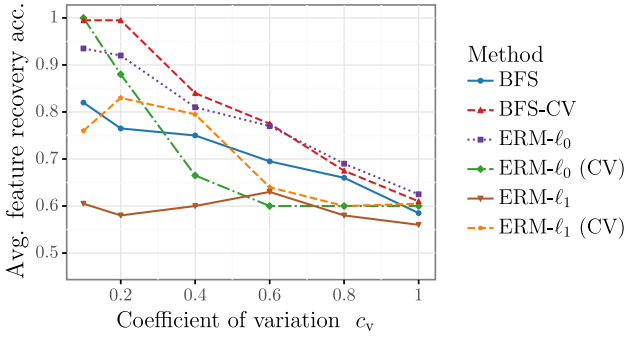


Fig. 5. Impact of different noise levels on the accuracy of feature recovery.

Shortage and holding costs. Fig. 7 displays results on the accuracy performance as a function of the newsvendor ratio $b/(b+h)$, by varying $(b, h) \in [1, 10]^2$. In general, BFS-CV has accuracy values consistently above 0.9 and outperforms the other methods.

Fig. 8 shows results on test cost performance of each method as a function of b, for large instances ($n \geq 1500$), where we fixed $h = 1$, corresponding to newsvendor ratios $b/(b+h) \geq 0.5$. ERM-ℓ_0 (CV) and ERM-ℓ_1 (CV) generally perform worse than BFS-CV, since the deviation values are often positive. For all considered methods, we observe that the test cost deviations range from −6% to 10% and the variance increases with increasing b. We observed similar results for cases with $b/(b+h) < 0.5$.

5.2. Nonlinear demand model

Demand may not be linearly related to the features. Therefore, we investigate how each method performs under nonlinear demand models, considering homoscedastic and heteroscedastic settings. In the following, we only present results regarding accuracy performance. Results on test cost performance did not provide new insights, since we observed similar patterns as in the case of linear instances (see Appendix C). Unless otherwise stated, we use the same reference configuration as in the previous section.

Instance size. Due to the nonlinear structure of the demand models, Fig. 9 shows considerably lower accuracy values compared to instances with linear demand. For heteroscedastic instances, BFS-CV outperforms existing methods, with accuracy values above 0.9 already for $n = 500$ samples. For the homoscedastic case, all methods present inferior

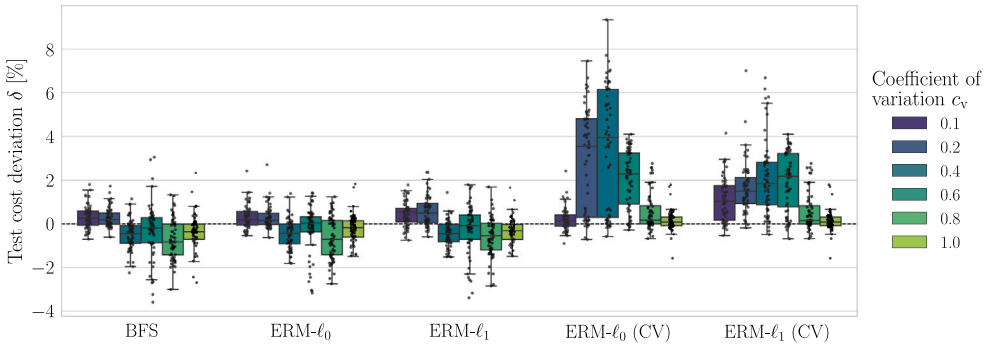


Fig. 6. Impact of noise level on the test cost performance of different methods, relative to BFS-CV.

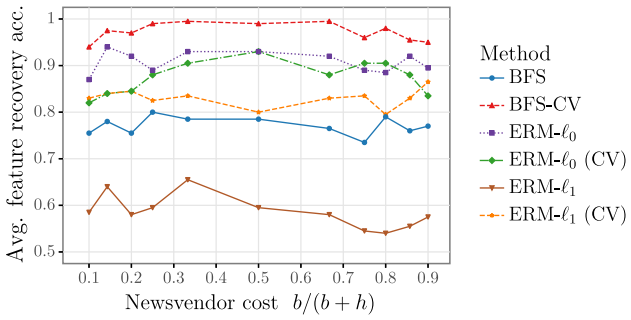


Fig. 7. Impact of shortage cost b and holding cost h on the accuracy of feature recovery.

accuracy compared to the heteroscedastic case. In this setting, BFS-CV and ERM-ℓ_0 present comparable performance, superior to the other considered methods.

Number of features. Fig. 10 shows the average feature recovery accuracy for varying $m \in \{8, 10, 12, 14\}$, for both homoscedastic and heteroscedastic demand. Similarly as for instances with linear demand, the number of features does not strongly influence the accuracy performance for nonlinear instances.

Noise level. For the heteroscedastic case, BFS-CV outperforms existing methods (Fig. 11). In this setting, varying the standard deviation σ_ϵ does not strongly affect the accuracy performance. For homoscedastic instances, all methods present inferior accuracy compared to heteroscedastic instances. Here, ERM-ℓ_0 is superior to other methods for most values of σ_ϵ, but BFS-CV often presents comparable accuracy.

Shortage and holding costs. In Fig. 12, for heteroscedastic instances, BFS-CV has superior accuracy than other methods when $b/(b+h) \geq 0.3$. We observe that all methods perform poorly when the newsvendor ratio is below 0.3. In particular, the considered methods often do not recover any features when $b/(b+h) \in [0.1, 0.2]$, leading to accuracy values between 0.5 and 0.65. For the homoscedastic case, the accuracy of all methods decreases with increasing newsvendor ratios, with values often below 0.8. In this setting, BFS-CV is superior to most methods, but is outperformed by ERM-ℓ_0 when $b/(b+h) \geq 0.8$ and by ERM-ℓ_1 (CV) when $b/(b+h) \in [0.2, 0.35]$.

One comment on these results is in order. Applying a linear decision function to nonlinear data may naturally lead to poor results, due to the inconsistent dependency structure with respect to the features. In some cases, we observed that all methods fail to achieve reasonable accuracy in feature recovery. However, for the vast majority of settings that we considered, BFS-CV presented good performance, outperforming the other methods. One possibility for dealing with such nonlinear instances would be to include additional features by considering nonlinear transformations of the original features, which we leave for future research.

6. Conclusion

We have presented a novel formulation based on bilevel optimization for incorporating feature selection in the feature-based newsvendor problem. The proposed BFS models provide an intuitive approach specifically designed for the task of feature selection, in which the upper-level problem directly selects the subset of relevant features. Our experimental results on synthetic data show that the proposed methods can accurately recover ground-truth informative features, leading to more explainable inventory decisions in comparison to previous methods.

There are many possibilities for follow-up works. First, research on tailored solution methods for BFS and BFS-CV, e.g., based on decomposition strategies for mixed integer programming, may allow to improve the scalability of the proposed methods when dealing with a large number of features. Second, tailoring heuristic methods to efficiently solve very large instances remains an interesting avenue for future research. In this context, our proposed solution methods provide useful benchmarks for evaluating the performance of heuristics with respect to solution quality. Third, other classes of data-driven optimization problems may benefit from an extension of the proposed methodology. As an example, the newsvendor problem also captures the fundamental trade-offs emerging in decisions related to capacity planning. Accordingly, an extension of the proposed BFS models may be applied to select features in stochastic, data-driven variations of such problems. Finally, from a broader perspective, incorporating concepts from machine learning into data-driven optimization problems may lead to crucial advances for both fields. As exemplified in this work, feature selection is one among possibly many machine learning tasks that can be integrated into the decision-making process, especially given the growing need for more explainable decisions.

Acknowledgments

The authors are grateful for the support received through the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the research group Advanced Optimization in a Networked Economy (AdONE, GRK2201/277991500).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.01.025.

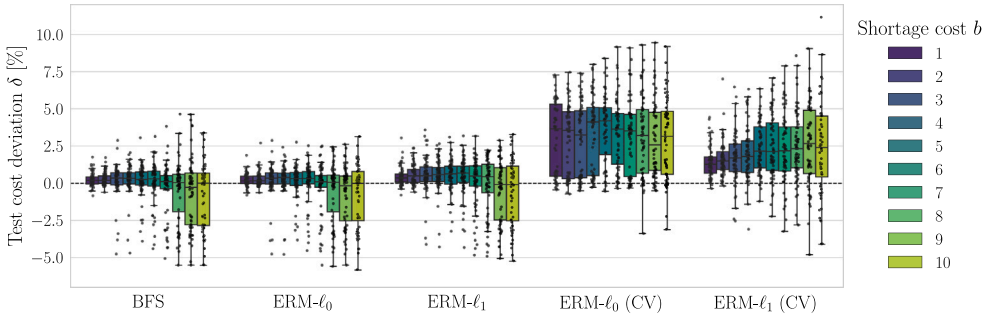


Fig. 8. Impact of shortage cost b on the test cost performance of different methods, relative to BFS-CV.

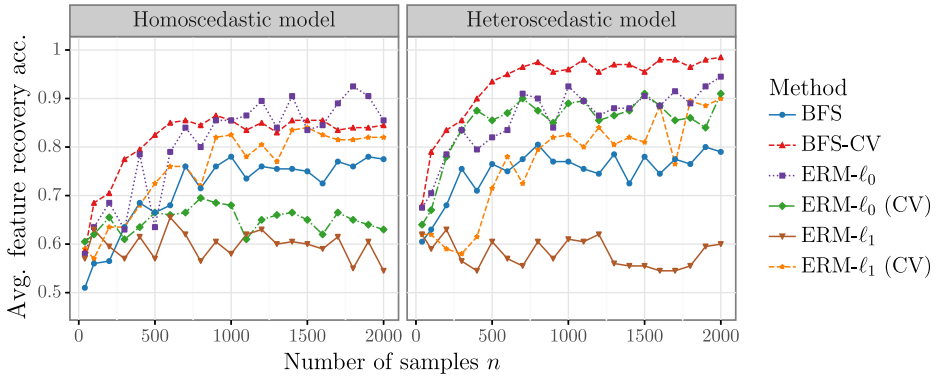


Fig. 9. Impact of instance size on the accuracy of feature recovery.

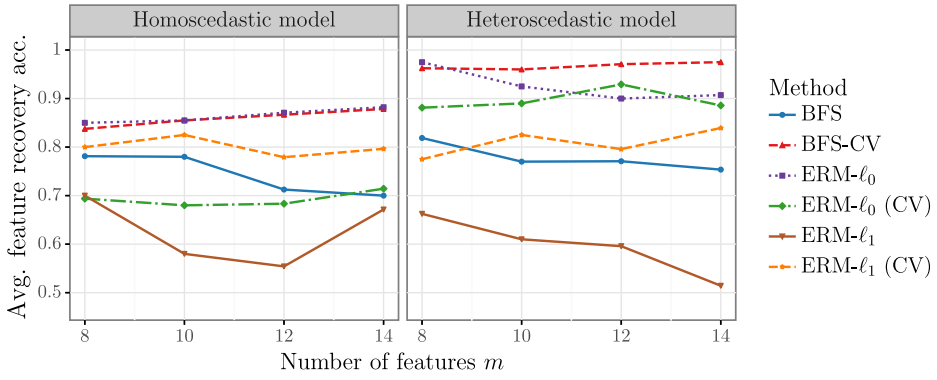


Fig. 10. Impact of number of features m on the accuracy of feature recovery.

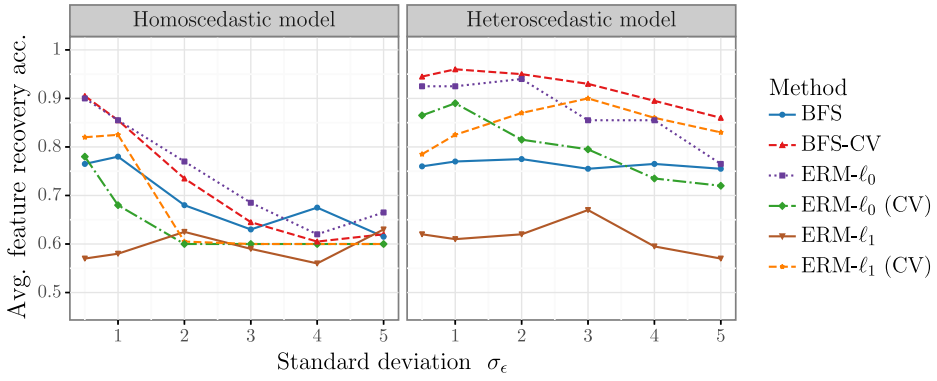


Fig. 11. Impact of different noise levels on the accuracy of feature recovery.

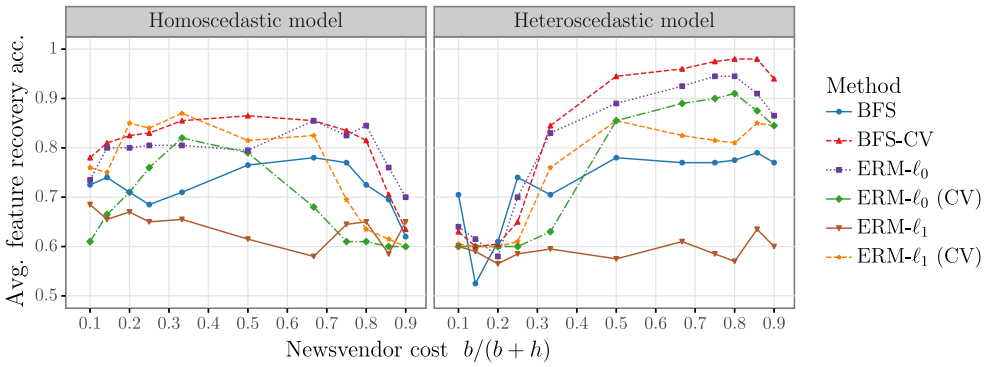


Fig. 12. Impact of shortage cost b and holding cost h on the accuracy of feature recovery.

References

Agor, J., & Özaltın, O. Y. (2019). Feature selection for classification models via bilevel optimization. *Computers & Operations Research*, *106*, 156–168.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Ban, G.-Y. (2020). Confidence intervals for data-driven inventory policies with demand censoring. *Operations Research*, *68*(2), 309–326.

Ban, G.-Y., & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, *67*(1), 90–108.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., & Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, *59*(2), 341–357.

Benati, S., & García, S. (2014). A mixed integer linear model for clustering with variable selection. *Computers & Operations Research*, *43*, 280–285.

Bennett, K. P., Hu, J., Ji, X., Kunapuli, G., & Pang, J.-S. (2006). Model selection via bilevel optimization. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 1922–1929).

Bennett, K. P., Kunapuli, G., Hu, J., & Pang, J.-S. (2008). Bilevel optimization and machine learning. In J. M. Zurada, G. G. Yen, & J. Wang (Eds.), *Computational intelligence: research frontiers: IEEE world congress on computational intelligence* (pp. 25–47). Springer, Berlin, Heidelberg.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(10), 281–305.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In S. Dasgupta, & D. McAllester (Eds.), *Proceedings of machine learning research: vol. 28*, *Proceedings of the 30th international conference on machine learning* (pp. 115–123). Atlanta, Georgia, USA: PMLR.

Bertsimas, D., Gupta, V., & Kallus, N. (2018). Robust sample average approximation. *Mathematical Programming*, *171*(1), 217–282.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, *44*(2), 813–852.

Bertsimas, D., & Thiele, A. (2005). *A data-driven approach to newsvendor problems*: *Technical Report*, Cambridge, MA: Citeseer, Massachusetts Institute of Technology.

Bertsimas, D., & Thiele, A. (2006). A robust optimization approach to inventory theory. *Operations Research*, *54*(1), 150–168.

Besbes, O., & Mouchtaki, O. (2023). How big should your data really be? Data-driven newsvendor: Learning one sample at a time. *Management Science*, *69*(10), 5848–5865.

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

Beutel, A.-L., & Minner, S. (2012). Safety stock planning under causal demand forecasting. *International Journal of Production Economics*, *140*(2), 637–645.

Cao, D., & Chen, M. (2006). Capacitated plant selection in a decentralized manufacturing environment: A bilevel optimization approach. *European Journal of Operational Research*, *169*(1), 97–110.

Chen, X., Sim, M., Simchi-Levi, D., & Sun, P. (2007). Risk aversion in inventory management. *Operations Research*, *55*(5), 828–842.

Cheung, W. C., & Simchi-Levi, D. (2019). Sampling-based approximation schemes for capacitated stochastic inventory control models. *Mathematics of Operations Research*, *44*(2), 668–692.

Choi, T.-M. (2012). *Handbook of newsvendor problems: models, extensions and applications*: *vol. 176*, Springer, New York.

Elmachtoub, A. N., & Grigas, P. (2021). Smart "predict, then optimize". *Management Science*, *68*(1), 9–26.

Fletcher, R., & Leyffer, S. (2002). Nonlinear programming without a penalty function. *Mathematical Programming*, *91*(2), 239–269.

Fontaine, P., & Minner, S. (2014). Benders decomposition for discrete–continuous linear bilevel problems with application to traffic network design. *Transportation Research, Part B (Methodological)*, *70*, 163–172.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., & Pontil, M. (2018). Bilevel programming for hyperparameter optimization and meta-learning. In J. Dy, & A. Krause (Eds.), *Proceedings of machine learning research*: *vol. 80*, *Proceedings of the 35th international conference on machine learning* (pp. 1568–1577). Stockholm, Sweden: PMLR.

Gallego, G., & Moon, I. (1993). The distribution free newsboy problem: Review and extensions. *Journal of the Operational Research Society*, *44*(8), 825–834.

Gallego, G., Ryan, J. K., & Simchi-Levi, D. (2001). Minimax analysis for finite-horizon inventory models. *IIE Transactions*, *33*(10), 861–874.

Gómez, A., & Prokopyev, O. A. (2021). A mixed-integer fractional optimization approach to best subset selection. *INFORMS Journal on Computing*, *33*(2), 551–565.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hanasusanto, G. A., Kuhn, D., Wallace, S. W., & Zymler, S. (2015). Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, *152*(1), 1–32.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*: *vol. 2*, Springer, New York.

Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, *278*(3), 904–915.

Khouja, M. (1999). The single-period (news-vendor) problem: Literature review and suggestions for future research. *Omega*, *27*(5), 537–553.

Kimura, K., & Waki, H. (2018). Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. *Optimization Methods & Software*, *33*(3), 633–649.

Kleywegt, A. J., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502.

Kogan, K., & Lou, S. (2003). Multi-stage newsboy problem: A dynamic model. *European Journal of Operational Research*, *149*(2), 448–458.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.

Lau, H.-S., & Lau, A. H.-L. (1996). The newsstand problem: A capacitated multiple-product single-period inventory problem. *European Journal of Operational Research*, *94*(1), 29–42.

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

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

Mackay, M., Vicol, P., Lorraine, J., Duvenaud, D., & Grosse, R. (2019). Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *International conference on learning representations*. New Orleans, Louisiana, USA.

Maldonado, S., Pérez, J., Weber, R., & Labbé, M. (2014). Feature selection for support vector machines via mixed integer linear programming. *Information Sciences*, *279*, 163–175.

Mandi, J., Demirović, E., Stuckey, P. J., & Guns, T. (2020). Smart predict-and-optimize for hard combinatorial optimization problems. *34*, In *Proceedings of the AAAI Conference on Artificial Intelligence* (2), (pp. 1603–1610).

Mandl, C., & Minner, S. (2023). Data-driven optimization for commodity procurement under price uncertainty. *Manufacturing & Service Operations Management*, *25*(2), 371–390.

Miyashiro, R., & Takano, Y. (2015). Mixed integer second-order cone programming formulations for variable selection in linear regression. *European Journal of Operational Research*, *247*(3), 721–731.

Molina, L., Belanche, L., & Nebot, A. (2002). Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the IEEE international conference on data mining* (pp. 306–313). Maebashi City, Japan.

Moon, I., & Gallego, G. (1994). Distribution free procedures for some inventory models. *Journal of the Operational Research Society*, *45*(6), 651–658.

Oroojlooyjadid, A., Snyder, L. V., & Takáč, M. (2020). Applying deep learning to the newsvendor problem. *IISE Transactions*, *52*(4), 444–463.

Park, Y. W., & Klabjan, D. (2020). Subset selection for multiple linear regression via optimization. *Journal of Global Optimization*, *77*(3), 543–574.

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

Petruzzi, N. C., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research*, *47*(2), 183–194.

Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., & Seref, M. M. (2011). The newsvendor problem: Review and directions for future research. *European Journal of Operational Research*, *213*(2), 361–374.

Sachs, A.-L., & Minner, S. (2014). The data-driven newsvendor with censored demand observations. *International Journal of Production Economics*, *149*, 28–36.

Scarf, H. (1958). A min-max solution of an inventory problem. In K. Arrow, S. Karlin, & H. Scarf (Eds.), *Studies in the mathematical theory of inventory and production* (pp. 201–209). Stanford: Stanford University Press.

See, C.-T., & Sim, M. (2010). Robust approximation to multiperiod inventory management. *Operations Research*, *58*(3), 583–594.

Shapiro, A. (2003). Monte Carlo sampling methods. In A. Ruszczynski, & A. Shapiro (Eds.), *Handbooks in operations research and management science*: *vol. 10*, *Stochastic programming* (pp. 353–425). Elsevier, Amsterdam.

Takano, Y., & Miyashiro, R. (2020). Best subset selection via cross-validation criterion. *TOP*, *28*(2), 475–488.

Tian, Y.-X., & Zhang, C. (2023). An end-to-end deep learning model for solving data-driven newsvendor problem with accessibility to textual review data. *International Journal of Production Economics*, *265*, Article 109016.

Wang, Z., Glynn, P. W., & Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science*, *13*(2), 241–261.

Wang, C. X., & Webster, S. (2009). The loss-averse newsvendor problem. *Omega*, *37*(1), 93–105.

Yue, J., Chen, B., & Wang, M.-C. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, *54*(6), 1128–1136.

Zhang, Y., & Gao, J. (2017). Assessing the performance of deep learning algorithms for newsvendor problem. In D. Liu, S. Xie, Y. Li, D. Zhao, & E.-S. M. El-Alfy (Eds.), *Neural information processing* (pp. 912–921). Springer, Cham.

Zhu, L., Huang, M., & Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, *22*(4), 1379–1401.