Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(9): 6011–6042.
DOI: 10.3934/era.2026266
Received: 22 April 2026
Revised: 10 June 2026
Accepted: 24 June 2026
Published: 13 July 2026

Research article

On the coefficients of prescriptiveness in contextual optimization

Bo Jiang, Xuecheng Tian* and Shuaian Wang

Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong 999077, China

* **Correspondence:** Email: xuecheng-simon.tian@connect.polyu.hk.

Abstract: This paper focused on the coefficients of prescriptiveness in the field of contextual optimization, which measure the efficacy of a prescriptive analytics method in leveraging contextual information for prescribing data-driven decisions. We identified two key deficiencies of the original coefficient of prescriptiveness proposed in the literature. First, this coefficient is problem-dependent, i.e., it lacks a universal scale: its asymptotic upper bound (less than or equal to 1) varies across problems, making cross-problem comparisons infeasible. Second, this coefficient is data-dependent, i.e., its calculation is contingent on specific training and test datasets, limiting its generalizability. To address the two deficiencies, we proposed two novel coefficients of prescriptiveness in this paper: (i) the adjusted coefficient of prescriptiveness, which replaces the deterministic perfect-foresight benchmark in the original coefficient with the full-information optimum; and (ii) the expected coefficient of prescriptiveness, derived by taking the expectation over the sampling distribution of the training dataset. Meanwhile, we developed their empirical estimators for practical applications. Moreover, rigorous theoretical analysis established two fundamental properties of these coefficients of prescriptiveness. Through extensive numerical experiments based on the newsvendor problem, we validated the effectiveness of our proposed coefficients of prescriptiveness in addressing the identified deficiencies. By refining the evaluation framework for prescriptive analytics, this work advances the generalizability and applicability of the coefficients of prescriptiveness.

Keywords: prescriptive analytics; contextual optimization; coefficient of prescriptiveness; decision analysis

1. Introduction

Traditionally, decision making under uncertainty in operations research and management science (OR/MS) has primarily focused on two primitives: (i) $Y \in \mathscr{Y} \subset \mathbb{R}^{d_Y}$ is a vector of uncertain parameters (where d_Y is the dimension of Y) with the underlying true distribution μ_Y and the historical observed

data $\{y_1, ..., y_N\}$ ($N \in \mathbb{Z}_{++}$); (ii) z is a decision constrained in $\mathscr{Z} \subset \mathbb{R}^{d_z}$ (where d_z is the dimension of z) with the objective of minimizing the expected cost $\mathbb{E}[c(z; Y)]$, where $c(\cdot; \cdot)$ is the cost function. We assume that $\mathscr{Z}$ is a non-empty compact set, that $c(z; Y)$ is well-defined and continuous in z for every $z \in \mathscr{Z}$ and μ_Y-almost-everywhere $Y \in \mathscr{Y}$, and that $\mathbb{E}[c(z; Y)]$ is well-defined and finite for every $z \in \mathscr{Z}$. Based on these, the classic stochastic optimization problem is defined as follows [1]:

$$v^{\text{stoch}} = \min_{z \in \mathscr{Z}} \mathbb{E}_Y[c(z; Y)], \tag{1.1}$$

$$z^{\text{stoch}} \in \arg\min_{z \in \mathscr{Z}} \mathbb{E}_Y[c(z; Y)]. \tag{1.2}$$

To solve optimization problem (1.2), the sample average approximation (SAA) is a common data-driven method, using data $\{y_1, ..., y_N\}$ as the empirical distribution of Y, to obtain an approximate solution [2, 3].

With the availability of rich data and the development of machine learning (ML) methods, OR/MS modeling increasingly makes use of auxiliary data $\{x_1, ..., x_N\}$ on covariates $X \in \mathscr{X} \subset \mathbb{R}^{d_X}$ (where d_X is the dimension of X) associated with the random variable Y, where x_i is concurrently observed with y_i ($i \in \{1, ..., N\}$). Let μ_X be the underlying true marginal distribution of X. Therefore, the contextual stochastic optimization (CSO) problem is introduced, using the corresponding dataset $s_N = \{(x_1, y_1), ..., (x_N, y_N)\}$, to prescribe data-driven actions/decisions for new observations of the covariates X. Specifically, for a given new observation $X = x$, the CSO problem is defined as follows [4, 5]:

$$v^*(x) = \min_{z \in \mathscr{Z}} \mathbb{E}_Y[c(z; Y)|X = x], \tag{1.3}$$

$$z^*(x) \in \arg\min_{z \in \mathscr{Z}} \mathbb{E}_Y[c(z; Y)|X = x]. \tag{1.4}$$

The solution $z^*(x)$ represents the full-information optimal decision with full knowledge of the conditional distribution $\mu_{Y|X=x}$ of Y given $X = x$. However, in general, the underlying true joint distribution $\mu_{X,Y}$ of (X, Y) is unknown in practice and only the historical dataset $s_N = \{(x_1, y_1), ..., (x_N, y_N)\}$ is available. In this context, we generally approximate $\mathbb{E}[c(z; Y)|X = x]$ based on s_N and obtain a data-driven predictive prescription $\hat{z}_N(x)$ (a feasible solution to optimization problem (1.4), i.e., $\hat{z}_N(x) \in \mathscr{Z}$), with the aim of its performance (out-of-sample cost) $\mathbb{E}[c(\hat{z}_N(x); Y)|X = x]$ being close to the full-information optimum $v^*(x)$. Existing data-driven methods that we might apply include the point-prediction method [6] and weighted SAA (wSAA) [4] (these are introduced in Section 2.1), among others.

To measure the efficacy of a predictive prescription $\hat{z}_N(\cdot)$ (note that $\hat{z}_N(x)$ is simplified to $\hat{z}_N(\cdot)$ for convenience), Bertsimas and Kallus [4] define a relative and unitless metric P, which is termed the coefficient of prescriptiveness. Specifically, given a fixed training dataset $s_N = \{(x_1, y_1), ..., (x_N, y_N)\}$ and a fixed test dataset $\bar{s}_M = \{(\bar{x}_1, \bar{y}_1), ..., (\bar{x}_M, \bar{y}_M)\}$ ($M \in \mathbb{Z}_{++}$), let $\hat{R}_M(\hat{z}_N)$ be the estimated out-of-sample cost of a predictive prescription $\hat{z}_N(\cdot)$, $\hat{R}_M(\hat{z}_N^{\text{SAA}})$ be the estimated out-of-sample cost of the data-poor SAA solution $\hat{z}_N^{\text{SAA}}$ (solely based on Y data), and $\hat{R}_M^*$ be the out-of-sample cost of the deterministic perfect-foresight decision (i.e., the decision made with foreknowledge of $\bar{y}_j$, $j \in \{1, ..., M\}$, without any uncertainties) (the three costs are further discussed in Section 2.2). Then, the coefficient of prescriptiveness is defined as

$$P = \frac{\hat{R}_M(\hat{z}_N^{\text{SAA}}) - \hat{R}_M(\hat{z}_N)}{\hat{R}_M(\hat{z}_N^{\text{SAA}}) - \hat{R}_M^*}. \tag{1.5}$$

As discussed by Bertsimas and Kallus [4], the coefficient P is a unitless quantity bounded above by the value of 1. That is, if Y is measurable with respect to X (i.e., Y is a function of X), then, under mild conditions, we have $\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N) = \lim_{M\to\infty} \hat{R}^*_M$, indicating that $\hat{R}_M(\hat{z}_N)$ equals $\hat{R}^*_M$, if we have an infinite number of both training and test data. In this case, we would have $\lim_{N,M\to\infty} P(N; M) = 1$ almost surely. Nevertheless, for most CSO problems in practice, Y may not be fully measurable with respect to X (for example, the data-generating process has noise or the underlying relationship cannot be learned), which means that P always has an asymptotic upper bound that is less than or equal to 1, even though we may have a sufficiently large number of data, i.e., $\lim_{N,M\to\infty} P(N; M) \leq 1$, almost surely. For example, in a two-stage shipment planning problem with random demand as discussed in the illustrative example provided by Bertsimas and Kallus [4], the P values of some good predictive prescriptions would converge to 0.46 approximately, instead of 1, as the training sample size increases. Note that P reaches its asymptotic upper bound when the predictive prescription $\hat{z}_N(\cdot)$ approximates the full-information optimal decision, with $\hat{R}_M(\hat{z}_N)$ approaching the full-information optimum. Generally, the asymptotic upper bound of P varies across different CSO problems, which may be related to the problem itself (e.g., the objective function and the parameter settings of the problem) and/or the underlying true joint distribution $\mu_{X,Y}$ of (X, Y). Furthermore, the asymptotic upper bound of P for a specific problem remains unknown to us, because (i) we may have limited real-world data on hand for training; and (ii) we may not have enough resources to compare numerous data-driven methods and observe whether their P values converge or approximately converge to a certain value. We denote this inherent, varying, and unknown asymptotic upper bound (less than or equal to 1) of P as the problem-dependent deficiency of the coefficient of prescriptiveness, i.e., P lacks a universal scale. This deficiency implies that we can only analyze and compare the P values for different prescriptive analytics methods in one particular CSO problem, instead of two or more problems, which we further discuss in Section 2.3.1.

Furthermore, the coefficient of prescriptiveness P is defined based on a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$. However, a training dataset of a given size N is essentially a random sample of N independent and identically distributed (i.i.d.) draws of (X, Y) from the underlying true joint distribution $\mu_{X,Y}$, and s_N is just a specific sample realization. The same applies to $\bar{s}_M$. Therefore, if the decision-maker uses another realized training or test dataset, the P value may change, and some decisions or insights derived from the original dataset may also change. We denote this dependency on specific datasets as the data-dependent deficiency of the coefficient of prescriptiveness, which we further discuss in Section 2.3.2.

1.1. Contributions

To address the problem-dependent deficiency, we propose a novel metric termed the adjusted coefficient of prescriptiveness, which serves as the primary contribution of this paper, supported by the following key points. First, We identify that there always exists an inherent deficiency of the original coefficient of prescriptiveness; that is, for different CSO problems, the asymptotic upper bounds of P, which are always less than or equal to 1, are inherent, varying, and unknown, which means that we can only analyze and compare P values of different methods in one particular CSO problem.

Second, to address the identified deficiency, we propose a novel metric termed the adjusted coefficient of prescriptiveness (Definition 5), whose asymptotic upper bound is theoretically always

equal to 1 for all possible CSO problems under appropriate conditions. Unlike P, the proposed adjusted coefficient uses the full-information optimum as the benchmark (see Formula (1.3), and the more formal definition given in Definition 2), instead of the deterministic perfect-foresight counterpart (Formula (2.6)). Moreover, for real-world applications, we propose the estimate of the adjusted coefficient of prescriptiveness (Definition 6) to solve the problem-dependent deficiency practically.

Third, we discuss some in-depth theoretical properties. To analyze the essential difference between the original coefficient of prescriptiveness and the adjusted coefficient, we define the gap between the full-information optimum and the deterministic perfect-foresight counterpart (Definition 8) and demonstrate its non-negativity (Proposition 1). We also discuss the convergence properties of the original and the adjusted coefficients of prescriptiveness, i.e., the relationship between them when the number of data tends to infinity (Proposition 2).

Fourth, we conduct a case study of the newsvendor problem in Section 5. Through extensive numerical experiments, we confirm the deficiency of the original coefficient of prescriptiveness and validate the effectiveness of our proposed adjusted coefficient.

To address the data-dependent deficiency, another contribution of this paper is a proposed novel metric termed the expected coefficient of prescriptiveness (Formula (3.9)), and more importantly, its estimate (Definition 7) for practical applications. That is, if we have a large number of real-world data in practice, we can use the estimate of the expected coefficient of prescriptiveness, which is more general than P, to more fairly evaluate the average performance of a predictive prescription $\hat{z}_N(\cdot)$ across all available training datasets of a given size for a particular CSO problem.

1.2. Related literature

Our work is primarily related to contextual optimization [7], also called prescriptive analytics [4] by some authors. In OR/MS, improvements in data availability and advancements in ML techniques are transforming traditional decision-making problems with uncertainty, which ignore correlated side information [1, 8], into contextual optimization problems, which consider the data of covariates [5, 6, 9]. Contextual optimization aims to prescribe actions/decisions for new observations of covariates based on the historical data of both covariates and the corresponding uncertain parameters. Existing studies propose many methodologies for contextual optimization, including decision rule optimization [10–12], sequential learning and optimization [13–16], and integrated learning and optimization [17–19]. Recent studies have also developed learning-assisted and data-driven optimization algorithms, such as data-driven evolutionary algorithms [20]. The real-world applications of data-driven optimization can be seen in various fields, such as logistics [21, 22], production [23, 24], and energy management [25, 26].

However, few studies explore the evaluation metrics for data-driven prescriptions, i.e., methods for measuring the quality of the prescribed actions/decisions, despite their importance for decision-makers. The literature on traditional predictive analytics describes both absolute metrics, including mean squared error (MSE) [27], root mean squared error (RMSE), and mean absolute error (MAE) [28], and relative and unitless metrics, including the coefficient of determination (R^2) [29] and the adjusted R^2 (R^2_{adj}) [30–32], to measure the performance of a predictor (e.g., linear regression). In prescriptive analytics, the decision loss, as an absolute metric, was first introduced to the OR/MS and ML communities in Bengio [33], and popularized in Elmachtoub and Grigas [34] for the recent

context of big data and huge computational power. Elmachtoub and Grigas [34] proposed a general framework of smart "predict, then optimize" (SPO), and introduced the SPO loss function and SPO+ loss function to measure the decision error induced by a prediction. Shah et al. [35] constructed the locally optimized decision loss to evaluate the efficacy of a predictor for downstream optimization tasks. Notably, Bertsimas and Kallus [4] developed a relative and unitless metric, the coefficient of prescriptiveness, to measure the efficacy of a predictive prescription. Departing from these existing metrics, our paper builds on the original coefficient of prescriptiveness but goes a step further; that is, our novel metric, the adjusted coefficient of prescriptiveness, addresses the problem-dependent deficiency of P, a topic not yet explored in the literature. In addition, recent work has identified a gap between data-driven policies generated by the wSAA framework and oracle policies [36]. This observation further motivates the need for transparent evaluation metrics that are not tied to a particular prescriptive method.

Our work also relates to the literature on the expected out-of-sample/generalization error in supervised learning, such as Bates et al. [37], Wager [38], and Yang [39]. Generally, we discuss the out-of-sample error for an ML model based on a fixed and realized training dataset; however, the training dataset of size N is essentially a random sample of N i.i.d. draws from the underlying true joint distribution of the feature and target. Therefore, the expected out-of-sample error can be obtained by taking the expectation over the random training dataset [37]. A similar metric, termed the expected coefficient of prescriptiveness, is proposed to address the data-dependent deficiency of P in this paper.

1.3. Organization of the paper

The remainder of this paper is organized as follows. Section 2 introduces common data-driven methods and the original coefficient of prescriptiveness, along with the two deficiencies of this coefficient. In Section 3, we propose two novel metrics for prescriptive analytics methods, i.e., the adjusted coefficient of prescriptiveness and the expected coefficient of prescriptiveness. Section 4 presents in-depth theoretical analyses of these proposed metrics and obtains three useful properties. Section 5 conducts comprehensive numerical experiments on the newsvendor problem to validate the effectiveness of our proposed metrics. Section 6 concludes the paper.

2. Preliminaries

We introduce some existing data-driven methods for CSO problems in Section 2.1, describe the definition of the original coefficient of prescriptiveness in Section 2.2, and identify the two deficiencies of P in Section 2.3. Recall that the CSO problem has been presented in Formulas (1.3) and (1.4) in Section 1.

2.1. Data-driven methods

2.1.1. SAA

The SAA [2, 3] is solely concerned with the marginal distribution of Y, thus ignoring the observed data on X. Specifically, the SAA uses the realizations $\{y_1, ..., y_N\}$, i.e., the empirical distribution, to

approximate the marginal distribution of Y, and solves the following data-driven optimization problem:

$$\hat{z}_N^{\mathrm{SAA}} \in \arg\min_{z\in\mathscr{Z}} \frac{1}{N}\sum_{i=1}^{N} c(z; y_i), \tag{2.1}$$

whose objective approximates $\mathbb{E}[c(z; Y)]$ instead of $\mathbb{E}[c(z; Y)|X = x]$.

2.1.2. Point-prediction method

In the point-prediction method [6], a predictive ML model (e.g., random forest) trained on the dataset s_N is often employed to provide a deterministic point-prediction $\hat{m}_N(x)$ for the expected value of Y given the new observation $X = x$, leading to the corresponding optimization problem:

$$\hat{z}_N^{\mathrm{point}}(x) \in \arg\min_{z\in\mathscr{Z}} c(z; \hat{m}_N(x)), \tag{2.2}$$

whose objective approximates $c(z; \mathbb{E}[Y|X = x])$ instead of $\mathbb{E}[c(z; Y)|X = x]$. Notably, if the cost function $c(z; Y)$ is linear in Y, we have $\mathbb{E}[c(z; Y)|X = x] = c(z; \mathbb{E}[Y|X = x])$. However, if the cost function $c(z; Y)$ is nonlinear in Y, approximating $c(z; \mathbb{E}[Y|X = x])$ is not equivalent to approximating $\mathbb{E}[c(z; Y)|X = x]$.

2.1.3. wSAA

The wSAA method [4], based on various predictive ML models, calculates weights $w_{N,i}(x)$ for samples (x_i, y_i) $(i \in \{1, ..., N\})$ in the dataset s_N given the new observation $X = x$ and solves the following data-driven optimization problem:

$$\hat{z}_N^{\mathrm{wSAA}}(x) \in \arg\min_{z\in\mathscr{Z}} \sum_{i=1}^{N} w_{N,i}(x)c(z; y_i), \tag{2.3}$$

whose objective approximates $\mathbb{E}[c(z; Y)|X = x]$. The calculation of the weights in the wSAA method, motivated by some common ML models, is given in Appendix A. The wSAA framework is flexible, but its performance is still affected by the quality of the learned weights, finite-sample variability, and the possible gap between a data-driven weighted empirical policy and an oracle policy that has access to the true conditional distribution. Therefore, in this paper, we use wSAA as an important representative prescriptive method, while our proposed coefficients are method-agnostic and can also be applied to other data-driven prescriptions.

In this paper, we restrict our consideration to the aforementioned three data-driven methods, as they represent the most commonly used and easily implementable approaches in practice; please refer to Sadana et al. [7] for alternative methods.

2.2. The original coefficient of prescriptiveness

According to Bertsimas and Kallus [4], the original coefficient of prescriptiveness involves three quantities given a fixed training dataset $s_N = \{(x_1, y_1), ..., (x_N, y_N)\}$ and a fixed test dataset $\bar{s}_M = \{(\bar{x}_1, \bar{y}_1), ..., (\bar{x}_M, \bar{y}_M)\}$, based on which the definition of P is proposed.

First, the estimated out-of-sample cost of a predictive prescription $\hat{z}_N(\cdot)$ is

$$\hat{R}_M(\hat{z}_N) = \frac{1}{M}\sum_{j=1}^{M} c(\hat{z}_N(\bar{x}_j); \bar{y}_j), \tag{2.4}$$

where we measure the performance of $\hat{z}_N(\cdot)$ (trained on s_N) on the test set $\bar{s}_M$. Here, we discuss the prescriptions considering feature X, with examples defined in Formulas (2.2) and (2.3).

Second, the estimated out-of-sample cost of the data-poor SAA solution $\hat{z}_N^{\mathrm{SAA}}$, which is defined in Formula (2.1) (solely based on Y data and not considering feature X), is

$$\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) = \frac{1}{M}\sum_{j=1}^{M} c(\hat{z}_N^{\mathrm{SAA}}; \bar{y}_j). \tag{2.5}$$

Third, the out-of-sample cost of the deterministic perfect-foresight counterpart problem (where one has foreknowledge of Y without any uncertainties) is

$$\hat{R}_M^* = \frac{1}{M}\sum_{j=1}^{M} \min_{z\in\mathscr{Z}} c(z; \bar{y}_j). \tag{2.6}$$

Here, let $z^{\mathrm{fore}}(\cdot)$ be the deterministic perfect-foresight decision, i.e., $z^{\mathrm{fore}}(\bar{y}_j) \in \arg\min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$. Note that $z^{\mathrm{fore}}(\cdot)$ and $\hat{R}_M^*$ rely solely on a realized test dataset, and thus we refer to $\hat{R}_M^*$ as the out-of-sample cost (without "estimated") of the deterministic perfect-foresight counterpart.

Definition 1 (The coefficient of prescriptiveness)**.** *Based on a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$, the coefficient of prescriptiveness P for a predictive prescription $\hat{z}_N(\cdot)$ is defined as follows:*

$$P = 1 - \frac{\hat{R}_M(\hat{z}_N) - \hat{R}_M^*}{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M^*} = \frac{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(\hat{z}_N)}{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M^*}. \tag{2.7}$$

Here, we assume that no feasible solution $\bar{z}$ exists that can minimize all cost functions $c(z; \bar{y}_j)$ ($j \in \{1, ..., M\}$), i.e., there is no $\bar{z} \in \mathscr{Z}$ such that $c(\bar{z}; \bar{y}_j) = \min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$ holds for every $j \in \{1, ..., M\}$, thereby ensuring that $\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) \neq \hat{R}_M^*$ and P is well-defined. This nondegeneracy assumption is closely related to the case in which Y retains residual uncertainty after conditioning on X. If Y is not measurable with respect to X and different realizations of Y lead to different ex-post minimizers with positive probability, then a single fixed decision cannot attain the deterministic perfect-foresight cost for all test samples. Conversely, in a degenerate problem where all realizations share the same ex-post minimizer, the denominator in Formula (2.7) may vanish and the coefficient is not informative.

This coefficient is a unitless quantity bounded above by the value of 1. P represents the extent of the potential that X has in a predictive prescription $\hat{z}_N(\cdot)$ to reduce costs in a particular problem based on the training dataset s_N and the test dataset $\bar{s}_M$. A low P indicates that X provides little useful information for prescribing a data-driven decision in a particular problem or that $\hat{z}_N(\cdot)$ is inefficient in leveraging the data on X. In contrast, a high P indicates that taking X into consideration significantly reduces costs and that $\hat{z}_N(\cdot)$ is efficient in leveraging the data on X for this purpose.

Here, we are discussing the P value for a predictive prescription $\hat{z}_N(\cdot)$ under the training dataset s_N and the test dataset $\bar{s}_M$. In other words, s_N and $\bar{s}_M$ are given and fixed in a particular problem; then, a particular predictive prescription $\hat{z}_N(\cdot)$ is trained on s_N and P measures the efficacy of a method (e.g., wSAA) based on $\bar{s}_M$, compared with an oracle that can solve the deterministic perfect-foresight counterpart problem. Therefore, when we calculate the P value for a predictive prescription $\hat{z}_N(\cdot)$, or compare the P values for different prescriptive analytics methods in a particular problem, we must first specify the training dataset s_N and the test dataset $\bar{s}_M$ that we use.

2.3. Two deficiencies of the original coefficient

2.3.1. The problem-dependent deficiency

As proposed in Section 1, the original coefficient of prescriptiveness always has an asymptotic upper bound that is less than or equal to 1 for a specific CSO problem, and may vary across different problems. In this section, we theoretically analyze the reasons for the existence of this deficiency, and propose our research question accordingly.

The original coefficient P (Formula (2.7)) given by Bertsimas and Kallus [4] is related to $\hat{R}_M^*$, i.e., the out-of-sample cost of the deterministic perfect-foresight counterpart problem. This means that the coefficient P is the fraction of the way that knowledge of X, leveraged in a predictive prescription $\hat{z}_N(\cdot)$, takes us from making a decision under full uncertainty about the value of Y (i.e., the data-poor SAA solution $\hat{z}_N^{\mathrm{SAA}}$) to making a decision in a completely deterministic setting without any uncertainty of Y (i.e., the deterministic perfect-foresight decision $z^{\mathrm{fore}}(\cdot)$).

For some potential predictive prescriptions $\hat{z}_N(\cdot)$ and under appropriate conditions [4]: (i) If Y is independent of X, we have

$$\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N) = \mathbb{E}_X[\min_{z\in\mathscr{Z}} \mathbb{E}_Y[c(z;Y)|X]] = \min_{z\in\mathscr{Z}} \mathbb{E}_Y[c(z;Y)] = \lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N^{\mathrm{SAA}});$$

hence, as N and M grow, we have $\lim_{N,M\to\infty} P(N;M) = 0$ almost surely; (ii) If Y is measurable with respect to X (that is, Y is a function of X), then we have

$$\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N) = \mathbb{E}_X[\min_{z\in\mathscr{Z}} \mathbb{E}_Y[c(z;Y)|X]] = \mathbb{E}_Y[\min_{z\in\mathscr{Z}} c(z;Y)] = \lim_{M\to\infty} \hat{R}_M^*;$$

hence, in this case, if we have an infinite number of both training and test data, we have $\lim_{N,M\to\infty} P(N;M) = 1$ almost surely.

However, in reality, Y may not be fully measurable with respect to X (because we are unable to capture and consider all underlying covariates related to Y in most problems); otherwise, we would be able to directly calculate the value of Y based on a deterministic function $Y = f(X)$ with a new observation of X, and then solve the corresponding deterministic optimization problem. Therefore, generally and practically, $\hat{R}_M(\hat{z}_N)$ is always larger than or equal to $\hat{R}_M^*$, even though we have a sufficiently large number of data. That is, P always has an asymptotic upper bound that is less than or equal to 1 for any possible CSO problem, i.e., $\lim_{N,M\to\infty} P(N;M) \leq 1$ almost surely. This upper bound may be related to the characteristic of the CSO problem at hand, and thus varies across different problems. Specifically, it may be related to the underlying true joint distribution $\mu_{X,Y}$ of (X,Y) and the dataset that we have, making it unknown to us.

Therefore, it may be problematic to compare the P values of a specific prescriptive analytics method for two particular CSO problems (see Example 1). Similarly, it may also be problematic to compare the means of P values for a series of CSO problems (generalizability) for two prescriptive analytics methods (see Example 2).

Example 1. Consider addressing two contextual newsvendor problems: for a given observation $X = x$, the random demands in Problems 1 and 2 are assumed to follow uniform distributions $Y_1 \sim U(5,15)$ and $Y_2 \sim U(9,11)$, respectively. For each problem, a prescriptive analytics method is applied, with the coefficients of prescriptiveness assumed to be $P_1 = 0.4$ and $P_2 = 0.6$, respectively. Does this mean that this method performs better on Problem 2 than on Problem 1 (i.e., the corresponding predictive

prescription $\hat{z}_N(\cdot)$ is more efficient in leveraging the data on X for Problem 2)? This is not necessarily the case. Assume that the asymptotic upper bounds of P for these two problems are 0.4 and 0.8, respectively (note that Problem 1 seems "harder" than Problem 2, due to its larger variance of Y given $X = x$). In this case, we can easily see that for Problem 1, this method provides prescription $\hat{z}_N(x) = 10$ (which is equal to the full-information optimal decision $z^*(x) = 10$, supposing that the per-unit overage cost is equal to the per-unit underage cost) and achieves an efficiency of 100% (= 0.4/0.4) in leveraging the data on X for prescribing a data-driven decision (i.e., P_1 reaches its asymptotic upper bound, likely due to the greater availability of data in Problem 1). However, for Problem 2, this method only achieves an efficiency of 75% (= 0.6/0.8). Therefore, this method performs better on Problem 1 in this case.

Example 2. Consider solving a series of CSO problems with two prescriptive analytics methods, namely, Method 1 and Method 2 (e.g., point-prediction and wSAA). Let $\bar{P}_1$ and $\bar{P}_2$ be the means of the coefficients of prescriptiveness for these problems obtained by the two methods, respectively, and suppose $\bar{P}_1 < \bar{P}_2$. Does this mean that Method 2 performs better on average than Method 1 for these problems (i.e., Method 2 is more efficient on average in leveraging the data on X for these problems)? This is not necessarily the case. For example, suppose that three CSO problems are solved using Methods 1 and 2, with the corresponding P values (denoted by P_1 and P_2, respectively) assumed to be as follows.

- Problem A (the asymptotic upper bound of P is 0.1): $P_1 = 0.1$ and $P_2 = 0.01$;
- Problem B (the asymptotic upper bound of P is 0.1): $P_1 = 0.1$ and $P_2 = 0.01$;
- Problem C (the asymptotic upper bound of P is 0.9): $P_1 = 0.7$ and $P_2 = 0.9$.

It is evident that $\bar{P}_1 = 0.9/3 < \bar{P}_2 = 0.92/3$. However, from the asymptotic upper bounds of the three problems, we can easily see that (i) for Problems A and B, Method 1 achieves an efficiency of 100% (= 0.1/0.1) in leveraging the data on X, whereas Method 2 only reaches 10% (= 0.01/0.1); and (ii) for Problem C, Method 1 achieves an efficiency of 77.8% ($\approx$ 0.7/0.9), while Method 2 reaches 100% (= 0.9/0.9). Therefore, the average efficiency of Method 1 (92.6%) is greater than that of Method 2 (40%). That is, in this case, Method 1 performs better on average for these problems.

From the above two examples, we can clearly see that there exist some shortcomings to applying P values when solving two or more CSO problems. These shortcomings always and definitely exist as long as we calculate P values based on the given datasets s_N and $\bar{s}_M$. This leads to the following question:

Q1. What metric can we use to compare the efficacy of a specific method across multiple CSO problems fairly, or to compare the generalizability of multiple prescriptive analytics methods across multiple CSO problems fairly?

To formally address Q1, we propose a novel metric, the adjusted coefficient of prescriptiveness, in Section 3.1.

2.3.2. The data-dependent deficiency

As mentioned in Section 1, the insights derived from the original coefficient of prescriptiveness P are contingent on specific training or test datasets. In this section, we present an illustrative example (Example 3) and propose our research question accordingly.

Example 3. In a particular CSO problem, suppose we have T ($T \in \mathbb{Z}_{++}$) training datasets of the same size N, denoted by $s_{N,1}, s_{N,2}, ..., s_{N,T}$. Consider that two prescriptive analytics methods (e.g., point-prediction and wSAA) are trained on the training dataset $s_{N,t}$ ($t \in \{1, ..., T\}$), with the corresponding predictive prescriptions denoted by $\hat{z}^1_{N,t}(\cdot)$ and $\hat{z}^2_{N,t}(\cdot)$, respectively. Given the same new observation $X = x$, it is evident that for two different training datasets $s_{N,t}$ and $s_{N,t'}$ ($t \neq t'$ and $t, t' \in \{1, ..., T\}$), we may have $\hat{z}^1_{N,t}(x) \neq \hat{z}^1_{N,t'}(x)$ and $\hat{z}^2_{N,t}(x) \neq \hat{z}^2_{N,t'}(x)$. Moreover, if we have a large number of test data, it is also possible that on a training dataset $s_{N,t}$, $\hat{z}^1_{N,t}(\cdot)$ may perform better (i.e., with a higher P value) than $\hat{z}^2_{N,t}(\cdot)$, whereas on a different training dataset $s_{N,t'}$, $\hat{z}^2_{N,t'}(\cdot)$ may perform better than $\hat{z}^1_{N,t'}(\cdot)$. Therefore, we are interested in knowing which method has the best average performance (evaluated with a large number of test data) over T training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$.

Q2. Given a large number of real-world data, which method, from the list of candidate prescriptive analytics methods, has the best average performance (evaluated with a large number of test data) over all available training datasets of a given size for a particular CSO problem?

To formally address Q2, we propose our novel metric, the expected coefficient of prescriptiveness, and its estimate for practical applications in Section 3.3.

3. Novel coefficients of prescriptiveness

In this section, we propose the adjusted coefficient of prescriptiveness in Section 3.1 and its estimate in Section 3.2. Then, Section 3.3 introduces the expected coefficient of prescriptiveness and its estimate.

3.1. The adjusted coefficient of prescriptiveness

In this section, we address Q1: What metric can we use to compare the efficacy of a specific method across multiple CSO problems fairly, or to compare the generalizability of multiple prescriptive analytics methods across multiple CSO problems fairly?

As mentioned in Section 2.3.1, in the definition of P, $\hat{R}^*_M$ is the benchmark for measuring the efficacy of a predictive prescription $\hat{z}_N(\cdot)$; however, it is always smaller (better) than or equal to $\hat{R}_M(\hat{z}_N)$, and thus we have $\lim_{N,M\to\infty} P(N; M) \leq 1$ almost surely. Therefore, the inherent and unknown asymptotic upper bound (less than or equal to 1) of P always exists.

Specifically, for each test data sample $(\bar{x}_j, \bar{y}_j)$ ($j \in \{1, ..., M\}$) in $\bar{s}_M$, the perfect-foresight oracle makes a decision $z^{\text{fore}}(\bar{y}_j)$ knowing $Y = \bar{y}_j$ without any uncertainty, leading to the cost of $\min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$. However, for a potentially "excellent" prescriptive analytics method, the best it can achieve (supposing we have infinite training data, i.e., $N \to \infty$, and some mild conditions hold) is to identify the underlying true conditional distribution $\mu_{Y|X=\bar{x}_j}$ of Y given $X = \bar{x}_j$, and provide the best predictive prescription, denoted by $\hat{z}^{\text{best}}_N(\bar{x}_j)$, which approximates the full-information optimal decision $z^*(\bar{x}_j)$ (see Formula (1.4)) rather than $z^{\text{fore}}(\bar{y}_j)$. As M approaches infinity, the estimated out-of-sample cost $\hat{R}_M(\hat{z}^{\text{best}}_N) := \frac{1}{M}\sum_{j=1}^{M} c(\hat{z}^{\text{best}}_N(\bar{x}_j); \bar{y}_j)$ (see Formula (2.4)) approaches the full-information optimum $\frac{1}{M}\sum_{j=1}^{M}\{\min_{z\in\mathscr{Z}} \mathbb{E}[c(z; Y)|X = \bar{x}_j]\}$ instead of the deterministic perfect-foresight counterpart $\frac{1}{M}\sum_{j=1}^{M} \min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$ (i.e., $\hat{R}^*_M$). That is, $\lim_{N,M\to\infty} \hat{R}_M(\hat{z}^{\text{best}}_N) = \lim_{M\to\infty} \frac{1}{M}\sum_{j=1}^{M}\{\min_{z\in\mathscr{Z}} \mathbb{E}[c(z; Y)|X = \bar{x}_j]\} \geq \lim_{M\to\infty} \hat{R}^*_M$ (recall that $\hat{R}_M(\hat{z}_N)$ is generally larger than or equal to $\hat{R}^*_M$).

Therefore, we propose that it is reasonable to use the full-information optimum as the benchmark to measure the efficacy of a predictive prescription; however, this can be done only for situations in which we know the true joint distribution $\mu_{X,Y}$ of (X, Y) and have an infinite number of training data. This is because (i) only when we know the joint distribution can we obtain the full-information optimum; and (ii) only when we have infinite training data ($N \to \infty$) can we guarantee that $\hat{z}_N^{\text{best}}(\cdot)$ approaches $z^*(\cdot)$ (note that we omit the feature x for convenience). However, these two conditions are hard to achieve in practice. Therefore, we can only use this benchmark in a theoretical manner, in which we assume that we have full knowledge of the joint distribution and, therefore, can also generate infinite synthetic training data in theory. Note that it is not necessary to generate synthetic test data because we can theoretically evaluate the out-of-sample costs of all decisions based on the assumed joint distribution.

Remark 1. Here, note that knowing the joint distribution $\mu_{X,Y}$ of (X, Y) does not mean that Y is measurable with respect to X. Recall that Y is measurable with respect to X if and only if there exists a deterministic function f such that $Y = f(X)$ without any uncertainties. However, for a general joint distribution $\mu_{X,Y}$, given X, the target Y is still a random variable that depends on the conditional distribution $\mu_{Y|X}$.

Based on the assumed joint distribution $\mu_{X,Y}$ of (X, Y), we can obtain the full-information optimal decision $z^*(\cdot)$ by solving optimization problem (1.4), and the corresponding out-of-sample cost (i.e., the full-information optimum) is defined as follows.

Definition 2 (Out-of-sample cost of the full-information optimal solution)**.** *The absolute performance of the full-information optimal decision $z^*(\cdot)$ is represented by its out-of-sample cost (i.e., the full-information optimum):*

$$R(z^*) = \mathbb{E}_{X\sim\mu_X}[\mathbb{E}_{Y\sim\mu_{Y|X}}[c(z^*(X); Y)|X]] = \mathbb{E}_{(X,Y)\sim\mu_{X,Y}}[c(z^*(X); Y)], \tag{3.1}$$

where $z^*(X) \in \arg\min_{z\in\mathscr{Z}} \mathbb{E}_{Y\sim\mu_{Y|X}}[c(z; Y)|X]$, which is defined in Formula (1.4) (provided that this stochastic optimization model can be solved to optimality or with high accuracy). Note that $R(z^*)$ can be obtained by taking the expectation of $v^*(x)$ (see Formula (1.3)) over X.

Similarly, based on the assumed joint distribution $\mu_{X,Y}$ of (X, Y) and a generated training dataset s_N containing infinite data samples (i.e., $N \to \infty$) in theory, we can define the out-of-sample costs of a predictive prescription $\hat{z}_N(\cdot)$ and the data-poor SAA solution $\hat{z}_N^{\text{SAA}}$ as follows.

Definition 3 (Out-of-Sample cost of a predictive prescription)**.** *The absolute performance of a predictive prescription $\hat{z}_N(\cdot)$ (considering feature X) trained on a generated training dataset s_N ($N \to \infty$) is represented by its out-of-sample cost:*

$$R(\hat{z}) = \lim_{N\to\infty} \mathbb{E}_{X\sim\mu_X}[\mathbb{E}_{Y\sim\mu_{Y|X}}[c(\hat{z}_N(X); Y)|X]] = \lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mu_{X,Y}}[c(\hat{z}_N(X); Y)]. \tag{3.2}$$

Definition 4 (Out-of-Sample cost of the SAA solution)**.** *The absolute performance of the data-poor SAA solution $\hat{z}_N^{\text{SAA}}$ (solely based on Y data and not considering feature X) trained on a generated training dataset s_N ($N \to \infty$) is represented by its out-of-sample cost:*

$$R(\hat{z}^{\text{SAA}}) = \lim_{N\to\infty} \mathbb{E}_{Y\sim\mu_Y}[c(\hat{z}_N^{\text{SAA}}; Y)]. \tag{3.3}$$

Here, note that $\hat{R}_M(\hat{z}_N)$ and $\hat{R}_M(\hat{z}_N^{\mathrm{SAA}})$ in Formulas (2.4) and (2.5) are all estimates of $R(\hat{z})$ and $R(\hat{z}^{\mathrm{SAA}})$, respectively, based on a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$.

Remark 2. Theoretically, in the above three definitions, the three expectations $R(z^*)$, $R(\hat{z})$, and $R(\hat{z}^{\mathrm{SAA}})$ can be obtained analytically through multi-dimensional integration based on the assumed joint distribution $\mu_{X,Y}$, and are not dependent on a test dataset. Therefore, here, we do not need to generate a test dataset and evaluate the out-of-sample costs of $z^*(\cdot)$, $\hat{z}_N(\cdot)$, and $\hat{z}_N^{\mathrm{SAA}}$ numerically (however, this can be used as an approximate numerical computation method, which we discuss in Section 3.2). Recall that the perfect-foresight decision $z^{\mathrm{fore}}(\cdot)$ and the corresponding out-of-sample cost $\hat{R}_M^*$ rely solely on a realized test dataset; therefore, without a realized test dataset, $\hat{R}_M^*$ is non-existent, and thus so is P.

Therefore, with an infinite number of training data generated from an assumed joint distribution, we introduce the adjusted coefficient of prescriptiveness to formally address Q1, which is defined as follows.

Definition 5 (The adjusted coefficient of prescriptiveness)**.** *Based on an assumed joint distribution $\mu_{X,Y}$ and a generated training dataset s_N containing infinite data samples ($N \to \infty$), the adjusted coefficient of prescriptiveness for a predictive prescription $\hat{z}_N(\cdot)$ is defined as follows:*

$$P_{\mathrm{adj}} = 1 - \frac{R(\hat{z}) - R(z^*)}{R(\hat{z}^{\mathrm{SAA}}) - R(z^*)} = \frac{R(\hat{z}^{\mathrm{SAA}}) - R(\hat{z})}{R(\hat{z}^{\mathrm{SAA}}) - R(z^*)}. \tag{3.4}$$

Here, we assume that there is no feasible solution $\bar{z} \in \mathscr{Z}$ such that $\mathbb{E}[c(\bar{z}; Y)] = \min_{z\in\mathscr{Z}} \mathbb{E}_{Y\sim\mu_{Y|X}}[c(z; Y)|X]$ holds for μ_X-almost-everywhere $X \in \mathscr{X}$, thereby ensuring that $R(\hat{z}^{\mathrm{SAA}}) \neq R(z^*)$ and P_{adj} is well-defined.

Remark 3. The above assumption also excludes the no-contextual-value case in which the full-information optimum coincides with the data-poor stochastic optimum. For example, if Y is independent of X, then the conditional distribution $\mu_{Y|X=x}$ equals the marginal distribution μ_Y for all x, and the full-information decision $z^*(x)$ reduces to the same decision as the data-poor SAA solution in the asymptotic limit. In that boundary case, $R(\hat{z}^{\mathrm{SAA}}) - R(z^*) = 0$, so P_{adj} and its estimate are not stable normalized scores. Therefore, when contextual information has no value, one should report the raw out-of-sample costs or the original coefficient P with the deterministic perfect-foresight benchmark, for which a consistent method has a numerator approaching zero.

This coefficient is a unitless quantity bounded above by the value of 1. P_{adj} is a theoretical metric, based on s_N ($N \to \infty$), which represents the fraction of the way that knowledge of X, leveraged in a predictive prescription $\hat{z}_N(\cdot)$, takes us from making a decision under full uncertainty about the value of Y (i.e., data-poor SAA solution $\hat{z}_N^{\mathrm{SAA}}$) to making a decision with full information of the underlying true joint distribution $\mu_{X,Y}$ of (X, Y) (i.e., the full-information optimal decision $z^*(\cdot)$). Therefore, the adjusted coefficient of prescriptiveness P_{adj} measures the efficacy of a method, based on an assumed joint distribution $\mu_{X,Y}$ and a training dataset s_N ($N \to \infty$) generated from it, compared with the full-information optimum. A low P_{adj} indicates that $\hat{z}_N(\cdot)$ trained on s_N ($N \to \infty$) is inefficient in leveraging the data on X for prescribing a data-driven decision in a problem. In contrast, a high P_{adj} indicates that $\hat{z}_N(\cdot)$ trained on s_N ($N \to \infty$) is efficient in leveraging the data on X and significantly reduces the out-of-sample cost.

Remark 4. Here, note that the definition of the adjusted coefficient of prescriptiveness P_{adj} is based on an assumed joint distribution $\mu_{X,Y}$ and a generated training dataset s_N ($N \to \infty$) containing infinite data samples, and is not dependent on a test dataset. In contrast, the coefficient of prescriptiveness P is defined based on a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$ (N and M are finite). Therefore, the prerequisites for these two coefficients P and P_{adj} are intrinsically different; thus, they are not comparable in theory.

For some potential predictive prescriptions $\hat{z}_N(\cdot)$ (e.g., $\hat{z}_N^{\text{best}}(\cdot)$) and under appropriate conditions [4], as N approaches infinity, we have $R(\hat{z}) = R(z^*)$; hence, in this case, if we generate an infinite number of training data (i.e., $N \to \infty$) based on the assumed joint distribution, we will have $P_{\text{adj}} = 1$ almost surely. This means that P_{adj} theoretically always has the same asymptotic upper bound of 1 for all possible CSO problems. Therefore, with the adjusted coefficient of prescriptiveness defined in Formula (3.4), we can, in theory, fairly compare both the efficacy of a specific prescriptive analytics method across multiple CSO problems and the generalizability of multiple prescriptive analytics methods (i.e., Q1 is addressed in theory).

3.2. The estimate of the adjusted coefficient of prescriptiveness

The computation and application of P_{adj} face two challenges. First, it is unclear how to achieve an infinite number of synthetic training data (i.e., $N \to \infty$) in reality, as $N \to \infty$ is a necessary condition to guarantee the asymptotic upper bound of 1 for P_{adj}. Second, although we can obtain the values of the three expectations $R(z^*)$, $R(\hat{z})$, and $R(\hat{z}^{\text{SAA}})$ in P_{adj} analytically through multi-dimensional integration based on the assumed $\mu_{X,Y}$, this may be very challenging and time-consuming for some problems or distributions, causing low efficiency in the computational process of P_{adj} and thus limiting its application.

Therefore, to achieve efficient computational experiments and real-world applications, we propose the estimate of the adjusted coefficient of prescriptiveness to address Q1 in practice. The estimate can be obtained through an efficient numerical computation method, detailed as follows.

(1) First, based on the assumed joint distribution $\mu_{X,Y}$ of (X, Y), we generate a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$, each containing a sufficiently large number (e.g., a billion) of data samples.

(2) Then, for each test sample $(\bar{x}_j, \bar{y}_j)$ ($j \in \{1, ..., M\}$) in $\bar{s}_M$, we calculate $\hat{z}_N(\bar{x}_j)$ and $\hat{z}_N^{\text{SAA}}$ (the same for all test samples) based on s_N, and accordingly calculate $\hat{R}_M(\hat{z}_N)$ and $\hat{R}_M(\hat{z}_N^{\text{SAA}})$ using Formulas (2.4) and (2.5), respectively.

(3) Next, we calculate an estimate of $R(z^*)$ based on $\bar{s}_M$, denoted by $\hat{R}_M(z^*)$:

$$\hat{R}_M(z^*) = \frac{1}{M} \sum_{j=1}^{M} c(z^*(\bar{x}_j); \bar{y}_j), \tag{3.5}$$

where $z^*(\bar{x}_j) \in \arg\min_{z \in \mathscr{Z}} \mathbb{E}_{Y \sim \mu_{Y|X=\bar{x}_j}}[c(z; Y)|X = \bar{x}_j]$ ($j \in \{1, ..., M\}$) (defined in Formula (1.4)). There are generally two approaches to calculating the expected cost $\mathbb{E}[c(z; Y)|X = \bar{x}_j]$ ($j \in \{1, ..., M\}$) for obtaining $z^*(\bar{x}_j)$:

- Assess $\mathbb{E}[c(z; Y)|X = \bar{x}_j]$ ($j \in \{1, ..., M\}$) analytically through multi-dimensional integration based on $\mu_{Y|X=\bar{x}_j}$, which may be very challenging for some problems or distributions, and then solve the corresponding deterministic optimization problem to exactly obtain $z^*(\bar{x}_j)$.

- Take a sufficiently large number of data samples $\{\bar{\bar{y}}_{j,1}, ..., \bar{\bar{y}}_{j,U}\}$ ($U \in \mathbb{Z}_{++}$) of Y from the conditional distribution $\mu_{Y|X=\bar{x}_j}$ of Y given $X = \bar{x}_j$, and solve the resulting SAA problem:

$$\hat{z}_U^*(\bar{x}_j) \in \arg\min_{z\in\mathscr{Z}} \frac{1}{U}\sum_{u=1}^{U} c(z; \bar{\bar{y}}_{j,u}), \quad j \in \{1, ..., M\}, \tag{3.6}$$

  where $\hat{z}_U^*(\bar{x}_j)$ is an estimate of $z^*(\bar{x}_j)$. The strong law of large numbers implies that for a random variable, as the sample size approaches infinity, the sample mean converges to the true mean with a probability of 1. Therefore, we have $\lim_{U\to\infty} \hat{z}_U^*(\bar{x}_j) = z^*(\bar{x}_j)$ almost surely.

Similarly, we also have $\lim_{M\to\infty} \hat{R}_M(z^*) = R(z^*)$, $\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N) = R(\hat{z})$, and $\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) = R(\hat{z}^{\mathrm{SAA}})$ almost surely.

(4) Finally, we can obtain an estimate of the adjusted coefficient of prescriptiveness, which is defined as follows.

Definition 6 (The estimate of the adjusted coefficient of prescriptiveness)**.** *Based on a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$ (N and M are sufficiently large numbers) generated from an assumed joint distribution $\mu_{X,Y}$ of (X, Y), the estimate of the adjusted coefficient of prescriptiveness for a predictive prescription $\hat{z}_N(\cdot)$ is defined as follows:*

$$\hat{P}_{\mathrm{adj}} = \frac{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(\hat{z}_N)}{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(z^*)}. \tag{3.7}$$

Similarly, we have $\lim_{N,M\to\infty} \hat{P}_{\mathrm{adj}}(N; M) = P_{\mathrm{adj}}$ almost surely, which means that based on the assumed joint distribution $\mu_{X,Y}$ of (X, Y), as long as N and M are large enough, $\hat{P}_{\mathrm{adj}}$ can approach P_{adj} with arbitrary accuracy almost surely. Note that because $\hat{P}_{\mathrm{adj}}$ is also calculated using a fixed training dataset s_N and a fixed test dataset $\bar{s}_M$ (although we need, in addition, to obtain the full-information optimal decision based on an assumed $\mu_{X,Y}$), P and $\hat{P}_{\mathrm{adj}}$ are comparable in practice.

As mentioned in Section 3.1, P_{adj} always has the same asymptotic upper bound of 1 for all possible CSO problems; however, this may not hold for $\hat{P}_{\mathrm{adj}}$, as N and M are fixed and finite albeit very large. Can we use $\hat{P}_{\mathrm{adj}}$ to address Q1 in practice? The answer is yes, for the following reasons.

That P_{adj} always has an asymptotic upper bound of 1 is a convergence property in theory with N approaching infinity. Therefore, the asymptotic upper bound of $\hat{P}_{\mathrm{adj}}$ may not exactly take the value of 1, because $\hat{P}_{\mathrm{adj}}$ is obtained numerically. However, it is reasonable to suppose that for a relatively large N and a relatively large M, there always exists a potentially "excellent" prescriptive analytics method whose value of $\hat{P}_{\mathrm{adj}}$ can approach 1 with arbitrary accuracy almost surely. Here, a relatively large N ensures that the training dataset s_N contains relatively abundant information about (X, Y) for a specific CSO problem, while a relatively large M ensures that the test dataset $\bar{s}_M$ is relatively effective in evaluating the estimated out-of-sample costs of all decisions. Therefore, from the perspective of computational experiments and practical applications, we can use $\hat{P}_{\mathrm{adj}}$ to address Q1 in practice. See the following Example 4.

Example 4. In a specific CSO problem without any real-world data, we first assume a joint distribution $\mu_{X,Y}$ of (X, Y), and then generate datasets s_N and $\bar{s}_M$, each containing a million data samples. After some attempts, we finally identify a potentially "excellent" prescriptive analytics method with its $\hat{P}_{\mathrm{adj}} =$

0.9999 (this possibility does exist). Therefore, from the perspective of computational experiments and practical applications, we have the confidence to address Q1 in practice. In contrast, based on the same datasets s_N and $\bar{s}_M$, we may have $P = 0.4999$ for the same potentially "excellent" prescriptive analytics method (here, the asymptotic upper bound of P may be approximately 0.5 for this CSO problem). Additionally, if we need higher precision (note that the computational cost would increase accordingly), we may set larger N and M (e.g., ten million), and would expect $\hat{P}_{\text{adj}}$ to get closer to the value of 1 (e.g., $\hat{P}_{\text{adj}} = 0.999999$) for this potentially "excellent" prescriptive analytics method.

Example 5 (A newsvendor illustration of the coefficients)**.** Consider the newsvendor cost

$$c(z; y) = c_o(z - y)^+ + c_u(y - z)^+,$$

where $c_o > 0$ and $c_u > 0$. If the conditional distribution $F_{Y|X=x}$ is known, the full-information optimal decision is the conditional critical fractile

$$z^*(x) = F^{-1}_{Y|X=x}\left(\frac{c_u}{c_o + c_u}\right).$$

In contrast, the data-poor SAA decision uses only the marginal information of Y and, asymptotically, targets the unconditional critical fractile $F_Y^{-1}(c_u/(c_o + c_u))$. The deterministic perfect-foresight benchmark is even stronger: after observing the realized demand y, it can choose $z = y$ and therefore attain zero realized newsvendor loss. Hence, in the newsvendor problem, the original coefficient P normalizes the improvement of a prescription against an ex-post perfect-foresight oracle, whereas P_{adj} and $\hat{P}_{\text{adj}}$ normalize it against the attainable full-information decision that knows $F_{Y|X=x}$ but not the realized value of Y. Similarly, $\hat{P}^E$ averages this evaluation over multiple training datasets of the same size, thereby measuring the average performance of a method rather than its performance on a single realized training set.

3.3. Expected coefficients of prescriptiveness

In this section, we address Q2: Given a large number of real-world data, which method, from the list of candidate prescriptive analytics methods, has the best average performance (evaluated with a large number of test data) over all available training datasets of a given size for a particular CSO problem?

Before introducing the expected coefficient of prescriptiveness (a theoretical metric that is hard to apply in practice), we first propose its estimate, i.e., the estimate of the expected coefficient of prescriptiveness, which can be obtained numerically and effectively for practical applications through the following detailed computational process.

(1) First, based on a large number of real-world data, we randomly build T ($T \in \mathbb{Z}_{++}$) training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$ of a given size N and a test dataset $\bar{s}_M$. Here, T and M are sufficiently large numbers. Considering the solution efficiency, N can be determined flexibly according to the complexity of the specific CSO problem to be solved and the prescriptive analytics method that we use.

(2) Then, for the training dataset $s_{N,t}$ ($t \in \{1, ..., T\}$), let $\hat{z}_{N,t}(\cdot)$ be the resulting predictive prescription and $\hat{z}^{\text{SAA}}_{N,t}$ be the resulting data-poor SAA solution. Correspondingly, let $\hat{R}_M(\hat{z}_{N,t})$ and $\hat{R}_M(\hat{z}^{\text{SAA}}_{N,t})$ be the estimated out-of-sample costs (based on $\bar{s}_M$) of $\hat{z}_{N,t}(\cdot)$ and $\hat{z}^{\text{SAA}}_{N,t}$, respectively, which can be calculated using Formulas (2.4) and (2.5).

(3) Next, we calculate $\hat{R}^*_M$ based on $\bar{s}_M$ using Formula (2.6).
(4) Finally, we obtain the estimate of the expected coefficient of prescriptiveness, which is defined as follows.

Definition 7 (The estimate of the expected coefficient of prescriptiveness)**.** *Based on T training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$ of a given size N and a test dataset $\bar{s}_M$ (T and M are sufficiently large numbers), the estimate of the expected coefficient of prescriptiveness is defined as follows:*

$$\hat{P}^E = \frac{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_M(\hat{z}^{\mathrm{SAA}}_{N,t}) - \frac{1}{T}\sum_{t=1}^{T}\hat{R}_M(\hat{z}_{N,t})}{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_M(\hat{z}^{\mathrm{SAA}}_{N,t}) - \hat{R}^*_M}. \tag{3.8}$$

This coefficient is a unitless quantity bounded above by the value of 1. $\hat{P}^E$ measures the average performance (on $\bar{s}_M$) of a predictive prescription $\hat{z}_N(\cdot)$ across T training datasets of a given size N, compared with an oracle that can solve the deterministic perfect-foresight counterpart problem. A low $\hat{P}^E$ denotes that $\hat{z}_N(\cdot)$ is inefficient on average in leveraging the data on X for prescribing a data-driven decision in a problem. In contrast, a high $\hat{P}^E$ denotes that $\hat{z}_N(\cdot)$ is efficient on average in leveraging the data on X and significantly reduces costs.

In practice, the value of T can be chosen by a sequential stability rule. One may start with a moderate number of training datasets, for example, $T = 20$ or 30, compute $\hat{P}^E_T$, and then increase T in blocks of size b. The computation may stop once the stabilized relative change $|\hat{P}^E_T - \hat{P}^E_{T-b}|/(|\hat{P}^E_T| + \epsilon_0)$ is below a prescribed tolerance, where $\epsilon_0 > 0$ is a small constant used only to avoid division by a value close to zero, or once the empirical standard error of the averaged numerator and denominator is sufficiently small for the intended decision analysis. When only one large dataset is available, bootstrap resampling or repeated subsampling can be used as a computationally cheaper approximation to independently generated training datasets. This approximation is useful for sensitivity analysis, but it is not identical to drawing new independent samples from the population and may underestimate variability when the original dataset is small or dependent.

Note that if we integrate all samples of T training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$ and thus construct a huge training dataset $s_{N\cdot T}$ of size $N \cdot T$, we may also calculate a P value based on $s_{N\cdot T}$ and $\bar{s}_M$, which is expected to be approximately the same as $\hat{P}^E$. However, considering that T is a sufficiently large number, it may be challenging and time-consuming, or even unachievable with reasonable time and computational resources, to train many ML models on a training dataset of a huge size $N \cdot T$, making the solution of the downstream optimization problem also impossible. Therefore, $\hat{P}^E$ is more practical, and to apply $\hat{P}^E$ effectively, we can set a reasonable value for N according to our specific CSO problem and data-driven method.

Therefore, for a list of candidate prescriptive analytics methods, using a large number of real-world data and $\hat{P}^E$, we can more fairly evaluate their average performances over all available training datasets of a given size and obtain more general insights for a particular CSO problem (i.e., Q2 is addressed).

Additionally, the $\hat{}$ notation above $\hat{P}^E$ indicates that $\hat{P}^E$ is an estimate based on the realized real-world data. The primitive value of $\hat{P}^E$ is denoted by P^E, i.e., the expected coefficient of prescriptiveness, whose definition is given in the following remark.

Remark 5. Let S_N be a random training sample of N i.i.d. draws of (X, Y) from the (unknown) underlying true joint distribution $\mu_{X,Y}$, $\hat{z}_N(\cdot|S_N)$ be the resulting predictive prescription, and $\hat{z}^{\mathrm{SAA}}_N(S_N)$ be the resulting data-poor SAA solution. Correspondingly, let

$R(\hat{z}_N(\cdot|S_N)) := \mathbb{E}_{(X,Y)\sim\mu_{X,Y}}[c(\hat{z}_N(X|S_N);Y)]$ and $R(\hat{z}_N^{\mathrm{SAA}}(S_N)) := \mathbb{E}_{Y\sim\mu_Y}[c(\hat{z}_N^{\mathrm{SAA}}(S_N);Y)]$ be the out-of-sample cost of $\hat{z}_N(\cdot|S_N)$ and $\hat{z}_N^{\mathrm{SAA}}(S_N)$, respectively. Note that $R(\hat{z}_N(\cdot|S_N))$ and $R(\hat{z}_N^{\mathrm{SAA}}(S_N))$ are random variables that depend on the random training sample S_N. Taking expectations over the sampling distribution of S_N, the expected coefficient of prescriptiveness is defined as

$$P^E = \frac{\mathbb{E}[R(\hat{z}_N^{\mathrm{SAA}}(S_N))] - \mathbb{E}[R(\hat{z}_N(\cdot|S_N))]}{\mathbb{E}[R(\hat{z}_N^{\mathrm{SAA}}(S_N))] - \mathbb{E}_{Y\sim\mu_Y}[\min_{z\in\mathscr{Z}} c(z;Y)]}. \tag{3.9}$$

Note that P^E is a theoretical metric and we cannot obtain its value in the context of Q2, because we do not know the underlying true joint distribution $\mu_{X,Y}$ of (X,Y). Instead, we only have a large number of real-world data, and thus, the value of $\hat{P}^E$ can be calculated.

Under the same size N of the training dataset, $\hat{P}^E$ is an estimate of P^E based on T training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$ and a test dataset $\bar{s}_M$ (T and M are sufficiently large numbers). As expressed in Formula (2.6), the strong law of large numbers implies that $\lim_{M\to\infty}\hat{R}_M^* = \mathbb{E}_{Y\sim\mu_Y}[\min_{z\in\mathscr{Z}} c(z;Y)]$ almost surely. Similarly, we have $\lim_{T,M\to\infty}\frac{1}{T}\sum_{t=1}^{T}\hat{R}_M(\hat{z}_{N,t}) = \mathbb{E}[R(\hat{z}_N(\cdot|S_N))]$ and $\lim_{T,M\to\infty}\hat{R}_M(\hat{z}_{N,t}^{\mathrm{SAA}}) = \mathbb{E}[R(\hat{z}_N^{\mathrm{SAA}}(S_N))]$ almost surely, and thus under the same N, we have $\lim_{T,M\to\infty}\hat{P}^E(T;N;M) = P^E(N)$ almost surely, which means that based on a large number of real-world data, as long as T and M are large enough, $\hat{P}^E$ can approach P^E with arbitrary accuracy almost surely.

Similarly, under the same size N of the training dataset, P is also an estimate of P^E based on a training dataset s_N and a test dataset $\bar{s}_M$, and we have $\lim_{N,M\to\infty}P(N;M) = \lim_{N\to\infty}P^E(N)$ almost surely.

Although P and $\hat{P}^E$ are both estimates of P^E (we have $\lim_{N,M\to\infty}P(N;M) = \lim_{T,N,M\to\infty}\hat{P}^E(T;N;M) = \lim_{N\to\infty}P^E(N)$ almost surely), $\hat{P}^E$ is more general than P, because $\hat{P}^E$ considers the average performance across T training datasets and is evaluated on a test dataset of a large size M. Therefore, in practice, if we have a large number of real-world data, we can not only calculate P but also calculate a more general metric $\hat{P}^E$ to more fairly evaluate and compare the efficacy of candidate prescriptive analytics methods for a particular CSO problem (however, if there are limited data, we can only use P).

Remark 6. Naturally, similar to the relationship between P and P^E, regarding P_{adj}, we may define the expected adjusted coefficient of prescriptiveness. Indeed, based on an assumed joint distribution $\mu_{X,Y}$ and a generated random training dataset S_N containing infinite data samples ($N\to\infty$), we define $R(\hat{z}(\cdot|S)) := \lim_{N\to\infty}R(\hat{z}_N(\cdot|S_N))$ and $R(\hat{z}^{\mathrm{SAA}}(S)) := \lim_{N\to\infty}R(\hat{z}_N^{\mathrm{SAA}}(S_N))$, and the expected adjusted coefficient of prescriptiveness has the following form:

$$P_{\mathrm{adj}}^E = \frac{\mathbb{E}[R(\hat{z}^{\mathrm{SAA}}(S))] - \mathbb{E}[R(\hat{z}(\cdot|S))]}{\mathbb{E}[R(\hat{z}^{\mathrm{SAA}}(S))] - R(z^*)}, \tag{3.10}$$

where the two expectations $\mathbb{E}[R(\hat{z}(\cdot|S))]$ and $\mathbb{E}[R(\hat{z}^{\mathrm{SAA}}(S))]$ are taken over the sampling distribution of S_N ($N\to\infty$). However, for P_{adj}, the synthetic training dataset s_N already contains infinite data samples ($N\to\infty$), which means that we would get the same results if we took expectations over the sampling distribution of S_N containing infinite data samples, i.e., $\mathbb{E}[R(\hat{z}(\cdot|S))] = R(\hat{z})$ and $\mathbb{E}[R(\hat{z}^{\mathrm{SAA}}(S))] = R(\hat{z}^{\mathrm{SAA}})$. Therefore, P_{adj}^E is equivalent to P_{adj}.

4. Theoretical analysis

In this section, we provide theoretical analyses of the coefficients of prescriptiveness discussed above. First, we summarize the relationships between these coefficients, which measure the efficacy of prescriptive analytics methods under different circumstances, as follows.

Remark 7 (A summary of the relationships between the coefficients)**.**

- In terms of P, P_{adj}, and $\hat{P}_{\text{adj}}$, we have:
  (1) P and P_{adj} are intrinsically different, thus, they are not comparable in theory. For some potential predictive prescriptions $\hat{z}_N(\cdot)$ (e.g., $\hat{z}_N^{\text{best}}(\cdot)$) and under appropriate conditions, we have $\lim_{N,M\to\infty} P(N;M) \leq 1$ and $P_{\text{adj}} = 1$ almost surely.
  (2) $\hat{P}_{\text{adj}}$ is an estimate of P_{adj} based on fixed datasets s_N and $\bar{s}_M$ (N and M are sufficiently large numbers), and we have $\lim_{N,M\to\infty} \hat{P}_{\text{adj}}(N;M) = P_{\text{adj}}$ almost surely. Note that P_{adj} is a theoretical metric and hard to apply in practice, while $\hat{P}_{\text{adj}}$ is a numerical metric and easy to calculate for real-world applications. Additionally, P and $\hat{P}_{\text{adj}}$ are comparable in practice.
- In terms of P, P^E, and $\hat{P}^E$, we have:
  (1) P and $\hat{P}^E$ are both estimates of P^E under the same size N of the training dataset. However, P is based on fixed datasets s_N and $\bar{s}_M$, and we have $\lim_{N,M\to\infty} P(N;M) = \lim_{N\to\infty} P^E(N)$ almost surely, while $\hat{P}^E$ is based on T training datasets $s_{N,1}, s_{N,2}, ..., s_{N,T}$ and a test dataset $\bar{s}_M$ (T and M are sufficiently large numbers), and we have $\lim_{T,M\to\infty} \hat{P}^E(T;N;M) = P^E(N)$ almost surely.
  (2) Note that P^E is a theoretical metric and hard to apply in practice, while P and $\hat{P}^E$ are numerical metrics and easy to calculate for real-world applications. However, $\hat{P}^E$ is more general than P, because $\hat{P}^E$ considers the average performance across T training datasets and is evaluated on a test dataset of a large size M. Additionally, we have $\lim_{N,M\to\infty} P(N;M) = \lim_{T,N,M\to\infty} \hat{P}^E(T;N;M) = \lim_{N\to\infty} P^E(N)$ almost surely. Therefore, if we have a large number of real-world data, $\hat{P}^E$ is a better and fairer metric than P (however, if there are limited real-world data, we can only use P).

Furthermore, we define the gap between the full-information optimum and the deterministic perfect-foresight counterpart (Definition 8), and accordingly propose two theoretical properties (Propositions 1 and 2) as follows.

Definition 8 (The gap between the full-information optimum and the deterministic perfect-foresight counterpart)**.** *Generally, for a CSO problem, given a test dataset $\bar{s}_M$ ($M \in \mathbb{Z}_{++}$) generated from an assumed joint distribution $\mu_{X,Y}$ of (X, Y) (suppose Y is not measurable with respect to X), let $\hat{\Delta}$ be the gap between the full-information optimum $\hat{R}_M(z^*)$ (based on $\bar{s}_M$, see Formula (3.5)) and the deterministic perfect-foresight counterpart $\hat{R}^*_M$ (see Formula (2.6)), i.e.,*

$$\hat{\Delta} = \hat{R}_M(z^*) - \hat{R}^*_M. \tag{4.1}$$

Remark 8. Here, the $\hat{}$ notation above $\hat{\Delta}$ indicates that $\hat{\Delta}$ is an estimate based on a fixed test dataset $\bar{s}_M$. The primitive value of $\hat{\Delta}$ is denoted by Δ, which measures the gap between the full-information optimum $R(z^*)$ (based on $\mu_{X,Y}$, see Formula (3.1)) and the expectation $\mathbb{E}_Y[\min_{z\in\mathscr{Z}} c(z;Y)]$, i.e.,

$$\Delta = R(z^*) - \mathbb{E}_Y[\min_{z\in\mathscr{Z}} c(z;Y)]. \tag{4.2}$$

However, note that Δ is a theoretical metric based on an assumed joint distribution $\mu_{X,Y}$ of (X, Y), which is challenging to calculate and apply in practice.

The value of Δ also clarifies when the asymptotic upper bound of the original coefficient P is strictly smaller than 1. Let

$$R^*_{\text{fore}} := \mathbb{E}_Y \left[\min_{z \in \mathscr{Z}} c(z; Y) \right].$$

If an excellent prescriptive analytics method asymptotically attains the full-information optimum $R(z^*)$, then the limiting best possible value of the original coefficient is

$$P^{\infty}_{\max} = \frac{R(\hat{z}^{\text{SAA}}) - R(z^*)}{R(\hat{z}^{\text{SAA}}) - R^*_{\text{fore}}} = \frac{R(\hat{z}^{\text{SAA}}) - R(z^*)}{R(\hat{z}^{\text{SAA}}) - R(z^*) + \Delta}.$$

Therefore, as long as $R(\hat{z}^{\text{SAA}}) > R(z^*)$, the asymptotic upper bound of P is strictly smaller than 1 if and only if $\Delta > 0$. The conditional distribution $\mu_{Y|X}$ affects this gap through the residual uncertainty of Y after conditioning on X. Greater conditional variance, heavier tails, or stronger tail asymmetry can increase the expected ex-post regret $c(z^*(X); Y) - \min_{z \in \mathscr{Z}} c(z; Y)$ when the cost function penalizes deviations from the realized perfect-foresight decision. For the newsvendor problem, $z^*(x)$ is a conditional critical-fractile quantile, while the deterministic perfect-foresight decision is $z = Y$; hence the conditional dispersion and tail behavior of $Y|X = x$ directly influence the gap between these two benchmarks.

Proposition 1. *For a CSO problem, given a test dataset $\bar{s}_M$ generated from an assumed joint distribution $\mu_{X,Y}$ of (X, Y) (suppose Y is not measurable with respect to X), as M approaches infinity, we have $\lim_{M\to\infty} \hat{\Delta}(M) = \Delta \geq 0$ almost surely.*

Proof. As M approaches infinity, we have $\lim_{M\to\infty} \hat{R}_M(z^*) = R(z^*) = \mathbb{E}_{(X,Y)\sim\mu_{X,Y}}[c(z^*(X); Y)]$ and $\lim_{M\to\infty} \hat{R}^*_M = \mathbb{E}_{Y\sim\mu_Y}[\min_{z\in\mathscr{Z}} c(z; Y)]$ almost surely; thus, we have $\lim_{M\to\infty} \hat{\Delta}(M) = \Delta$ almost surely. For each test data sample $(\bar{x}_j, \bar{y}_j)$ $(j \in \{1, ..., M\})$ in $\bar{s}_M$, the perfect-foresight oracle makes a decision $z^{\text{fore}}(\bar{y}_j)$ knowing $Y = \bar{y}_j$ without any uncertainty, leading to the cost of $\min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$. However, because Y is not measurable with respect to X, even though $X = \bar{x}_j$ is given, the full-information optimal decision $z^*(\bar{x}_j) \in \arg\min_{z\in\mathscr{Z}} \mathbb{E}_{Y\sim\mu_{Y|X=\bar{x}_j}}[c(z; Y)|X = \bar{x}_j]$ $(j \in \{1, ..., M\})$, which is made under uncertainty depending on the conditional distribution $\mu_{Y|X=\bar{x}_j}$, is just a feasible solution for the deterministic perfect-foresight counterpart problem. Therefore, $c(z^*(\bar{x}_j); \bar{y}_j) \geq \min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$. As M approaches infinity, we have $\mathbb{E}_{(X,Y)\sim\mu_{X,Y}}[c(z^*(X); Y)] \geq \mathbb{E}_{Y\sim\mu_Y}[\min_{z\in\mathscr{Z}} c(z; Y)]$, i.e., $\Delta \geq 0$, accordingly. Therefore, we have $\lim_{M\to\infty} \hat{\Delta}(M) = \Delta \geq 0$ almost surely.

Proposition 2. *For a prescriptive analytics method applied in a CSO problem, let s_N and $\bar{s}_M$ $(N, M \in \mathbb{Z}_{++})$ be a training dataset and a test dataset, respectively, generated from an assumed joint distribution $\mu_{X,Y}$ of (X, Y) (suppose Y is not measurable with respect to X). Suppose that when $N \to \infty$, the efficacy of the prescriptive analytics method is better than that of the SAA method on $\bar{s}_M$, i.e., $\lim_{N\to\infty} \hat{R}_M(\hat{z}^{\text{SAA}}_N) > \lim_{N\to\infty} \hat{R}_M(\hat{z}_N)$. Then, as N and M approach infinity, the following inequalities hold almost surely:*

$$0 < \lim_{N,M\to\infty} P(N; M) \leq \lim_{N,M\to\infty} \hat{P}_{\text{adj}}(N; M) \leq 1. \tag{4.3}$$

Note that if N is fixed and relatively small, we may have $P \leq 0$ and $\hat{P}_{\text{adj}} \leq 0$, and if M is fixed and relatively small, we may have $P = 1$ and $\hat{P}_{\text{adj}} > 1$.

Proof. Recall that in Definitions 1 and 6, we have

$$P = \frac{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(\hat{z}_N)}{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M^*},$$

$$\hat{P}_{\mathrm{adj}} = \frac{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(\hat{z}_N)}{\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) - \hat{R}_M(z^*)}.$$

First, when $N \to \infty$, we have $\lim_{N\to\infty} \hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) > \lim_{N\to\infty} \hat{R}_M(\hat{z}_N)$, which ensures that the numerators of P and $\hat{P}_{\mathrm{adj}}$ are positive. When $M \to \infty$, we generally have $\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) > \hat{R}_M^*$ and $\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) > \hat{R}_M(z^*)$, which ensure that the denominators of P and $\hat{P}_{\mathrm{adj}}$ are positive as well. Therefore, we have $\lim_{N,M\to\infty} P(N; M) > 0$ and $\lim_{N,M\to\infty} \hat{P}_{\mathrm{adj}}(N; M) > 0$ almost surely.

Second, according to Proposition 1, we have $\lim_{M\to\infty} \hat{R}_M(z^*) \geq \lim_{M\to\infty} \hat{R}_M^*$ almost surely. Therefore, we have $\lim_{N,M\to\infty} P(N; M) \leq \lim_{N,M\to\infty} \hat{P}_{\mathrm{adj}}(N; M)$ almost surely.

Third, as mentioned in Section 3.1, for some potential predictive prescriptions $\hat{z}_N(\cdot)$ and under appropriate conditions, as N approaches infinity, we have $R(\hat{z}) = R(z^*)$. Therefore, for a general prescriptive analytics method, as N approaches infinity, we have $R(\hat{z}) \geq R(z^*)$, and thus we have $\lim_{N,M\to\infty} \hat{R}_M(\hat{z}_N) \geq \lim_{M\to\infty} \hat{R}_M(z^*)$ almost surely. Therefore, we have $\lim_{N,M\to\infty} \hat{P}_{\mathrm{adj}}(N; M) \leq 1$ almost surely.

Finally, if N is fixed and relatively small, the efficacy of the prescriptive analytics method may be no better than that of the SAA method on $\bar{s}_M$, i.e., $\hat{R}_M(\hat{z}_N^{\mathrm{SAA}}) \leq \hat{R}_M(\hat{z}_N)$. If the denominators of P and $\hat{P}_{\mathrm{adj}}$ are still positive, then we may have $P \leq 0$ and $\hat{P}_{\mathrm{adj}} \leq 0$. Moreover, if M is fixed and relatively small, (i) a predictive prescription $\hat{z}_N(\bar{x}_j)$ might coincidentally be the same as the deterministic perfect-foresight decision $z^{\mathrm{fore}}(\bar{y}_j)$ ($j \in \{1, ..., M\}$) and, accordingly, we would have $\hat{R}_M(\hat{z}_N) = \hat{R}_M^*$, resulting in $P = 1$; (ii) the average performance of a predictive prescription $\hat{z}_N(\cdot)$ might be even better than the full-information optimum (note that $\hat{z}_N(\cdot)$ and $z^*(\cdot)$ are feasible solutions to the deterministic perfect-foresight counterpart problem), i.e., $\hat{R}_M(z^*) > \hat{R}_M(\hat{z}_N) \geq \hat{R}_M^*$, resulting in $\hat{P}_{\mathrm{adj}} > 1$.

Remark 9. Proposition 2 should be interpreted as an asymptotic almost-sure comparison. For any realized test sample, the inequality $c(z^*(\bar{x}_j); \bar{y}_j) \geq \min_{z\in\mathscr{Z}} c(z; \bar{y}_j)$ holds because $z^*(\bar{x}_j)$ is feasible for the deterministic perfect-foresight problem; hence $\hat{R}_M(z^*) \geq \hat{R}_M^*$ holds for every realized test dataset. However, with limited data, the ordering of the normalized coefficients can still be unstable because the predictive prescription may perform no better than SAA, the denominators may be close to zero, or sampling noise may make $\hat{R}_M(\hat{z}_N)$ appear better than $\hat{R}_M(z^*)$ on a particular finite test set. Therefore, when N or M is small, P and $\hat{P}_{\mathrm{adj}}$ should be interpreted together with the raw out-of-sample costs and the magnitudes of their denominators.

5. Case study: The newsvendor problem

In this section, we describe the classic newsvendor problem in Section 5.1, conduct comprehensive computational experiments using synthetic data in Section 5.2, and numerically demonstrate the deficiencies of P and the strengths of $\hat{P}_{\mathrm{adj}}$.

5.1. Problem description

In the newsvendor problem, we suppose the product to be sold is ice cream. Let $Y \in \mathscr{Y} \subset \mathbb{R}_+$ be tomorrow's random demand for ice cream. Consider two feature variables related to Y: tomorrow's temperature (°C) and humidity (%RH) (note that we can know the values of these two meteorological indicators in advance through weather forecasts with high accuracy), which are denoted by $X_{(1)}$ and $X_{(2)}$, respectively. Generally, the higher the temperature and the lower the humidity (i.e., the drier the weather), the higher the sales of ice cream. Thus, we have the covariate vector $X = (X_{(1)}, X_{(2)}) \in \mathscr{X} \subset \mathbb{R}^2$.

The underlying true joint distribution of (X, Y) is denoted by $F_{X,Y}$, i.e., the cumulative distribution function (cdf). The conditional distribution of Y given X is denoted by $F_{Y|X}$, which is assumed to have a bounded first moment. Therefore, we define the set $\mathcal{F}$ as the collection of all possible joint distributions that satisfy these conditions:

$$\mathcal{F} = \{F \mid X \in \mathscr{X} \subset \mathbb{R}^2,\ Y \in \mathscr{Y} \subset \mathbb{R}_+,\ \mathbb{E}_F[Y|X] < \infty\}, \tag{5.1}$$

and the data of (X, Y) are assumed to be i.i.d. draws from a distribution in $\mathcal{F}$ (i.e., $F_{X,Y} \in \mathcal{F}$), ensuring that the expected cost is well-defined [40].

The decision-maker needs to decide today's order quantity of ice cream for tomorrow after observing the values of X. Let $z \in \mathscr{Z} \subset \mathbb{R}_+$ be the decision variable (i.e., today's order quantity). Let $c_o > 0$ be the per-unit overage cost for each unsold ice cream, and $c_u > 0$ be the per-unit underage cost for each unit of unmet demand. Therefore, the decision-maker needs to solve the following CSO problem with the objective of minimizing the conditional expected cost:

$$v^{\text{newsvd}}(x) = \min_{z \in \mathscr{Z}} \mathbb{E}_{Y \sim F_{X,Y}}[c_o(z - Y)^+ + c_u(Y - z)^+ | X = x], \tag{5.2}$$

$$z^{\text{newsvd}}(x) \in \arg\min_{z \in \mathscr{Z}} \mathbb{E}_{Y \sim F_{X,Y}}[c_o(z - Y)^+ + c_u(Y - z)^+ | X = x], \tag{5.3}$$

where $(\cdot)^+ = \max\{\cdot, 0\}$.

To solve optimization problem (5.3), it is well-established that an ordering quantity z is optimal if $F_{Y|X=x}(z)$ equals $c_u/(c_u + c_o)$. However, in general, there may exist multiple optimal solutions, or no solution z may satisfy that $F_{Y|X=x}(z)$ exactly equals $c_u/(c_u + c_o)$. Nevertheless, a full-information optimal solution can always be defined using the inverse cdf $F^{-1}_{Y|X=x}(\cdot)$:

$$z^*_{F_{Y|X=x}} = F^{-1}_{Y|X=x}\left(\frac{c_u}{c_u + c_o}\right) = \min\left\{z \geq 0 : F_{Y|X=x}(z) \geq \frac{c_u}{c_u + c_o}\right\}, \tag{5.4}$$

which selects the smallest among multiple optimal solutions [41]. Specifically, $z^*_{F_{Y|X=x}}$ corresponds to the $c_u/(c_u + c_o)$ quantile of $F_{Y|X=x}(\cdot)$, commonly referred to as the critical quantile or the newsvendor quantile [42].

5.2. Numerical experiments

For the newsvendor problem mentioned above, we consider two scenarios with different settings of the cost parameters and the assumed joint distributions of (X, Y) (which we refer to as two specific

problems), where we calculate the values of P and $\hat{P}_{\text{adj}}$ for some prescriptive analytics methods to demonstrate the above-discussed deficiencies of P and strengths of $\hat{P}_{\text{adj}}$.

Scenario 1: Suppose that $c_o = 1$, $c_u = 5$, and the joint distribution $F_{X,Y}$ of (X, Y) is defined as follows. The feature variables $X = (X_{(1)}, X_{(2)}) \in \mathscr{X} \subset \mathbb{R}^2$ follow uniform distributions: $X_{(1)} \sim U(20, 38)$ and $X_{(2)} \sim U(20, 80)$. The random demand $Y \in \mathscr{Y} \subset \mathbb{R}_+$ follows a truncated normal distribution, where the mean is $\mu = 10X_{(1)} - 2X_{(2)}$, the standard deviation is $\sigma = 10$, and the truncated interval is $[0, 2\mu]$.

Scenario 2: Suppose that $c_o = 1$, $c_u = 10$, and the joint distribution $F_{X,Y}$ of (X, Y) is defined as follows. The feature variables $X = (X_{(1)}, X_{(2)}) \in \mathscr{X} \subset \mathbb{R}^2$ follow uniform distributions: $X_{(1)} \sim U(30, 40)$ and $X_{(2)} \sim U(20, 50)$. The random demand $Y \in \mathscr{Y} \subset \mathbb{R}_+$ follows a truncated normal distribution, where the mean is $\mu = (X_{(1)})^2 - 10X_{(2)}$, the standard deviation is $\sigma = -X_{(1)} + 3X_{(2)}$, and the truncated interval is $[0, 2\mu]$.

These two scenarios are nondegenerate contextual cases in which the conditional distribution of Y changes with X, so the full-information optimum improves upon the data-poor SAA benchmark and $\hat{P}_{\text{adj}}$ has a meaningful denominator. In the boundary case $Y \perp X$, contextual information has no prescriptive value; asymptotically, the full-information and data-poor stochastic decisions coincide, and thus the denominator of $\hat{P}_{\text{adj}}$ vanishes. We therefore treat this case as a theoretical boundary case rather than as a regular numerical scenario.

Under each scenario, based on the assumed joint distribution, we generate several training datasets $s_N = \{(x_1, y_1), ..., (x_N, y_N)\}$ with different sizes of N, ranging from 10 to 10^4 (where $x_i = (x_{i,(1)}, x_{i,(2)}) \in \mathbb{R}^2$, $y_i \in \mathbb{R}_+$, $i \in \{1, ..., N\}$), and a test dataset $\bar{s}_M = \{(\bar{x}_1, \bar{y}_1), ..., (\bar{x}_M, \bar{y}_M)\}$ containing $M = 10^3$ samples. Using each generated training dataset, we adopt four common ML models: k-nearest-neighbors regression (kNN), kernel regression (KR), classification and regression tree (CART), and random forest (RF). The dataset s_N is partitioned into training and validation sets at a ratio of 4 : 1. Optimal hyperparameters are determined using grid search and 5-fold cross-validation on the validation set, with MSE used as the training metric. The detailed settings for model training and hyperparameter tuning results are described in Appendix B. The numerical tests were performed on a laptop equipped with an Intel(R) Core(TM) Ultra 7255 H processor and 32 GB of RAM. The code was implemented in Python 3.11.0, using scikit-learn for the kNN, KR, CART, and RF models. The four ML models are selected not to benchmark state-of-the-art prediction software, but to provide transparent and commonly used base learners for constructing prescriptions and wSAA weights. They represent local averaging, kernel smoothing, tree partitioning, and ensemble tree partitioning, respectively, which is sufficient for isolating the behavior of the proposed metrics. The proposed coefficients are independent of the specific ML engine and can also evaluate prescriptions generated by more advanced ML or AutoML pipelines. Similarly, grid search is used because it is transparent and reproducible for the controlled experiments. More advanced hyperparameter optimization tools, such as Bayesian optimization or AutoML packages, could be substituted without changing the definitions or theoretical properties of the proposed coefficients.

Then, to solve the newsvendor problem, we implement the point-prediction method (using RF) and four wSAA methods (using kNN, KR, CART, and RF), with the corresponding predictive prescriptions denoted by $\hat{z}_N^{\text{point}}(\cdot)$, $\hat{z}_N^{k\text{NN}}(\cdot)$, $\hat{z}_N^{\text{KR}}(\cdot)$, $\hat{z}_N^{\text{CART}}(\cdot)$, and $\hat{z}_N^{\text{RF}}(\cdot)$, respectively. Here, RF has two distinct roles. In the point-prediction method, RF produces a deterministic forecast $\hat{m}_N(x)$ that is plugged into the downstream optimization problem. In wSAA (RF), RF is instead used to construct

distributional weights through tree-leaf co-membership, so the weighted empirical distribution approximates the conditional distribution of Y given $X = x$. Since the newsvendor decision depends on a conditional quantile rather than only on a conditional mean, the distributional use of RF in wSAA can exploit conditional dispersion and tail information that the point-prediction method discards. Next, based on the training datasets s_N with different sizes N and the test dataset $\bar{s}_M$, we calculate the values of P and $\hat{P}_{\text{adj}}$ for these five prescriptive analytics methods through the following process:

(1) For each test sample $(\bar{x}_j, \bar{y}_j)$ $(j \in \{1, ..., M\})$ in $\bar{s}_M$, calculate $\hat{z}_N^{\text{SAA}}$, $\hat{z}_N^{\text{point}}(\bar{x}_j)$, and $\hat{z}_N^{\text{wSAA}}(\bar{x}_j)$ based on Formulas (2.1), (2.2), and (2.3), respectively, using the training datasets s_N.

(2) Calculate $\hat{R}_M(\hat{z}_N)$, $\hat{R}_M(\hat{z}_N^{\text{SAA}})$, and $\hat{R}_M^*$ based on Formulas (2.4)–(2.6), respectively. Then, we calculate the values of P for the five prescriptive analytics methods based on Formula (2.7).

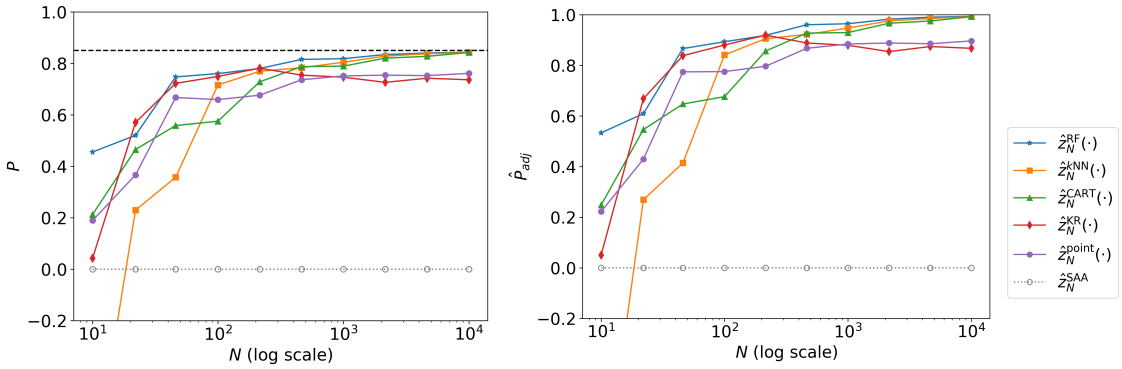
(3) For each test sample $(\bar{x}_j, \bar{y}_j)$ $(j \in \{1, ..., M\})$ in $\bar{s}_M$, calculate $z^*_{F_{Y|X=\bar{x}_j}}$ based on Formula (5.4) (here, note that we obtain the full-information optimal decision analytically based on the assumed distribution).

(4) Calculate $\hat{R}_M(z^*)$ based on Formula (3.5). Then we can calculate the values of $\hat{P}_{\text{adj}}$ for the five prescriptive analytics methods based on Formula (3.7).

The experimental results of Scenario 1 are illustrated in Table 1 and Figure 1, and those of Scenario 2 are shown in Table 2 and Figure 2.

Table 1. The values of P and $\hat{P}_{\text{adj}}$ under Scenario 1.

Metric	Method	N (training sample size)									
		10	22	46	100	215	464	1000	2154	4642	10,000
P	point (RF)	0.189	0.366	0.667	0.659	0.676	0.736	0.751	0.754	0.752	0.761
	wSAA (kNN)	−0.803	0.229	0.357	0.716	0.769	0.783	0.804	0.828	0.837	0.842
	wSAA (KR)	0.042	0.570	0.722	0.749	0.781	0.754	0.746	0.726	0.742	0.736
	wSAA (CART)	0.212	0.465	0.558	0.575	0.728	0.787	0.789	0.820	0.827	0.842
	wSAA (RF)	0.455	0.520	0.747	0.760	0.780	0.815	0.818	0.834	0.840	0.843
$\hat{P}_{\text{adj}}$	point (RF)	0.222	0.429	0.774	0.775	0.796	0.867	0.884	0.888	0.885	0.896
	wSAA (kNN)	−0.942	0.269	0.414	0.841	0.905	0.922	0.947	0.975	0.985	0.992
	wSAA (KR)	0.050	0.668	0.838	0.880	0.919	0.888	0.879	0.854	0.874	0.867
	wSAA (CART)	0.248	0.546	0.647	0.676	0.856	0.927	0.929	0.966	0.975	0.992
	wSAA (RF)	0.533	0.609	0.866	0.893	0.918	0.960	0.964	0.982	0.990	0.993

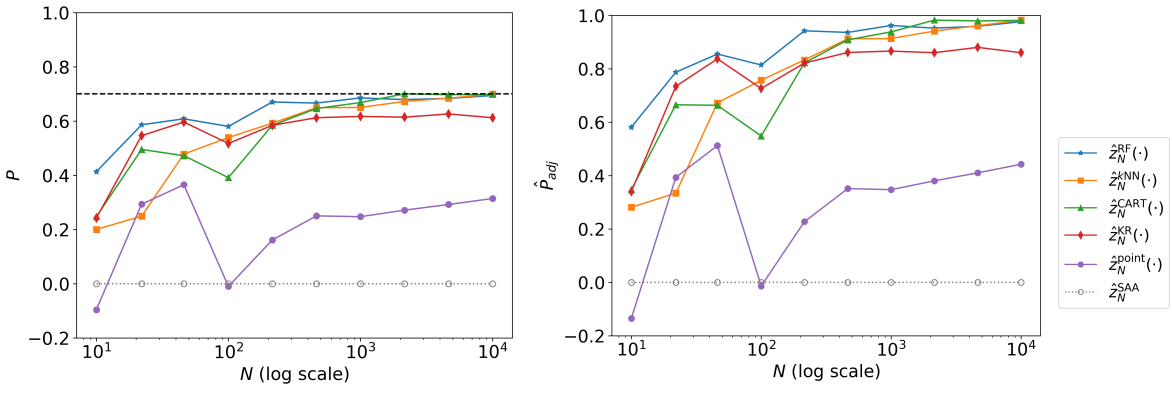


(a) P **(b)** $\hat{P}_{\text{adj}}$

Figure 1. The values of P and $\hat{P}_{\text{adj}}$ under Scenario 1.

Table 2. The values of P and $\hat{P}_{\text{adj}}$ under Scenario 2.

Metric	Method	N (training sample size)									
		10	22	46	100	215	464	1000	2154	4642	10,000
P	point (RF)	−0.096	0.293	0.365	−0.010	0.161	0.250	0.247	0.271	0.292	0.314
	wSAA (kNN)	0.200	0.249	0.478	0.539	0.592	0.649	0.650	0.672	0.684	0.699
	wSAA (KR)	0.242	0.546	0.596	0.517	0.584	0.612	0.617	0.614	0.626	0.612
	wSAA (CART)	0.247	0.495	0.472	0.391	0.584	0.646	0.668	0.700	0.697	0.698
	wSAA (RF)	0.413	0.586	0.608	0.580	0.670	0.666	0.685	0.679	0.683	0.694
$\hat{P}_{\text{adj}}$	point (RF)	−0.135	0.393	0.512	−0.014	0.227	0.351	0.347	0.380	0.410	0.442
	wSAA (kNN)	0.281	0.334	0.672	0.757	0.833	0.912	0.913	0.941	0.961	0.982
	wSAA (KR)	0.340	0.734	0.837	0.726	0.821	0.861	0.866	0.860	0.880	0.860
	wSAA (CART)	0.347	0.665	0.663	0.548	0.822	0.908	0.938	0.982	0.979	0.981
	wSAA (RF)	0.581	0.787	0.855	0.814	0.942	0.936	0.962	0.952	0.959	0.976



(a) P **(b)** $\hat{P}_{\text{adj}}$

Figure 2. The values of P and $\hat{P}_{\text{adj}}$ under Scenario 2.

From Figures 1 and 2, it is evident that as the training sample size N increases, the following occurs. First, the P and $\hat{P}_{\text{adj}}$ values of the point-prediction method all quickly converge to suboptimal values under the two scenarios, because the method does not consider the full spectrum of uncertainty corresponding to each feature X.

Second, the P values of the wSAA methods using kNN, CART, and RF quickly converge to approximately 0.85 and 0.7 under the two scenarios, respectively (although KR has a slower convergence speed). That is, 0.85 and 0.7 are the approximate asymptotic upper bounds of P under the two scenarios, where the settings of the cost parameters and the underlying joint distributions are different, which illustrates that the asymptotic upper bound of P indeed varies across different problems. Here, note that we still cannot know the exact values of the asymptotic upper bound of P under the two scenarios, because (i) we may not be able to find a potentially "excellent" prescriptive analytics method and obtain the corresponding best predictive prescription; and (ii) we cannot generate infinite training data. Although we have a sufficiently large number of training data based on the assumed joint distribution, the number is still fixed and finite.

Third, the $\hat{P}_{\text{adj}}$ values of the wSAA methods using kNN, CART, and RF all quickly converge to approximately 1 under the two scenarios (although KR has a slower convergence speed). Therefore, due to the arbitrary settings of the cost parameters and the joint distributions for the two scenarios, it can be concluded that the approximate asymptotic upper bounds of $\hat{P}_{\text{adj}}$ are all equal to 1 for all possible CSO problems.

In summary, through the above experiments, we numerically demonstrate the deficiencies of P and strengths of $\hat{P}_{\text{adj}}$. Specifically, the asymptotic upper bounds (less than or equal to 1) of P are different and unknown across different CSO problems, whereas the asymptotic upper bounds of $\hat{P}_{\text{adj}}$ are all approximately equal to 1 in all possible CSO problems. Therefore, using $\hat{P}_{\text{adj}}$, we can address Q1 practically. In other words, using $\hat{P}_{\text{adj}}$, we can fairly compare both the efficacy of a specific prescriptive analytics method across multiple CSO problems and the generalizability of multiple prescriptive analytics methods in practice.

6. Conclusions

This paper primarily identifies and addresses two inherent deficiencies of the original coefficient of prescriptiveness P, and accordingly proposes two novel metrics. The problem-dependent deficiency is that P lacks a universal scale, i.e., there always exists an inherent, varying, and unknown asymptotic upper bound (less than or equal to 1) of P in CSO problems. This deficiency results in P only being analyzable and comparable in one particular CSO problem instead of two or more problems. We propose a novel metric, termed the adjusted coefficient of prescriptiveness P_{adj}, that always has the same asymptotic upper bound of 1 theoretically, to address this deficiency, and its estimate $\hat{P}_{\text{adj}}$, which can be obtained numerically, for practical applications. Using $\hat{P}_{\text{adj}}$, we can compare the efficacy of a data-driven method across multiple CSO problems, and further compare the generalizability of multiple prescriptive analytics methods fairly.

Additionally, when applying the original coefficient of prescriptiveness P in a CSO problem, the training and test datasets we use are realizations of random samples in essence, causing some insights derived from the P value to depend on specific datasets, which we denote as the data-dependent deficiency. We propose another novel metric, termed the expected coefficient of prescriptiveness P^E,

to address this deficiency in theory, and its estimate $\hat{P}^E$, which can be calculated numerically, for practical applications. Using $\hat{P}^E$, we can more fairly evaluate the efficacy of candidate prescriptive analytics methods and obtain more general insights for a particular CSO problem.

Moreover, we rigorously analyze and demonstrate some theoretical properties of these coefficients. A case study of the newsvendor problem, along with comprehensive numerical experiments, illustrates and validates the effectiveness of our proposed metrics. The present numerical study focuses on validating the proposed metrics rather than exhaustively comparing all prescriptive algorithms. Since the coefficients are method-agnostic, future work can apply them to prescriptions generated by more advanced ML models, AutoML or Bayesian hyperparameter-optimization pipelines, empirical residuals-based SAA, leave-one-out SAA, and other data-driven optimization methods.

In summary, by establishing these theoretical and numerical coefficients of prescriptiveness, this paper offers new perspectives on contextual optimization.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Shuaian Wang is an editorial board member for Electronic Research Archive and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

1. J. R. Birge, F. Louveaux, *Introduction to Stochastic Programming*, Springer Science & Business Media, New York, 2011. https://doi.org/10.1007/978-1-4614-0237-4
2. A. J. Kleywegt, A. Shapiro, T. Homem-de Mello, The sample average approximation method for stochastic discrete optimization, *SIAM J. Optim.*, **12** (2002), 479–502. https://doi.org/10.1137/S1052623499363220
3. A. Shapiro, D. Dentcheva, A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2021. https://doi.org/10.1137/1.9781611976595
4. D. Bertsimas, N. Kallus, From predictive to prescriptive analytics, *Manage. Sci.*, **66** (2020), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253
5. V. V. Mišić, G. Perakis, Data analytics in operations management: A review, *Manuf. Serv. Oper. Manage.*, **22** (2020), 158–169. https://doi.org/10.1287/msom.2019.0805
6. M. Qi, Z. J. Shen, Integrating prediction/estimation and optimization with applications in operations management, in *Tutorials in Operations Research: Emerging and Impactful Topics in Operations*, Informs, (2022), 36–58. https://doi.org/10.1287/educ.2022.0249
7. U. Sadana, A. Chenreddy, E. Delage, A. Forel, E. Frejinger, T. Vidal, A survey of contextual optimization methods for decision-making under uncertainty, *Eur. J. Oper. Res.*, **320** (2025), 271–289. https://doi.org/10.1016/j.ejor.2024.03.020

8. G. B. Dantzig, Linear programming under uncertainty, *Manage. Sci.*, **1** (1955), 197–206. https://doi.org/10.1287/mnsc.1.3-4.197

9. J. Kotary, F. Fioretto, P. Van Hentenryck, B. Wilder, End-to-end constrained optimization learning: A survey, preprint, arXiv:2103.16378.

10. G. Y. Ban, C. Rudin, The big data newsvendor: Practical insights from machine learning, *Oper. Res.*, **67** (2019), 90–108. https://doi.org/10.1287/opre.2018.1757

11. D. Bertsimas, N. Koduri, Data-driven optimization: A reproducing kernel Hilbert space approach, *Oper. Res.*, **70** (2022), 454–471. https://doi.org/10.1287/opre.2020.2069

12. A. Oroojlooyjadid, L. V. Snyder, M. Takác, Applying deep learning to the newsvendor problem, *IISE Trans.*, **52** (2020), 444–463. https://doi.org/10.1080/24725854.2019.1632502

13. Y. Deng, S. Sen, Predictive stochastic programming, *Comput. Manage. Sci.*, **19** (2022), 65–98. https://doi.org/10.1007/s10287-021-00400-0

14. J. Mandi, J. Kotary, S. Berden, M. Mulamba, V. Bucarey, T. Guns, et al., Decision-focused learning: Foundations, state of the art, benchmark and future opportunities, *J. Artif. Intell. Res.*, **80** (2024), 1623–1701. https://doi.org/10.1613/jair.1.15320

15. R. Kannan, G. Bayraksan, J. R. Luedtke, Residuals-based distributionally robust optimization with covariate information, *Math. Program.*, **207** (2024), 369–425. https://doi.org/10.1007/s10107-023-02014-7

16. D. Bertsimas, B. Van Parys, Bootstrap robust prescriptive analytics, *Math. Program.*, **195** (2022), 39–78. https://doi.org/10.1007/s10107-021-01679-2

17. N. Kallus, X. Mao, Stochastic optimization forests, *Manage. Sci.*, **69** (2023), 1975–1994. https://doi.org/10.1287/mnsc.2022.4458

18. L. Kong, J. Cui, Y. Zhuang, R. Feng, B. A. Prakash, C. Zhang, End-to-end stochastic optimization with energy-based model, preprint, arXiv:2211.13837.

19. M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-López, et al., Efficient and modular implicit differentiation, preprint, arXiv:2105.15183.

20. Q. Dang, Q. Liu, S. Yang, X. He, Data-driven evolutionary algorithm based on inductive graph neural networks for multimodal multiobjective optimization, *IEEE Trans. Evol. Comput.*, **30** (2026), 186–198.

21. X. Tian, R. Yan, Y. Liu, S. Wang, A smart predict-then-optimize method for targeted and cost-effective maritime transportation, *Transp. Res. Part B Methodol.*, **172** (2023), 32–52. https://doi.org/10.1016/j.trb.2023.03.009

22. N. Liu, W. Tang, A. Chen, Y. Lan, A new data-driven robust optimization method for sustainable waste-to-energy supply chain network design problem, *Inf. Sci.*, **699** (2025), 121780. https://doi.org/10.1016/j.ins.2024.121780

23. G. Perakis, M. Sim, Q. Tang, P. Xiong, Robust pricing and production with information partitioning and adaptation, *Manage. Sci.*, **69** (2023), 1398–1419. https://doi.org/10.1287/mnsc.2022.4446

24. Y. G. Kim, B. Do Chung, Data-driven wasserstein distributionally robust dual-sourcing inventory model under uncertain demand, *Omega*, **127** (2024), 103112. https://doi.org/10.1016/j.omega.2024.103112

25. A. Stratigakos, S. Camal, A. Michiorri, G. Kariniotakis, Prescriptive trees for integrated forecasting and optimization applied in trading of renewable energy, *IEEE Trans. Power Syst.*, **37** (2022), 4696–4708. https://doi.org/10.1109/TPWRS.2022.3152667

26. A. Oneto, Á. Lorca, E. Ferrario, A. Poulos, J. C. De La Llera, M. Negrete-Pincetic, Data-driven optimization for seismic-resilient power network planning, *Comput. Oper. Res.*, **166** (2024), 106628. https://doi.org/10.1016/j.cor.2024.106628

27. D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, *Int. J. Forecasting*, **13** (1997), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

28. T. O. Hodson, Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not, *Geosci. Model Dev.*, **15** (2022), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

29. J. P. Barrett, The coefficient of determination—some limitations, *Am. Stat.*, **28** (1974), 19–20. https://doi.org/10.1080/00031305.1974.10479056

30. A. K. Srivastava, V. K. Srivastava, A. Ullah, The coefficient of determination and its adjusted version in linear regression models, *Econom. Rev.*, **14** (1995), 229–240. https://doi.org/10.1080/07474939508800317

31. N. R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, 1998. https://doi.org/10.1002/9781118625590

32. J. Liao, D. McGee, Adjusted coefficients of determination for logistic regression, *Am. Stat.*, **57** (2003), 161–165. https://doi.org/10.1198/0003130031964

33. Y. Bengio, Using a financial training criterion rather than a prediction criterion, *Int. J. Neural Syst.*, **8** (1997), 433–443. https://doi.org/10.1142/S0129065797000422

34. A. N. Elmachtoub, P. Grigas, Smart "predict, then optimize", *Manage. Sci.*, **68** (2022), 9–26. https://doi.org/10.1287/mnsc.2020.3922

35. S. Shah, K. Wang, B. Wilder, A. Perrault, M. Tambe, Decision-focused learning without decision-making: Learning locally optimized decision losses, preprint, arXiv:2203.16067.

36. S. Wang, X. Tian, A deficiency of the weighted sample average approximation (wSAA) framework: Unveiling the gap between data-driven policies and oracles, *Appl. Sci.*, **13** (2023), 8355.

37. S. Bates, T. Hastie, R. Tibshirani, Cross-validation: What does it estimate and how well does it do it, *J. Am. Stat. Assoc.*, **119** (2024), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

38. S. Wager, Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani, *J. Am. Stat. Assoc.*, **115** (2020), 157–160. https://doi.org/10.1080/01621459.2020.1727235

39. Y. Yang, Consistency of cross validation for comparing regression procedures, *Ann. Stat.*, **35** (2007), 2450–2473. https://doi.org/10.1214/009053607000000514

40. O. Besbes, O. Mouchtaki, How big should your data really be? data-driven newsvendor: Learning one sample at a time, *Manage. Sci.*, **69** (2023), 5848–5865. https://doi.org/10.1287/mnsc.2023.4725
41. M. Lin, W. T. Huh, H. Krishnan, J. Uichanco, Data-driven newsvendor problem: Performance of the sample average approximation, *Oper. Res.*, **70** (2022), 1996–2012. https://doi.org/10.1287/opre.2022.2307
42. R. Levi, G. Perakis, J. Uichanco, The data-driven newsvendor problem: New bounds and insights, *Oper. Res.*, **63** (2015), 1294–1306. https://doi.org/10.1287/opre.2015.1422
43. L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Science & Business Media, New York, 2013. https://doi.org/10.1007/978-1-4612-0711-5

Appendix

A. Calculation of the weights in the wSAA method

The calculation of the weights in the weighted sample average approximation (wSAA) method motivated by four common and practically effective machine learning (ML) models is as follows [4]:

(1) Motivated by k-nearest-neighbors regression (kNN), the weights of sample (x_i, y_i) $(i \in \{1, ..., N\})$ in the dataset s_N for a new observation $X = x$ are

$$w_{N,i}^{k\mathrm{NN}}(x) = \frac{1}{k}\mathbb{I}[x_i \text{ is a } k\text{NN of } x], \tag{A.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Ties among equidistant data points are broken either randomly or by a lower-index-first rule.

(2) Motivated by kernel regression (KR), which uses a kernel K to measure distances to a new observation $X = x$, the weights of sample (x_i, y_i) $(i \in \{1, ..., N\})$ in the dataset s_N are

$$w_{N,i}^{\mathrm{KR}}(x) = \frac{K((x_i - x)/h_N)}{\sum_{j=1}^{N} K((x_j - x)/h_N)}, \tag{A.2}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ satisfies $\int K < \infty$ and $h_N > 0$, known as the bandwidth.

(3) Motivated by classification and regression trees (CART), given any map $\mathscr{R} : \mathscr{X} \to \{1, ..., r\}$, the weights of sample (x_i, y_i) $(i \in \{1, ..., N\})$ for a new observation $X = x$ in the dataset s_N are

$$w_{N,i}^{\mathrm{CART}}(x) = \frac{\mathbb{I}[\mathscr{R}(x) = \mathscr{R}(x_i)]}{|\{j : \mathscr{R}(x_j) = \mathscr{R}(x)\}|}, \tag{A.3}$$

where $j \in \{1, ..., N\}$. The map $\mathscr{R}$ corresponds to the disjoint partition $\mathscr{X} = \mathscr{R}^{-1}(1) \sqcup \cdots \sqcup \mathscr{R}^{-1}(r)$.

(4) Motivated by random forest (RF), given T maps $\mathscr{R}^t : \mathscr{X} \to \{1, ..., r^t\}$ $(t = 1, ..., T)$, the weights of sample (x_i, y_i) $(i \in \{1, ..., N\})$ for a new observation $X = x$ in the dataset s_N are

$$w_{N,i}^{\mathrm{RF}}(x) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}[\mathscr{R}^t(x) = \mathscr{R}^t(x_i)]}{|\{j : \mathscr{R}^t(x_j) = \mathscr{R}^t(x)\}|}, \tag{A.4}$$

where $j \in \{1, ..., N\}$.

B. Settings for model training and hyperparameter tuning results

All experiments are performed on a laptop equipped with an Intel(R) Core(TM) Ultra 7255 H processor (2.00 GHz) and 32 GB of RAM. The codebase, implemented in Python 3.11.0, leverages scikit-learn libraries* for the kNN, KR, CART, and RF algorithms. The detailed settings for model training are as follows:

(1) kNN: We use the Euclidean distance metric to measure distances between samples. We normalize each feature to mitigate the influence of different feature value scales on the distance measurement. k is the only hyperparameter to be considered. Referring to Devroye et al. [43], it is advisable to choose $k = \lfloor \sqrt{N_{\text{train}}} \rfloor$ as the starting number, where N_{train} denotes the number of samples in the training set. Based on this starting point, and considering that our training sample size N ranges from 10 to 10^4, we divide the range of N into four intervals and set the search ranges for k, respectively, with the details shown in Table B.1.

(2) KR: We use the Gaussian kernel function to fit the data, and `gamma`, which is used to control the bandwidth, is the only hyperparameter to be considered. We also normalize each feature to mitigate the influence of different feature value scales on the distance measurement. We divide the range of N into three intervals and then set the search ranges for `gamma`, respectively, which are given in Table B.1.

(3) CART: We consider five hyperparameters: `max_depth`, `max_features`, `min_impurity_decrease`, `min_samples_leaf`, and `min_samples_split`. As there are only two features in our case, the number of `max_features` is selected from $\{1, 2\}$. For the remaining four hyperparameters, we divide the range of N into three intervals and then set their respective search ranges, which are listed in Table B.1.

(4) RF: We consider five hyperparameters: `max_depth`, `max_features`, `min_samples_leaf`, `min_samples_split`, and `n_estimators`. The number of `max_features` is also selected from $\{1, 2\}$. Regarding `min_samples_leaf` and `min_samples_split`, we search them from $\{1, 2, 4\}$ and $\{2, 5, 10\}$, respectively. The search ranges for `max_depth` and `n_estimators` are presented in Table B.1.

*Available at https://scikit-learn.org/stable/.

Table B.1. Hyperparameter search ranges for four ML models.

Model	Hyperparameter	Range of N	Search range
kNN	n_neighbors	$N < 50$	{1, 2, ..., 8}
		$50 \leq N < 1000$	{1, 2, ..., 20}
		$1000 \leq N < 10{,}000$	{1, 2, ..., 50}
		$N \geq 10{,}000$	{1, 2, ..., 100}
KR	gamma	$N < 100$	{0.001, 0.01, 0.1}
		$100 \leq N < 1000$	{0.01, 0.1, 1}
		$N \geq 1000$	{0.1, 1, 10}
CART	max_depth	$N < 1000$	{3, 5, 10, None}
		$1000 \leq N < 10{,}000$	{5, 10, 20, None}
		$N \geq 10{,}000$	{10, 20, 30, None}
	max_features	$10 \leq N \leq 10{,}000$	{1, 2}
	min_impurity_decrease	$N < 1000$	{0.0, 0.1}
		$1000 \leq N < 10{,}000$	{0.0, 0.05}
		$N \geq 10{,}000$	{0.0, 0.01}
	min_samples_leaf	$N < 1000$	{3, 5, 10}
		$1000 \leq N < 10{,}000$	{5, 10, 20}
		$N \geq 10{,}000$	{10, 20, 50}
	min_samples_split	$N < 1000$	{5, 10, 20}
		$1000 \leq N < 10{,}000$	{10, 20, 50}
		$N \geq 10{,}000$	{20, 50, 100}
RF	max_depth	$N < 1000$	{3, 5, 10, None}
		$1000 \leq N < 10{,}000$	{10, 20, 30, None}
		$N \geq 10{,}000$	{20, 30, 50, None}
	max_features	$10 \leq N \leq 10{,}000$	{1, 2}
	min_samples_leaf	$10 \leq N \leq 10{,}000$	{1, 2, 4}
	min_samples_split	$10 \leq N \leq 10{,}000$	{2, 5, 10}
	n_estimators	$N < 1000$	{50, 100, 200}
		$N \geq 1000$	{100, 200, 300}

Based on a training dataset s_N, we use grid search and 5-fold cross-validation for hyperparameter tuning with mean squared error (MSE) as the metric during the training of the four ML models (kNN, KR, CART, and RF), and the hyperparameter tuning results under the two scenarios are shown in Tables B.2 and B.3, respectively.

Table B.2. The hyperparameter tuning results of four ML models under Scenario 1.

Model	Hyperparameter	*N* (training sample size)									
		10	22	46	100	215	464	1000	2154	4642	10,000
*k*NN	n_neighbors	1	1	1	3	4	4	8	13	22	40
KR	gamma	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CART	max_depth	3	3	5	10	10	10	10	20	10	20
	max_features	1	2	2	1	2	2	2	2	2	2
	min_impurity_decrease	0.0	0.0	0.0	0.0	0.1	0.0	0.05	0.0	0.0	0.01
	min_samples_leaf	3	3	3	3	3	3	5	5	5	10
	min_samples_split	5	5	5	5	5	5	10	10	20	20
RF	max_depth	None	None	None	None	None	None	None	10	10	10
	max_features	1	2	2	2	2	2	2	2	2	2
	min_samples_leaf	1	1	1	1	1	2	4	4	4	4
	min_samples_split	2	2	2	2	2	2	2	10	10	10
	n_estimators	200	50	50	100	100	200	50	300	300	300

Table B.3. The hyperparameter tuning results of four ML models under Scenario 2.

Model	Hyperparameter	*N* (training sample size)									
		10	22	46	100	215	464	1000	2154	4642	10,000
*k*NN	n_neighbors	2	2	2	5	6	12	11	27	43	88
KR	gamma	0.01	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CART	max_depth	3	3	5	10	5	10	10	10	10	10
	max_features	1	1	1	2	2	2	2	2	2	2
	min_impurity_decrease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.05	0.0
	min_samples_leaf	3	3	3	3	5	5	10	20	20	50
	min_samples_split	5	5	5	10	5	5	10	10	50	20
RF	max_depth	3	None	None	5	5	5	10	10	10	10
	max_features	2	2	2	2	2	2	2	2	1	2
	min_samples_leaf	1	1	1	1	4	4	4	4	4	4
	min_samples_split	2	2	2	2	2	10	10	10	10	2
	n_estimators	50	100	200	200	200	200	200	300	300	300

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)