From Predictive to Prescriptive Analytics

Dimitris Bertsimas
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, dbertsim@mit.edu
Nathan Kallus
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, kallus@mit.edu

In this paper, we combine ideas from machine learning (ML) and operations research and management science (OR/MS) in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems. In a departure from other work on data-driven optimization and reflecting our practical experience with the data available in applications of OR/MS, we consider data consisting, not only of observations of quantities with direct effect on costs/revenues, such as demand or returns, but predominantly of observations of associated auxiliary quantities. The main problem of interest is a conditional stochastic optimization problem, given imperfect observations, where the joint probability distributions that specify the problem are unknown. We demonstrate that our proposed solution methods, which are inspired by ML methods such as local regression (LOESS), classification and regression trees (CART), and random forests (RF), are generally applicable to a wide range of decision problems. We prove that they are computationally tractable and asymptotically optimal under mild conditions even when data is not independent and identically distributed (iid) and even for censored observations. As an analogue to the coefficient of determination R^2, we develop a metric P termed the coefficient of prescriptiveness to measure the prescriptive content of data and the efficacy of a policy from an operations perspective. To demonstrate the power of our approach in a real-world setting we study an inventory management problem faced by the distribution arm of an international media conglomerate, which ships an average of 1 billion units per year. We leverage both internal data and public online data harvested from IMDb, Rotten Tomatoes, and Google to prescribe operational decisions that outperform baseline measures. Specifically, the data we collect, leveraged by our methods, accounts for an 88% improvement as measured by our coefficient of prescriptiveness.

1. Introduction

In today's data-rich world, many problems of operations research and management science (OR/MS) can be characterized by three primitives:

a) Data $\{y^1, \ldots, y^N\}$ on uncertain quantities of interest $Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$ such as simultaneous demands.

b) Auxiliary data $\{x^1, \ldots, x^N\}$ on associated covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ such as recent sale figures, volume of Google searches for a products or company, news coverage, or user reviews, where x^i is concurrently observed with y^i.

c) A decision z constrained in $\mathcal{Z} \subset \mathbb{R}^{d_z}$ made after some observation $X = x$ with the objective of minimizing the *uncertain* costs $c(z; Y)$.

Traditionally, decision-making under uncertainty in OR/MS has largely focused on the problem

$$v^{\text{stoch}} = \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\right], \qquad z^{\text{stoch}} \in \arg\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\right] \tag{1}$$

and its multi-period generalizations and addressed its solution under a priori assumptions about the distribution μ_Y of Y (cf. Birge and Louveaux (2011)), or, at times, in the presence of data $\{y^1, \dots, y^n\}$ in the assumed form of independent and identically distributed (iid) observations drawn from μ_Y (cf. Shapiro (2003), Shapiro and Nemirovski (2005), Kleywegt et al. (2002)). (We will discuss examples of (1) in Section 1.1.) By and large, auxiliary data $\{x^1, \dots, x^N\}$ has not been extensively incorporated into OR/MS modeling, despite its growing influence in practice.

From its foundation, machine learning (ML), on the other hand, has largely focused on supervised learning, or the prediction of a quantity Y (usually univariate) as a function of X, based on data $\{(x^1, y^1), \dots, (x^N, y^N)\}$. By and large, ML does not address optimal decision-making under uncertainty that is appropriate for OR/MS problems.

At the same time, an explosion in the availability and accessibility of data and advances in ML have enabled applications that predict, for example, consumer demand for video games (Y) based on online web-search queries (X) (Choi and Varian (2012)) or box-office ticket sales (Y) based on Twitter chatter (X) (Asur and Huberman (2010)). There are many other applications of ML that proceed in a similar manner: use large-scale auxiliary data to generate predictions of a quantity that is of interest to OR/MS applications (Goel et al. (2010), Da et al. (2011), Gruhl et al. (2005, 2004), Kallus (2014)). However, it is not clear how to go from a good prediction to a good decision. A good decision must take into account uncertainty wherever present. For example, in the absence of auxiliary data, solving (1) based on data $\{y^1, \dots, y^n\}$ but using only the sample mean $\overline{y} = \frac{1}{N}\sum_{i=1}^N y^i \approx \mathbb{E}\left[Y\right]$ and ignoring all other aspects of the data would generally lead to inadequate solutions to (1) and an unacceptable waste of good data.

In this paper, we combine ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems that leverage auxiliary observations. Specifically, the problem of interest is

$$v^*(x) = \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right], \qquad z^*(x) \in \mathcal{Z}^*(x) = \arg\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right], \tag{2}$$

where the underlying distributions are unknown and only data S_N is available, where

$$S_N = \left\{(x^1, y^1), \dots, (x^N, y^N)\right\}.$$

The solution $z^*(x)$ to (2) represents the full-information optimal decision, which, via full knowledge of the unknown joint distribution $\mu_{X,Y}$ of (X, Y), leverages the observation $X = x$ to the fullest

possible extent in minimizing costs. We use the term *predictive prescription* for any function $z(x)$ that prescribes a decision in anticipation of the future given the observation $X = x$. Our task is to use S_N to construct a data-driven predictive prescription $\hat{z}_N(x)$. Our aim is that its performance in practice, $\mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right]$, is close to that of the full-information optimum, $v^*(x)$.

Our key contributions include:

a) We propose various ways for constructing predictive prescriptions $\hat{z}_N(x)$ The focus of the paper is predictive prescriptions that have the form

$$\hat{z}_N(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x)c(z;y^i), \tag{3}$$

where $w_{N,i}(x)$ are weight functions derived from the data. We motivate specific constructions inspired by a great variety of predictive ML methods, including for example and random forests (RF; Breiman (2001)). We briefly summarize a selection of these constructions that we find the most effective below.

b) We also consider a construction motivated by the traditional empirical risk minimization (ERM) approach to ML. This construction has the form

$$\hat{z}_N(\cdot) \in \operatorname*{arg\,min}_{z(\cdot)\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} c(z(x^i);y^i), \tag{4}$$

where $\mathcal{F}$ is some class of functions. We extend the standard ML theory of out-of-sample guarantees for ERM to the case of multivariate-valued decisions encountered in OR/MS problems. We find, however, that in the specific context of OR/MS problems, the construction (4) suffers from some limitations that do not plague the predictive prescriptions derived from (3).

c) We show that that our proposals are computationally tractable under mild conditions.

d) We study the asymptotics of our proposals under sampling assumptions more general than iid. Under appropriate conditions and for certain predictive prescriptions $\hat{z}_N(x)$ we show that costs with respect to the true distributions converge to the full information optimum, i.e.,

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right] = v^*(x),$$

and that the limit points of the decision itself are optimizers of the full information problem (2), i.e.,

$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \mathcal{Z}^*(x),$$

both for almost everywhere x and almost surely. We also extend our results to the case of censored data (such as observing demand via sales).

e) We introduce a new metric P, termed *the coefficient of prescriptiveness*, in order to measure the efficacy of a predictive prescription and to assess the prescriptive content of covariates X, that is, the extent to which observing X is helpful in reducing costs. An analogue to the coefficient of determination R^2 of predictive analytics, P is a unitless quantity that is (eventually) bounded between 0 (not prescriptive) and 1 (highly prescriptive).

f) We demonstrate in a real-world setting the power of our approach. We study an inventory management problem faced by the distribution arm of an international media conglomerate. This entity manages over 0.5 million unique items at some 50,000 retail locations around the world, with which it has vendor-managed inventory (VMI) and scan-based trading (SBT) agreements. On average it ships about 1 billion units a year. We leverage both internal company data and, in the spirit of the aforementioned ML applications, large-scale public data harvested from online sources, including IMDb, Rotten Tomatoes, and Google Trends. These data combined, leveraged by our approach, lead to large improvements in comparison to baseline measures, in particular accounting for an 88% improvement toward the deterministic perfect-foresight counterpart.

Of our proposed constructions of predictive prescriptions $\hat{z}_N(x)$, the ones that we find to be generally the most broadly and practically effective are the following:

a) Motivated by k-nearest-neighbors regression (kNN; Altman (1992)),

$$\hat{z}_N^{k\text{NN}}(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i\in\mathcal{N}_k(x)} c(z;y^i), \tag{5}$$

where $\mathcal{N}_k(x) = \left\{ i=1,\ldots,N : \sum_{j=1}^N \mathbb{I}\left[||x-x_i|| \geq ||x-x_j||\right] \leq k \right\}$ is the neighborhood of the k data points that are closest to x.

b) Motivated by local linear regression (LOESS; Cleveland and Devlin (1988)),

$$\hat{z}_N^{\text{LOESS}}(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i=1}^n k_i(x) \left(1 - \sum_{j=1}^n k_j(x)(x^j-x)^T \Xi(x)^{-1}(x^i-x)\right) c(z;y^i), \tag{6}$$

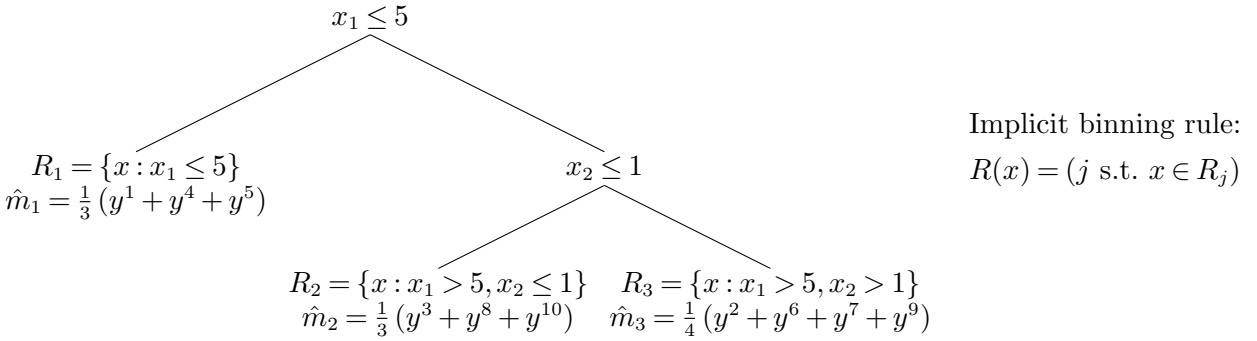
where $\Xi(x) = \sum_{i=1}^n k_i(x)(x^i-x)(x^i-x)^T$, $k_i(x) = \left(1-\left(||x^i-x||/h_N(x)\right)^3\right)^3 \mathbb{I}\left[||x^i-x|| \leq h_N(x)\right]$, and $h_N(x) > 0$ is the distance to the k-nearest point from x. Although this form may seem complicated, it corresponds to the simple idea of approximating $\mathbb{E}\left[c(z;Y)\middle|X=x\right]$ locally by a linear function in x, which we will discuss at greater length in Section 2.

c) Motivated by classification and regression trees (CART; Breiman et al. (1984)),

$$\hat{z}_N^{\text{CART}}(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \sum_{i:R(x^i)=R(x)} c(z;y^i), \tag{7}$$

where $R(x)$ is the binning rule implied by a regression tree trained on the data S_N as shown in an example in Figure 1.

Figure 1 A regression tree is trained on data $\{(x^1, y^1), \ldots, (x^{10}, y^{10})\}$ and partitions the X data into regions defined by the leaves. The Y prediction $\hat{m}(x)$ is $\hat{m}_j$, the average of Y data at the leaf in which $X = x$ ends up. The implicit binning rule is $R(x)$, which maps x to the identity of the leaf in which it ends up.



d) Motivated by random forests (RF; Breiman (2001)),

$$\hat{z}_N^{\mathrm{RF}}(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Z}} \sum_{t=1}^{T} \frac{1}{|\{j : R^t(x^j) = R^t(x)\}|} \sum_{i : R^t(x^i) = R^t(x)} c(z; y^i), \tag{8}$$

where where $R^t(x)$ is the binning rule implied by the t^{th} tree in a random forest trained on the data S_N.

Further detail and other constructions are given in Sections 2 and 6.

In this paper, we focus on the single-period problem (2), in which uncertain quantities are realized at one point in the problem. Such problems include, for example, two-stage decision problems where one set of decisions is made before uncertain quantities are realized and another set of decisions, known as the recourse, after. We study the more general and more intricate multi-period extensions to this problem, where uncertain quantities are realized at several points and in between subsequent decisions, in the multi-period extension to the present paper, Bertsimas and Kallus (2015).

1.1. Two Illustrative Examples

In this section, we illustrate various possible approaches to data-driven decision making and the value of auxiliary data in two examples. We illustrate this with synthetic data but, in Section 5, we study a real-world problem and use real-world data.

We first consider the mean-conditional-value-at-risk portfolio allocation problem. Here, our decision is how we would split our budget among each of d_y securities of which our portfolio may consist. The uncertain quantities of interest are $Y \in \mathbb{R}^{d_y}$, the returns of each of the securities, and the decisions are $z \in \mathbb{R}^{d_y}$, the allocation of the budget to each security. We are interested in maximizing the mean return while minimizing the conditional value at risk at level ϵ (CVaR_ϵ) of losses (negative return), which is the conditional expectation above the $1 - \epsilon$ quantile. Following the reformulation of CVaR_ϵ due to Rockafellar and Uryasev (2000) and using an exchange rate λ

between CVaR_ϵ and mean return, we can write this problem using an extra decision variable $\beta \in \mathbb{R}$, the following cost function for a realization y of returns

$$c((z,\beta);y) = \beta + \frac{1}{\epsilon}\max\left\{-z^T y - \beta, 0\right\} - \lambda z^T y,$$

and the feasible set

$$\mathcal{Z} = \left\{ (z,\beta) \in \mathbb{R}^{d_y \times 1} \,:\, \beta \in \mathbb{R},\, z \geq 0,\, \sum_{i=1}^{d_y} z_i = 1 \right\}.$$

The second example we consider is a two-stage shipment planning problem. Here we have a network of d_z warehouses that we use in order to satisfy the demand for a product at d_y locations. We consider two stages of the problem. In the first stage, some time in advance, we choose amounts $z_i \geq 0$ of units of product to produce and store at each warehouse i, at a cost of $p_1 > 0$ per unit produced. In the second stage, demand $Y \in \mathbb{R}^{d_y}$ realizes at the locations and we must ship units to satisfy it. We can ship from warehouse i to location j at a cost of c_{ij} per unit shipped (recourse variable $s_{ij} \geq 0$) and we have the option of using last-minute production at a cost of $p_2 > p_1$ per unit (recourse variable t_i). The overall problem has the cost function

$$\begin{aligned} c(z;y) = p_1 \sum_{i=1}^{d_z} z_i \quad + \quad \min \quad & \left(p_2 \sum_{i=1}^{d_z} t_i + \sum_{i=1}^{d_z}\sum_{j=1}^{d_y} c_{ij} s_{ij} \right) \\ \text{s.t. } & t_i \geq 0 && \forall i \\ & s_{ij} \geq 0 && \forall i, j \\ & \sum_{i=1}^{d_z} s_{ij} \geq y_j && \forall j \\ & \sum_{j=1}^{d_y} s_{ij} \leq z_i + t_i && \forall i \end{aligned}$$

and the feasible set

$$\mathcal{Z} = \left\{ z \in \mathbb{R}^{d_z} \,:\, z \geq 0 \right\}.$$

The key in each problem is that we do not know Y or its distribution. We consider the situation where we only have data $S_N = ((x^1, y^1), \ldots, (x^N, y^N))$ consisting of observations of Y along with concurrent observations of some auxiliary quantities X that may be associated with the future value of Y. For example, in the portfolio allocation problem, X may include past security returns, behavior of underlying securities, analyst ratings, or volume of Google searches for a company together with keywords like "merger." In the shipment planning problem, X may include, for example, past product sale figures at each of the different retail locations, weather forecasts at the locations, or volume of Google searches for a product to measure consumer attention.

Let us consider two possible extant data-driven approaches to leveraging such data for making a decision. One approach is the sample average approximation of stochastic optimization (SAA, for short). SAA only concerns itself with the marginal distribution of Y, thus ignoring data on X, and solves the following data-driven optimization problem

$$\hat{z}_N^{\text{SAA}} \in \arg\min_{z\in\mathcal{Z}} \frac{1}{N}\sum_{i=1}^{N} c(z;y^i). \tag{9}$$

The objective approximates $\mathbb{E}\left[c(z;Y)\right]$.

Machine learning, on the other hand, leverages the data on X as it tries to predict Y given observations $X=x$. Consider for example a random forest trained on the data S_N. It provides a point prediction $\hat{m}_N(x)$ for the value of Y when $X=x$. Given this prediction, one possibility is to consider the approximation of the random variable Y by our best-guess value $\hat{m}_N(x)$ and solve the corresponding optimization problem,

$$\hat{z}_N^{\text{point-pred}} \in \arg\min_{z\in\mathcal{Z}} c(z;\hat{m}_N(x)). \tag{10}$$

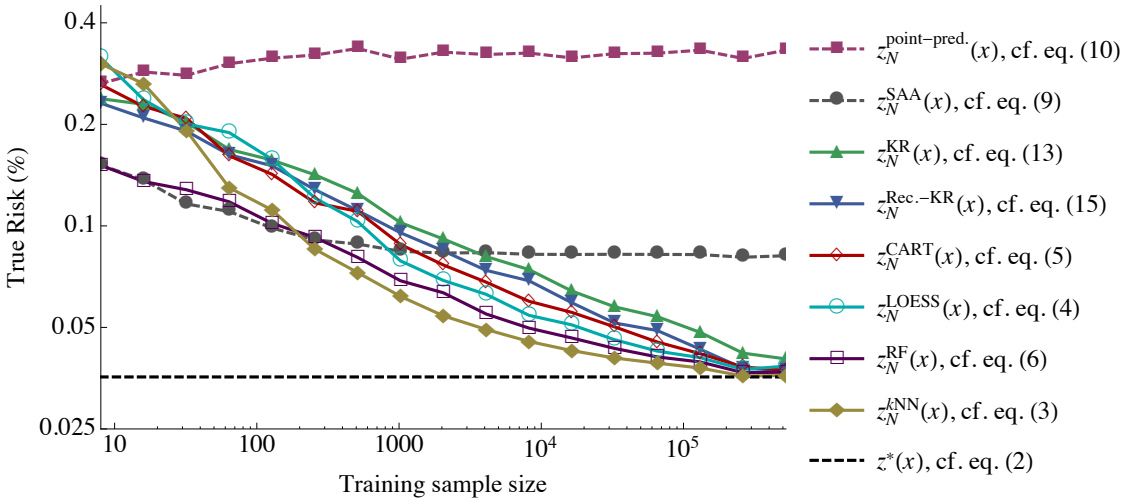
The objective approximates $c\left(z;\mathbb{E}\left[Y\middle|X=x\right]\right)$. We call (10) a point-prediction-driven decision.

If we knew the full joint distribution of Y and X, then the optimal decision having observed $X=x$ is given by (2). Let us compare SAA and the point-prediction-driven decision (using a random forest) to this optimal decision in the two decision problems presented. Let us also consider our proposals (5)-(8) and others that will be introduced in Section 2.

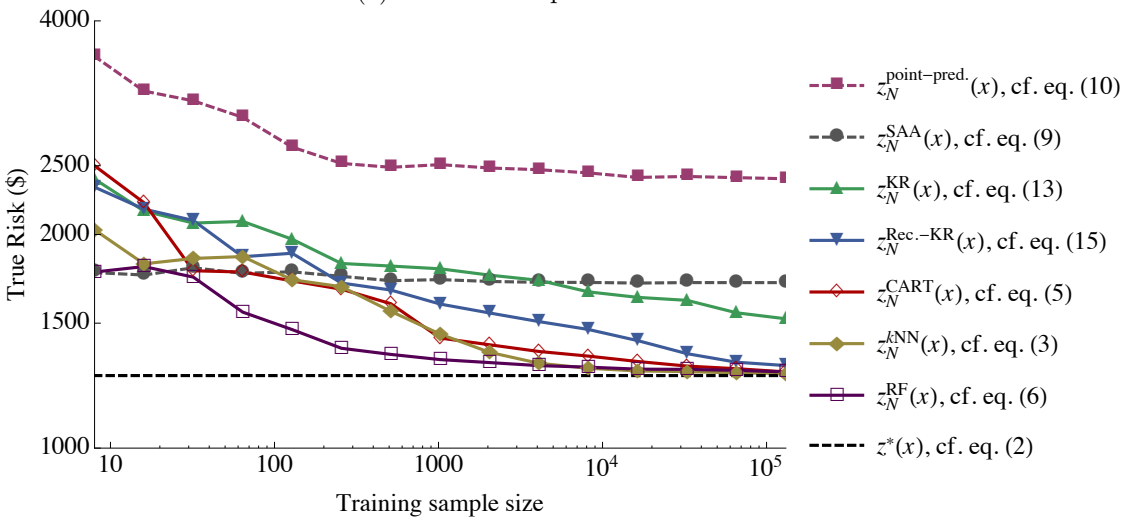
We consider a particular instance of the mean-CVaR portfolio allocation problem with $d_y=12$ securities, where we observe some predictive market factors X before making our investment. We also consider a particular instance of the two-stage shipment planning problem with $d_z=5$ warehouses and $d_y=12$ locations, where we observe some features predictive of demand. In both cases we consider $d_x=3$ and data S_N that, instead of iid, is sampled from a multidimensional evolving process in order to simulate real-world data collection. We give the particular parameters of the problems in the supplementary Section 11. In Figure 2, we report the average performance of the various solutions with respect to the *true* distributions.

The full-information optimum clearly does the best with respect to the true distributions, as expected. The SAA and point-prediction-driven decisions have performances that quickly converge to suboptimal values. The former because it does not use observations on X and the latter because it does not take into account the remaining uncertainty after observing $X=x$.[1] In comparison, we find that our proposals converge upon the full-information optimum given sufficient data. In

[1] Note that the uncertainty of the point prediction in estimating the conditional expectation, gleaned e.g. via the bootstrap, is the wrong uncertainty to take into account, in particular because it shrinks to zero as $N\to\infty$.



(a) Mean-CVaR portfolio allocation.



(b) Two-stage shipment planning.

Figure 2 Performance of various prescriptions with respect to true distributions, by sample size and averaged over samples and new observations x (lower is better). Note the horizontal and vertical log scales.

Section 4.2, we study the general asymptotics of our proposals and prove that the convergence observed here empirically is generally guaranteed under only mild conditions.

Inspecting the figures further, it seems that ignoring X and using only the data on Y, as SAA does, is appropriate when there is very little data; in both examples, SAA outperforms other data-driven approaches for N smaller than $\sim$64. Past that point, our constructions of predictive prescriptions, in particular (5)-(8), leverage the auxiliary data effectively and achieve better, and eventually optimal, performance. The predictive prescription motivated by RF is notable in particular for performing no worse than SAA in the small N regime, and better in the large N regime.

In both examples, the dimension d_x of the observations x was relatively small at $d_x = 3$. In many practical problems, this dimension may well be bigger, potentially inhibiting performance. E.g.,

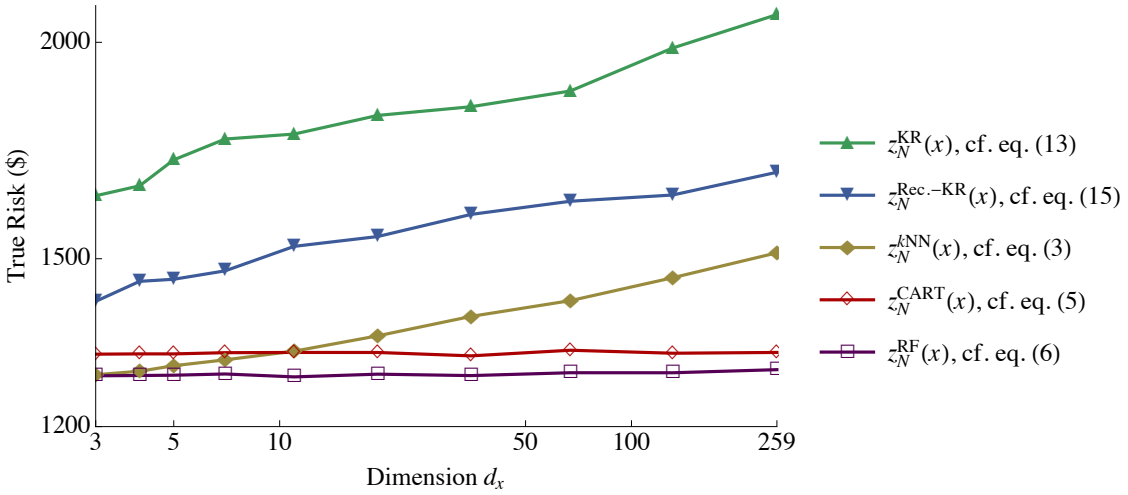


Figure 3 The dependence of performance on the dimension d_x of observations x in the two-stage shipment planning example ($N = 16384$).

in our real-world application in Section 5, we have $d_x = 91$. To study the effect of the dimension of x on the performance of our proposals, we consider polluting x with additional dimensions of uninformative components distributed as independent normals. The results, shown in Figure 3, show that while some of the predictive prescriptions show deteriorating performance with growing dimension d_x, the predictive prescriptions based on CART and RF are largely unaffected, seemingly able to detect the 3-dimensional subset of features that truly matter.

These examples serve as an illustration of the problems and data we tackle, existing approaches, and the gaps filled by our approach. In Section 5, we study an application of our approach to real-world problem and real – not synthetic – data.

1.2. Relevant Literature

Stochastic optimization as in (1) has long been the focus of decision making under uncertainty in OR/MS problems (cf. Birge and Louveaux (2011)) as has its multi-period generalization known commonly as dynamic programming (cf. Bertsekas (1995)). The solution of stochastic optimization problems as in (1) in the presence of data $\{y^1, \ldots, y^N\}$ on the quantity of interest is a topic of active research. The traditional approach is the sample average approximation (SAA) where the true distribution is replaced by the empirical one (cf. Shapiro (2003), Shapiro and Nemirovski (2005), Kleywegt et al. (2002)). Other approaches include stochastic approximation (cf. Robbins and Monro (1951), Nemirovski et al. (2009)), robust SAA (cf. Bertsimas et al. (2014)), and data-driven mean-variance distributionally-robust optimization (cf. Delage and Ye (2010), Calafiore and El Ghaoui (2006)). A notable alternative approach to decision making under uncertainty in OR/MS problems is robust optimization (cf. Ben-Tal et al. (2009), Bertsimas et al. (2011)) and its data-driven variants (cf. Bertsimas et al. (2013), Calafiore and Campi (2005)). There is also a

vast literature on the tradeoff between the collection of data and optimization as informed by data collected so far (cf. Robbins (1952), Lai and Robbins (1985), Besbes and Zeevi (2009)). In all of these methods for data-driven decision making under uncertainty, the focus is on data in the assumed form of iid observations of the parameter of interest Y. On the other hand, ML has attached great importance to the problem of supervised learning wherein the conditional expectation (regression) or mode (classification) of target quantities Y given auxiliary observations $X = x$ is of interest (cf. Trevor et al. (2001), Mohri et al. (2012)).

Statistical decision theory is generally concerned with the optimal selection of statistical estimators (cf. Berger (1985), Lehmann and Casella (1998)). Following the early work of Wald (1949), a loss function such as sum of squared errors or of absolute deviations is specified and the corresponding admissibility, minimax-optimality, or Bayes-optimality are of main interest. Statistical decision theory and ML intersect most profoundly in the realm of regression via empirical risk minimization (ERM), where a regression model is selected on the criterion of minimizing empirical average of loss. A range of ML methods arise from ERM applied to certain function classes and extensive theory on function-class complexity has been developed to analyze these (cf. Bartlett and Mendelson (2003), Vapnik (2000, 1992)). Such ML methods include ordinary linear regression, ridge regression, the LASSO of Tibshirani (1996), quantile regression, and ℓ_1-regularized quantile regression of Belloni and Chernozhukov (2011).

In certain OR/MS decision problems, one can employ ERM to select a decision policy, conceiving of the loss as costs. Indeed, the loss function used in quantile regression is exactly equal to the cost function of the newsvendor problem of inventory management. Rudin and Vahn (2014) consider this loss function and the selection of a univariate-valued linear function with coefficients restricted in ℓ_1-norm in order to solve a newsvendor problem with auxiliary data, resulting in a method similar to Belloni and Chernozhukov (2011). Kao et al. (2009) study finding a convex combination of two ERM solutions, the least-cost decision and the least-squares predictor, which they find to be useful when costs are quadratic. In more general OR/MS problems where decisions are constrained, we show in Section 6 that ERM is not applicable. Even when it is, a linear decision rule may be inappropriate as we show by example. For the limited problems where ERM is applicable, we generalize the standard function-class complexity theory and out-of-sample guarantees to multivariate decision rules since most OR/MS problems involve multivariate decisions.

Instead of ERM, we are motivated more by a strain of non-parametric ML methods based on local learning, where predictions are made based on the mean or mode of past observations that are in some way similar to the one at hand. The most basic such method is kNN (cf. Altman (1992)), which define the prediction as a locally constant function depending on which k data points lie closest. A related method is Nadaraya-Watson kernel regression (KR) (cf. Nadaraya (1964), Watson

(1964)), which is notable for being highly amenable to theoretical analysis but sees less use in practice. KR weighting for solving conditional stochastic optimization problems as in (2) has been considered in Hanasusanto and Kuhn (2013), Hannah et al. (2010) but these have not considered the more general connection to a great variety of ML methods used in practice and neither have they considered asymptotic optimality rigorously. A more widely used local learning regression method than KR is local regression (Cameron and Trivedi (2005) pg. 311) and in particular the LOESS method of Cleveland and Devlin (1988). Even more widely used are recursive partitioning methods, most often in the form of trees and most notably CART of Breiman et al. (1984). Ensembles of trees, most notably RF of Breiman (2001), are known to be very flexible and have competitive performance in a great range of prediction problems. The former averages locally over a partition designed based on the data (the leaves of a tree) and the latter combines many such averages. While there are many tree-based methods and ensemble methods, we focus on CART and RF because of their popularity and effectiveness in practice.

2. From Data to Predictive Prescriptions

Recall that we are interested in the conditional-stochastic optimization problem (2) of minimizing uncertain costs $c(z;Y)$ after observing $X=x$. The key difficulty is that the true joint distribution $\mu_{X,Y}$, which specifies problem (2), is unknown and only data S_N is available. One approach may be to approximate $\mu_{X,Y}$ by the empirical distribution $\hat{\mu}_N$ over the data S_N where each datapoint (x^i, y^i) is assigned mass $1/N$. This, however, will in general fail unless X has small and finite support; otherwise, either $X=x$ has not been observed and the conditional expectation is undefined with respect to $\hat{\mu}_N$ or it has been observed, $X=x=x^i$ for some i, and the conditional distribution is a degenerate distribution with a single atom at y^i without any uncertainty. Therefore, we require some way to generalize the data to reasonably estimate the conditional expected costs for any x. In some ways this is similar to, but more intricate than, the prediction problem where $\mathbb{E}[Y|X=x]$ is estimated from data for any possible $x\in\mathcal{X}$. We are therefore motivated to consider predictive methods and their adaptation to our cause.

In the next subsections we propose a selection of constructions of predictive prescriptions $\hat{z}_N(x)$, each motivated by a local-learning predictive methodology. All the constructions in this section will take the common form of defining some data-driven weights

$$w_{N,i}(x) \quad : \quad \text{the weight associated with } y^i \text{ when observing } X=x, \tag{11}$$

and optimizing the decision $\hat{z}_N$ against a re-weighting of the data, as in (3):

$$\hat{z}_N^{\text{local}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^N w_{N,i}(x)c(z;y^i).$$

In some cases the weights are nonnegative and can be understood to correspond to an estimated conditional distribution of Y given $X=x$. But, in other cases, some of the weights may be negative and this interpretation breaks down.

2.1. *k*NN

Motivated by k-nearest-neighbor regression we propose

$$w_{N,i}^{k\text{NN}}(x) = \begin{cases} 1/k, & \text{if } x^i \text{ is a } k\text{NN of } x, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

giving rise to the predictive prescription (5). Ties among equidistant data points are broken either randomly or by a lower-index-first rule. Finding the kNNs of x without pre-computation can clearly be done in $O(Nd)$ time. Data-structures that speed up the process at query time at the cost of pre-computation have been developed (cf. Bentley (1975)) and there are also approximate schemes that can significantly speed up queries (c.f. Arya et al. (1998)).

A variation of nearest neighbor regression is the radius-weighted k-nearest neighbors where observations in the neighborhood are weighted by a decreasing function f in their distance:

$$w_{N,i}^{\text{radius-}k\text{NN}}(x) = \frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^N \tilde{w}_{N,j}(x)}, \quad \tilde{w}_{N,i}(x) = \begin{cases} f(||x^i - x||), & \text{if } x^i \text{ is a } k\text{NN of } x, \\ 0, & \text{otherwise.} \end{cases}$$

2.2. Kernel Methods

The Nadaraya-Watson kernel regression (KR; cf. Nadaraya (1964), Watson (1964)) estimates $m(x) = \mathbb{E}[Y|X=x]$ by

$$\hat{m}_N(x) = \frac{\sum_{i=1}^N y^i K\left((x^i - x)/h_N\right)}{\sum_{i=1}^N K\left((x^i - x)/h_N\right)},$$

where $K : \mathbb{R}^d \to \mathbb{R}$, known as the kernel, satisfies $\int K < \infty$ (and often unitary invariance) and $h_N > 0$, known as the bandwidth. For nonnegative kernels, KR is the result of the conditional distribution estimate that arises from the Parzen-window density estimates (cf. Parzen (1962)) of $\mu_{X,Y}$ and μ_X (i.e., their ratio). In particular, using the same conditional distribution estimate, the following weights lead to a predictive prescription as in (3):

$$w_{N,i}^{\text{KR}}(x) = \frac{K\left((x^i - x)/h_N\right)}{\sum_{j=1}^N K\left((x^j - x)/h_N\right)}. \tag{13}$$

Some common choices of nonnegative kernels are:

a) Naïve: $K(x) = \mathbb{I}\left[||x|| \le 1\right]$.
b) Epanechnikov: $K(x) = (1 - ||x||^2)\mathbb{I}\left[||x|| \le 1\right]$.
c) Tri-cubic: $K(x) = \left(1 - ||x||^3\right)^3 \mathbb{I}\left[||x|| \le 1\right]$.
d) Gaussian: $K(x) = \exp\left(-||x||^2/2\right)$.

(14)

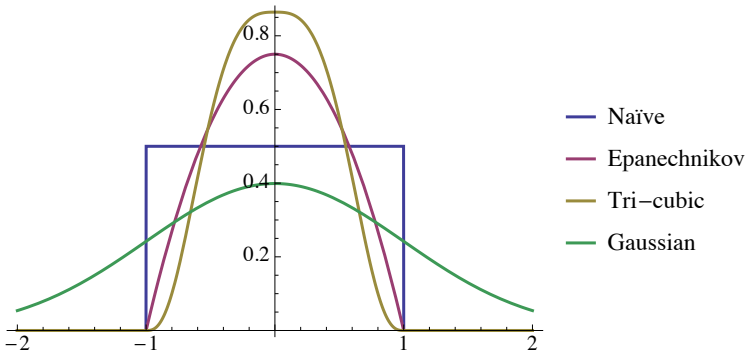


Figure 4 Comparison of the different kernels (each normalized to integrate to 1).

Note that the naïve kernel with bandwidth h_N corresponds directly to uniformly weighting all neighbors of x that are within a radius h_N. A comparison of different kernels is shown in Figure 4.

It is these weights (13) that are used in Hanasusanto and Kuhn (2013), Hannah et al. (2010) (without formally considering asymptotic optimality of $\hat{z}_N(x)$). A problem with KR is that it can be very biased in high dimensions, especially at the boundaries of the data (estimates will tend toward the outliers at the boundaries). While KR is particularly amenable to theoretical analysis due to its simplicity, it is not widely used in practice. We will next consider local regression, which is a related, more widely used approach. Before we proceed we first introduce a recursive modification to (13) that is motivated by an alternative kernel regressor introduced by Devroye and Wagner (1980):

$$w_{N,i}^{\text{recursive-KR}}(x) = \frac{K\left((x^i - x)/h_i\right)}{\sum_{j=1}^N K\left((x^j - x)/h_j\right)}, \tag{15}$$

where now the bandwidths h_i are selected per-data-point and independent of N. The benefits of (15) include the simplicity of an update when accumulating additional data since all previous weights remain unchanged as well as smaller variance under certain conditions (cf. Roussas (1992)). Moreover, from a theoretical point of view, much weaker conditions are necessary to ensure good asymptotic behavior of (15) compared to (13), as we will see in the next section.

2.3. Local Linear Methods

An alternative interpretation of KR predictions is they solves the locally-weighted least squares problem for a constant predictor:

$$\hat{m}_N(x) = \arg\min_{\beta_0} \sum_{i=1}^N k_i(x)\left(y^i - \beta_0\right)^2.$$

One can instead consider a predictive method that solves a similar locally-weighted least squares problem for a linear predictor:

$$\hat{m}_N(x) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^N k_i(x)\left(y^i - \beta_0 - \beta_1^T(x^i - x)\right)^2.$$

In prediction, local linear methods are known to be preferable over KR (cf. Fan (1993)). Combined with a particular choice of $k_i(x)$, this results in the linear version of the LOESS variant of local regression developed in Cleveland and Devlin (1988). If, instead, we use this idea to locally approximate the conditional costs $\mathbb{E}\left[c(z;Y)\middle|X=x\right]$ by a linear function we will arrive at a functional estimate and a predictive prescription as in (3) with the weights

$$w_{N,i}^{\text{LOESS}}(x)=\frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^N \tilde{w}_{N,j}(x)},\quad \tilde{w}_{N,i}(x)=k_i(x)\left(1-\sum_{j=1}^n k_j(x)(x^j-x)^T\Xi(x)^{-1}(x^i-x)\right), \tag{16}$$

where $\Xi(x)=\sum_{i=1}^n k_i(x)(x^i-x)(x^i-x)^T$ and $k_i(x)=K\left((x^i-x)/h_N(x)\right)$. In the LOESS method, K is the tri-cubic kernel and $h_N(x)$ is not fixed, as it is in (13), but chosen to vary with x so that at each query point x the same number of data points taken into consideration; in particular, $h_N(x)$ is chosen to be the distance to x's k-nearest neighbor where k, in turn, is fixed. These choices lead to the form of $\hat{z}_N^{\text{LOESS}}(x)$ presented in Section 1.

2.4. Trees

Tree-based methods recursively split the sample S_N into regions in $\mathcal{X}$ so to gain reduction in "impurity" of the response variable Y within each region. The most well known tree-based predictive method is CART developed in Breiman et al. (1984). There are different definitions of "impurity," such as Gini or entropy for classification and variance reduction for regression, and different heuristics to choose the best split, different combinations resulting in different algorithms. Multivariate impurity measures are usually the component-wise average of univariate impurities. Splits are usually restricted to axis-aligned half-spaces. Such splits combined with the Gini impurity for classification and variance reduction for regression results in the original CART algorithm of Breiman et al. (1984). Once a tree is constructed, the value of $\mathbb{E}[Y|X=x]$ (or, the most likely class) is then estimated by the average (or, the mode) of y^i's associated with the x^i's that reside in the same region as x. The recursive splitting is most often represented as a tree with each non-leaf node representing an intermediate region in the algorithm (see Figure 1). With axis-aligned splits, the tree can be represented as subsequent inquiries about whether a particular component of the vector x is larger or smaller than a value. For a thorough review of tree-based methods and their computation see §9.2 of Trevor et al. (2001).

Regardless of the particular method chosen, the final partition can generally be represented as a binning rule identifying points in $\mathcal{X}$ with the disjoint regions, $\mathcal{R}:\mathcal{X}\to\{1,\dots,r\}$. The partition is then the disjoint union $\mathcal{R}^{-1}(1)\sqcup\cdots\sqcup\mathcal{R}^{-1}(r)=\mathcal{X}$. The tree regression and classification estimates correspond directly to taking averages or modes over the uniform distribution of the data points residing in the region $R(x)$.

For our prescription problem, we propose to use the binning rule to construct weights as follows for a predictive prescription of the form (3):

$$w_{N,i}^{\mathrm{CART}}(x) = \frac{\mathbb{I}\left[\mathcal{R}(x) = \mathcal{R}(x^i)\right]}{\left|\left\{j : R(x^j) = R(x)\right\}\right|}. \tag{17}$$

Notice that the weights (17) are piecewise constant over the partitions and therefore the recommended optimal decision $\hat{z}_N(x)$ is also piecewise constant. Therefore, solving r optimization problems after the recursive partitioning process, the resulting predictive prescription can be fully compiled into a decision tree, with the decisions that are truly decisions. This also retains CART's lauded interpretability. Apart from being interpretable, tree-based methods are also known to be useful in learning complex interactions and to perform well with large datasets.[2]

2.5. Ensembles

A random forest, developed in Breiman (2001), is an ensemble of trees each trained on a random subset of components of X and a random subsample of the data. This makes them more uncorrelated and therefore their average have lower variance. Random forests are one of the most flexible tools of ML and is extensively used in predictive applications. For a thorough review of random forests and their computation see §15 of Trevor et al. (2001).

After training such a random forest of trees, we can extract the partition rules $\mathcal{R}_t$ $t = 1, \ldots, T$, one for each tree in the forest. We propose to use these to construct the following weights as follows for a predictive prescription of the form (3):

$$w_{N,i}^{\mathrm{RF}}(x) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\left[\mathcal{R}^t(x) = \mathcal{R}^t(x^i)\right]}{\left|\left\{j : R^t(x^j) = R^t(x)\right\}\right|}. \tag{18}$$

There are also other tree-ensembles methods. RF essentially combines the ideas from bagged (bootstrap-aggregated) forests (cf. Breiman (1996)) and random-subspace forests (cf. Ho (1998)). Other forest ensembles include extremely randomized trees developed in Geurts et al. (2006). The weights extracted from alternative tree ensembles would have the same form as (18).

In practice, RF is known as a flexible prediction algorithm that can perform competitively in almost any problem instance (cf. Breiman et al. (2001)). For our prescription problem, in Section 1.1 we saw that our predictive prescription based on RF, given in eq. (8), performed well overall in two different problems, for a range of sample sizes, and for a range of dimensions d_x. Based on this evidence of flexible performance, we choose our predictive prescription based on RF for our real-world application, which we study in Section 5.

[2] A more direct application of tree methods to the prescription problem would have us consider the impurities being minimized in each split to be equal to the mean cost $c(z; y)$ of taking the best constant decision z in each side of the split. However, since we must consider splitting on each variable and at each data point to find the best split (cf. pg. 307 of Trevor et al. (2001)), this can be overly computationally burdensome for all but the simplest problems that admit a closed form solution such as least sum of squares or the newsvendor problem.

3. Metrics of Prescriptiveness

In this section, we develop a relative, unitless measure of the efficacy of a predictive prescription. An absolute measure of efficacy is marginal expected costs,

$$R(\hat{z}_N) = \mathbb{E}\left[\mathbb{E}\left[c\left(\hat{z}_N(X);Y\right)\middle|X\right]\right] = \mathbb{E}\left[c\left(\hat{z}_N(X);Y\right)\right].$$

Given a validation data set $\tilde{S}_{N_v} = ((\tilde{x}^1,\tilde{y}^1),\cdots,(\tilde{x}^{N_v},\tilde{y}^{N_v}))$, we can estimate $R(\hat{z}_N)$ as the sample average:

$$\hat{R}_{N_v}(\hat{z}_N) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\tilde{x}^i);\tilde{y}^i\right).$$

If $\tilde{S}_{N_v} = S_N$ then this an in-sample estimate, which biases in favor of overfitting, and if $\tilde{S}_{N_v}$ is disjoint, then this is an out-of-sample estimate that provides an unbiased estimate of $R(\hat{z}_N)$.

While an absolute measure allows one to compare two predictive prescriptions for the same problem and data, a relative measure can quantify the overall prescriptive content of the data and the efficacy of a prescription on a universal scale. For example, in predictive analytics, the coefficient of determination R^2 – rather than the absolute root-mean-squared error – is a unitless quantity used to quantify the overall quality of a prediction and the predictive content of data X. R^2 measures the fraction of variance of Y reduced, or "explained," by the prediction based on X. Another way of interpreting R^2 is as the fraction of the way that X and a particular predictive model take us from a data-poor prediction (the sample average) to a perfect-foresight prediction that knows Y in advance.

We define an analogous quantity for the predictive prescription problem, which we term *the coefficient of prescriptiveness*. It involves three quantities. First,

$$\hat{R}_{N_v}\left(\hat{z}_N(x)\right) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\tilde{x}^i);\tilde{y}^i\right)$$

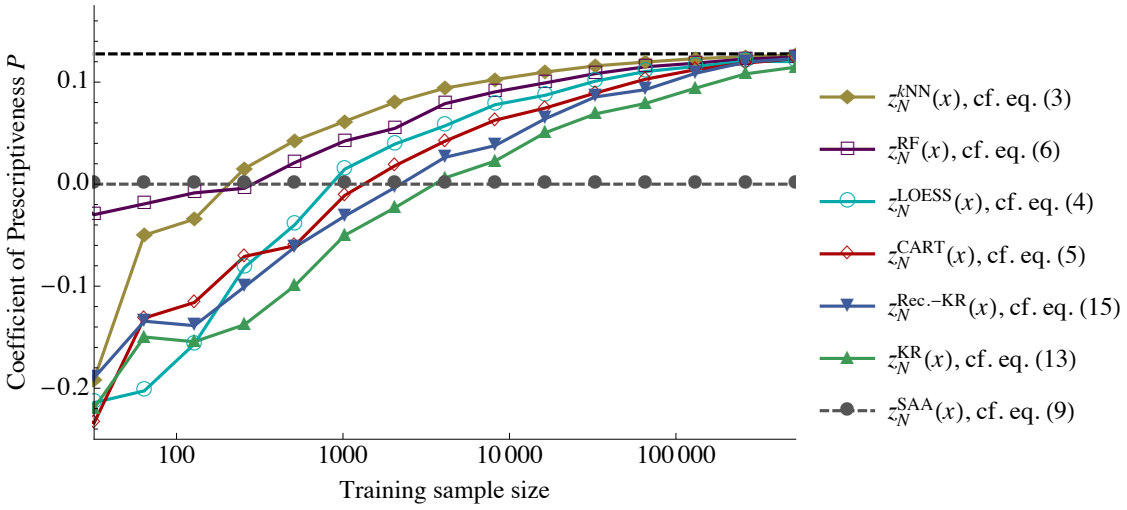
is the estimated expected costs due to our predictive prescription. Second,

$$\hat{R}^*_{N_v} = \frac{1}{N_v}\sum_{i=1}^{N_v} \min_{z\in\mathcal{Z}} c\left(z;\tilde{y}^i\right)$$

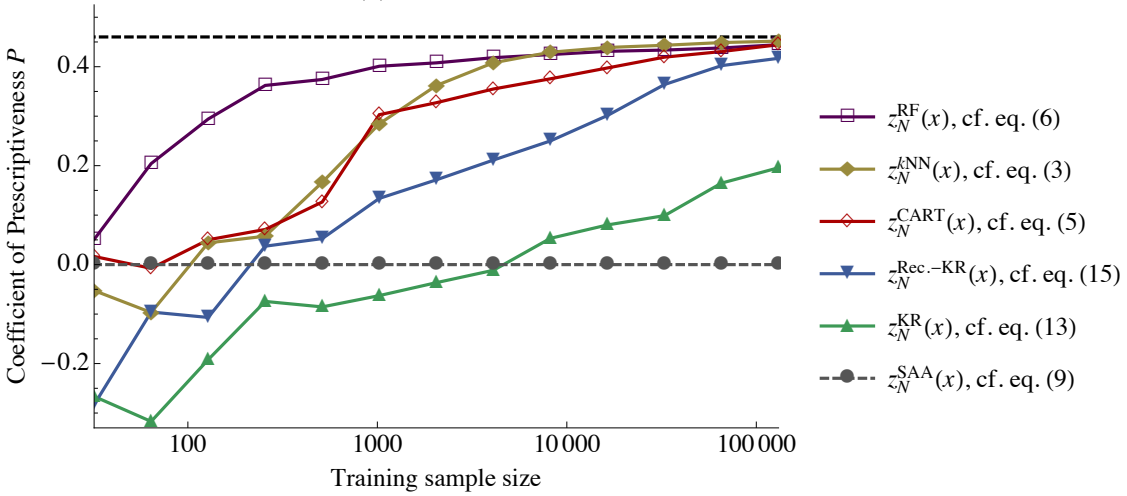
is the estimated expected costs in the deterministic perfect-foresight counterpart problem, in which one has foreknowledge of Y without any uncertainty (note the difference to the full-information optimum, which does have uncertainty). Third,

$$\hat{R}_{N_v}\left(\hat{z}^{\mathrm{SAA}}_N\right) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}^{\mathrm{SAA}}_N;\tilde{y}^i\right) \quad \text{where} \quad \hat{z}^{\mathrm{SAA}}_N \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \frac{1}{N}\sum_{i=1}^{N} c\left(z;y^i\right)$$

is the estimated expected costs of a data-driven prescription that is data poor, based only on Y data. This is the SAA solution to the prescription problem, which serves as the analogue to the



(a) Mean-CVaR portfolio allocation.



(b) Two-stage shipment planning.

Figure 5 The coefficient of prescriptiveness P in each of the two examples from Section 1.1, measured out of sample. The black horizontal line denotes the theoretical limit

sample average as a data-poor solution to the prediction problem. Using these three quantities, we define the coefficient of prescriptiveness P as follows:

$$P = 1 - \frac{\hat{R}_{N_v}\left(\hat{z}_N(x)\right) - \hat{R}^*_{N_v}}{\hat{R}_{N_v}\left(\hat{z}_N^{\mathrm{SAA}}\right) - \hat{R}^*_{N_v}} \tag{19}$$

When $\tilde{S}_{N_v} = S_N$ (in-sample) we can write

$$P = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} c\left(\hat{z}_N(x^i); y^i\right) - \frac{1}{N}\sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c\left(z; y^i\right)}{\min_{z\in\mathcal{Z}} \frac{1}{N}\sum_{i=1}^{N} c\left(z; y^i\right) - \frac{1}{N}\sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c\left(z; y^i\right)}.$$

The coefficient of prescriptiveness P is a unitless quantity bounded above by 1. A low P denotes that X provides little useful information for the purpose of prescribing an optimal decision in the particular problem at hand or that $\hat{z}_N(x)$ is ineffective in leveraging the information in X. A high P denotes that taking X into consideration has a significant impact on reducing costs and that $\hat{z}_N$ is effective in leveraging X for this purpose.

Let us consider the coefficient of prescriptiveness in the two examples from Section 1.1. For each of our predictive prescriptions and for each N, we measure the out of sample P on a validation set of size $N_v = 200$ and plot the results in Figure 5. Notice that even when we converge to the full-information optimum, P does not approach 1 as N grows. Instead we see that for the same methods that converged to the full-information optimum in Figure 2, we have a P that approaches 0.13 in the portfolio allocation example and 0.46 in the shipment planning example. This number represents the extent of the potential that X has to reduce costs in this particular problem. It is the fraction of the way that knowledge of X, leveraged correctly, takes us from making a decision under full uncertainty about the value of Y to making a decision in a completely deterministic setting. As is the case with R^2, what magnitude of P denotes a successful application depends on the context. In our real-world application in Section 5, we find an out-of-sample P of 0.88.

4. Properties of Local Predictive Prescriptions

In this section, we study two important properties of local predictive prescriptions: computational tractability and asymptotic optimality. All proofs are given in the E-companion.

4.1. Tractability

In Section 2, we considered a variety of predictive prescriptions $\hat{z}_N(x)$ that are computed by solving the optimization problem (3). An important question is then when is this optimization problem computationally tractable to solve. As an optimization problem, problem (3) differs from the problem solved by the standard SAA approach (9) only in the weights given to different observations. Therefore, it is similar in its computational complexity and we can defer to computational studies of SAA such as Shapiro and Nemirovski (2005) to study the complexity of solving problem (3). For completeness, we develop sufficient conditions for problem (3) to be solvable in polynomial time.

THEOREM 1. *Fix x and weights $w_{N,i}(x) \geq 0$. Suppose $\mathcal{Z}$ is a closed convex set and let a separation oracle for it be given. Suppose also that $c(z;y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z. Then for any x we can find an ϵ-optimal solution to (3) in time and oracle calls polynomial in $N_0, d, \log(1/\epsilon)$ where $N_0 = \sum_{i=1}^{N} \mathbb{I}[w_{N,i}(x) > 0] \leq N$ is the effective sample size.*

Note that all of the weights presented in Section 2 have been necessarily all nonnegative with the exception of local regression (16). As the spans $h_N(x)$ shrink and the number of data points increases, these weights will always become nonnegative. However, for a fixed problem the weights $w_{N,i}^{\text{LOESS}}(x)$ may be negative for some i and x, in which case the optimization problem (3) may not be polynomially solvable. In particular, in the case of the portfolio example presented in Section 1.1, if some weights are negative, we formulate the corresponding optimization problem as a mixed integer-linear optimization problem.

4.2. Asymptotic Optimality

In Section 1.1, we saw that our predictive prescriptions $\hat{z}_N(x)$ converged to the full-information optimum as the sample size N grew. Next, we show that this anecdotal evidence is supported by mathematics and that such convergence is guarantees under only mild conditions. We define *asymptotic optimality* as the desirable asymptotic behavior for $\hat{z}_N(x)$.

Definition 1. We say that $\hat{z}_N(x)$ is *asymptotically optimal* if, with probability 1, we have that for μ_X-almost-everywhere $x \in \mathcal{X}$, as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right] = v^*(x),$$
$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \mathcal{Z}^*(x),$$

where $L(A)$ denotes the limit points of A.

Asymptotic optimality depends on our choice of $\hat{z}_N(x)$, the structure of the decision problem (cost function and feasible set), and on how we accumulate our data S_N. The traditional assumption on data collection is that it constitutes an iid process. This is a strong assumption and is often only a modeling approximation. The velocity and variety of modern data collection often means that historical observations do not generally constitute an iid sample in any real-world application. We are therefore motivated to consider an alternative model for data collection, that of mixing processes. These encompass such processes as ARMA, GARCH, and Markov chains, which can correspond to sampling from evolving systems like prices in a market, daily product demands, or the volume of Google searches on a topic. While many of our results extend to such settings, we present only the iid case in the main text to avoid cumbersome exposition and defer these extensions to the supplemental Section 8.2. For the rest of the section let us assume that S_N is generated by iid sampling.

As mentioned, asymptotic optimality also depends on the structure of the decision problem. Therefore, we will also require the following conditions.

Assumption 1 **(Existence)**. *The full-information problem* (2) *is well defined:* $\mathbb{E}\left[|c(z;Y)|\right] < \infty$ *for every* $z \in \mathcal{Z}$ *and* $\mathcal{Z}^*(x) \neq \varnothing$ *for almost every* x.

ASSUMPTION 2 **(Continuity)**. *$c(z;y)$ is equicontinuous in z: for any $z \in \mathcal{Z}$ and $\epsilon > 0$ there exists $\delta > 0$ such that $|c(z;y) - c(z';y)| \le \epsilon$ for all z' with $||z - z'|| \le \delta$ and $y \in \mathcal{Y}$.*

ASSUMPTION 3 **(Regularity)**. *$\mathcal{Z}$ is closed and nonempty and in addition either*

1. *$\mathcal{Z}$ is bounded,*
2. *$\mathcal{Z}$ is convex and $c(z;y)$ is convex in z for every $y \in \mathcal{Y}$, or*
3. *$\liminf_{||z||\to\infty} \inf_{y\in\mathcal{Y}} c(z;y) > -\infty$ and for every $x \in \mathcal{X}$, there exists $D_x \subset \mathcal{Y}$ such that $\lim_{||z||\to\infty} c(z;y) \to \infty$ uniformly over $y \in D_x$ and $\mathbb{P}\left(y \in D_x \middle| X = x\right) > 0$.*

Under these conditions, we have the following sufficient conditions for asymptotic optimality.

THEOREM 2 **(kNN)**. *Suppose Assumptions 1, 2, and 3 hold. Let $w_{N,i}(x)$ be as in (12) with $k = \min\left\{\lceil CN^\delta \rceil, N-1\right\}$ for some $C > 0$, $0 < \delta < 1$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal.*

THEOREM 3 **(Kernel Methods)**. *Suppose Assumptions 1, 2, and 3 hold and that $\mathbb{E}\left[|c(z;Y)| \max\left\{\log|c(z;Y)|, 0\right\}\right] < \infty$ for each z. Let $w_{N,i}(x)$ be as in (13) with K being any of the kernels in (14) and $h_N = CN^{-\delta}$ for $C > 0$, $0 < \delta < 1/d_x$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal.*

THEOREM 4 **(Recursive Kernel Methods)**. *Suppose Assumptions 1, 2, and 3 hold. Let $w_{N,i}(x)$ be as in (15) with K being the naïve kernel and $h_i = Ci^{-\delta}$ for some $C > 0$, $0 < \delta < 1/(2d_x)$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal.*

THEOREM 5 **(Local Linear Methods)**. *Suppose Assumptions 1, 2, and 3 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X, and that costs are bounded over y for each z (i.e., $|c(z;y)| \le g(z)$). Let $w_{N,i}(x)$ be as in (16) with K being any of the kernels in (14) and with $h_N = CN^{-\delta}$ for some $C > 0$, $0 < \delta < 1/d_x$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal.*

Although we do not have firm theoretical results on the asymptotic optimality of the predictive prescriptions based on CART (eq. (7)) and RF (eq. (8)), we have observed them to converge empirically in Section 1.1.

5. A Real-World Application

In this section, we apply our approach to a real-world problem faced by the distribution arm of an international media conglomerate (the vendor) and demonstrate that our approach, combined with extensive data collection, leads to significant advantages. The vendor has asked us to keep its identity confidential as well as data on sale figures and specific retail locations. Some figures are therefore shown on relative scales.

5.1. Problem Statement

The vendor sells over 0.5 million entertainment media titles on CD, DVD, and BluRay at over 50,000 retailers across the US and Europe. On average they ship 1 billion units in a year. The retailers range from electronic home goods stores to supermarkets, gas stations, and convenience stores. These have vendor-managed inventory (VMI) and scan-based trading (SBT) agreements with the vendor. VMI means that the inventory is managed by the vendor, including replenishment (which they perform weekly) and planogramming. SBT means that the vendor owns all inventory until sold to the consumer. Only at the point of sale does the retailer buy the unit from the vendor and sell it to the consumer. This means that retailers have no cost of capital in holding the vendor's inventory.

The cost of a unit of entertainment media consists mainly of the cost of production of the content. Media-manufacturing and delivery costs are secondary in effect. Therefore, the primary objective of the vendor is simply to sell as many units as possible and the main limiting factor is inventory capacity at the retail locations. For example, at many of these locations, shelf space for the vendor's entertainment media is limited to an aisle endcap display and no back-of-the-store storage is available. Thus, the main loss incurred in over-stocking a particular product lies in the loss of potential sales of another product that sold out but could have sold more. In studying this problem, we will restrict our attention to the replenishment and sale of video media only and to retailers in Europe.

Apart from the limited shelf space the other main reason for the difficulty of the problem is the particularly high uncertainty inherent in the initial demand for new releases. Whereas items that have been sold for at least one period have a somewhat predictable decay in demand, determining where demand for a new release will start is a much less trivial task. At the same time, new releases present the greatest opportunity for high demand and many sales.

We now formulate the full-information problem. Let $r=1,\ldots,R$ index the locations, $t=1,\ldots,T$ index the replenishment periods, and $j=1,\ldots,d$ index the products. Denote by z_j the order quantity decision for product j, by Y_j the uncertain demand for product j, and by K_r the overall inventory capacity at location r. Considering only the *main* effects on revenues and costs as discussed in the previous paragraph, the problem decomposes on a per-replenishment-period, per-location basis. We therefore wish to solve, for each t and r, the following problem:

$$\begin{aligned} v^*(x_{tr}) = \max \quad & \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j,\, z_j\} \,\middle|\, X = x_{tr}\right] = \sum_{j=1}^{d} \mathbb{E}\left[\min\{Y_j,\, z_j\} \,\middle|\, X_j = x_{tr}\right] \qquad (20)\\ \text{s.t.} \quad & \sum_{j=1}^{d} z_j \le K_r \\ & z_j \ge 0 \qquad\qquad \forall j = 1,\ldots,d, \end{aligned}$$

where x_{tr} denotes auxiliary data available at the beginning of period t in the $(t,r)^{\text{th}}$ problem.

Note that had there been no capacity constraint in problem (20) and a per-unit ordering cost were added, the problem would decompose into d separate newsvendor problems, the solution to each being exactly a quantile regression on the regressors x_{tr}. As it is, the problem is coupled, but, fixing x_{tr}, the capacity constraint can be replaced with an equivalent per-unit ordering cost λ via Lagrangian duality and the optimal solution is attained by setting each z_j to the λ^{th} conditional quantile of Y_j. However, the reduction to quantile regression does not hold since the dual optimal value of λ depends *simultaneously* on all of the conditional distributions of Y_j for $j=1,\dots,d$.

5.2. Applying Predictive Prescriptions to Censored Data

In applying our approach to problem (20), we face the issue that we have data on sales, not demand. That is, our data on the quantity of interest Y is right-censored. In this section, we develop a modification of our approach to correct for this. The results in this section apply generally.

Suppose that instead of data $\{y^1,\dots,y^N\}$ on Y, we have data $\{u^1,\dots,u^N\}$ on $U=\min\{Y,V\}$ where V is an observable random threshold, data on which we summarize via $\delta=\mathbb{I}[U<V]$. For example, in our application, V is the on-hand inventory level at the beginning of the period. Overall, our data consists of $\tilde{S}_N=\{(x^1,u^1,\delta^1),\dots,(x^N,u^N,\delta^N)\}$.

In order to correct for the fact that our observations are in fact censored, we develop a conditional variant of the Kaplan-Meier method (cf. Kaplan and Meier (1958), Huh et al. (2011)) to transform our weights appropriately. Let (i) denote the ordering $u^{(1)}\le\cdots\le u^{(N)}$. Given the weights $w_{N,i}(x)$ generated based on the naïve assumption that $y^i=u^i$, we transform these into the weights

$$w_{N,(i)}^{\text{Kaplan-Meier}}(x)=\begin{cases}\left(\dfrac{w_{N,(i)}(x)}{\sum_{\ell=i}^N w_{N,(\ell)}(x)}\right)\displaystyle\prod_{k\le i-1\,:\,\delta^{(k)}=1}\left(\dfrac{\sum_{\ell=k+1}^N w_{N,(\ell)}(x)}{\sum_{\ell=k}^N w_{N,(\ell)}(x)}\right), & \text{if } \delta^{(i)}=1,\\ 0, & \text{if } \delta^{(i)}=0.\end{cases} \tag{21}$$

We next show that the transformation (21) preserves asymptotic optimality under certain conditions. The proof is in the E-companion.

THEOREM 6. *Suppose that Y and V are conditionally independent given X, that Y and V share no atoms, that for every $x\in\mathcal{X}$ the upper support of V given $X=x$ is greater than the upper support of Y given $X=x$, and that costs are bounded over y for each z (i.e., $|c(z;y)|\le g(z)$). Let $w_{N,i}(x)$ be as in* (12), (13), (15), *or* (16) *and suppose the corresponding assumptions of Theorem 2, 3, 4, or 5 apply. Let $\hat{z}_N(x)$ be as in* (3) *but using the transformed weights* (21). *Then $\hat{z}_N(x)$ is asymptotically optimal.*

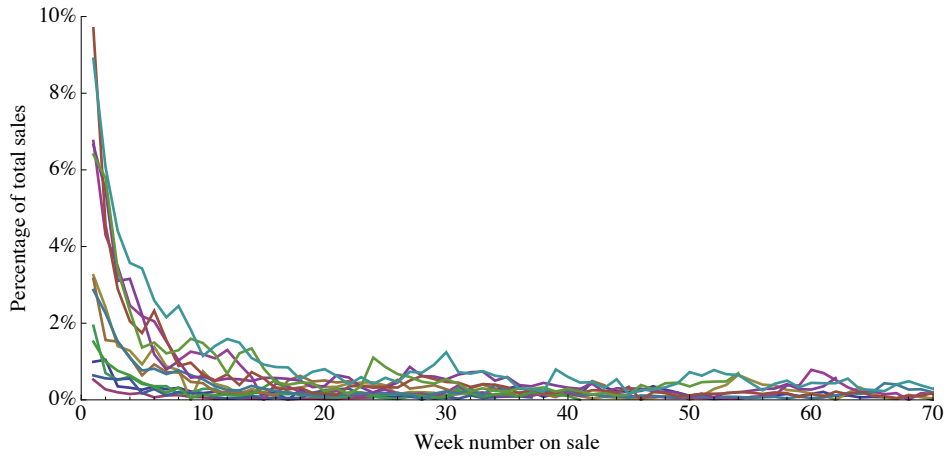


Figure 6 The percentage of all sales in the German state of Berlin taken up by each of 13 selected titles, starting from the point of release of each title to HE sales.

The assumption that Y and V share no atoms (which holds in particular if either is continuous) provides that $\delta \overset{a.s.}{=} \mathbb{I}[Y \leq V]$ so that the event of censorship is observable. In applying this to problem (20), the assumption that Y and V are conditionally independent given X will hold if X captures at least all of the information that past stocking decisions, which are made before Y is realized, may have been based on. The assumption on bounded costs applies to problem (20) because the cost (negative of the objective) is bounded in $[-K_r, 0]$.

5.3. Data

In this section, we describe the data collected. To get at the best data-driven predictive prescription, we combine both internal company data and public data harvested from online sources. The predictive power of such public data has been extensively documented in the literature (cf. Asur and Huberman (2010), Choi and Varian (2012), Goel et al. (2010), Da et al. (2011), Gruhl et al. (2005, 2004), Kallus (2014)). Here we study its prescriptive power.

Internal Data. The internal company data consists of 4 years of sale and inventory records across the network of retailers, information about each of the locations, and information about each of the items.

We aggregate the sales data by week (the replenishment period of interest) for each feasible combination of location and item. As discussed above, these sales-per-week data constitute a right-censored observation of weekly demand, where censorship occurs when an item is sold out. We developed the transformed weights (21) to tackle this issue exactly. Figure 6 shows the sales life cycle of a selection of titles in terms of their marketshare when they are released to home entertainment (HE) sales and onwards. Since new releases can attract up to almost 10% of sales in their first week of release, they pose a great sales opportunity, but at the same time significant demand uncertainty.

Information about retail locations includes to which chain a location belongs and the address of the location. To parse the address and obtain a precise position of the location, including country and subdivision, we used the Google Geocoding API (Application Programming Interface).[3]

Information about items include the medium (e.g. DVD or BluRay) and an item "title." The title is a short descriptor composed by a local marketing team in charge of distribution and sales in a particular region and may often include information beyond the title of the underlying content. For example, a hypothetical film titled *The Film* sold in France may be given the item title "THE FILM DVD + LIVRET - EDITION FR", implying that the product is a French edition of the film, sold on a DVD, and accompanied by a booklet (*livret*), whereas the same film sold in Germany on BluRay may be given the item title "FILM, THE (2012) - BR SINGLE", indicating it is sold on a single BluRay disc.

Public Data: Item Metadata, Box Office, and Reviews. We sought to collect additional data to characterize the items and how desirable they may be to consumers. For this we turned to the Internet Movie Database (IMDb; `www.imdb.com`) and Rotten Tomatoes (RT; `www.rottentomatoes.com`). IMDb is an online database of information on films and TV series. RT is a website that aggregates professional reviews from newspapers and online media, along with user ratings, of films and TV series. Example screen shots from IMDb and RT showing details about the 2012 movie *Skyfall* are shown in Figure 7.

In order to harvest information from these sources on the items being sold by the vendor, we first had to disambiguate the item entities and extract original content titles from the item titles. Having done so, we extract the following information from IMDb:

1. type (film, TV, other/unknown);
2. US original release date of content (e.g. in theaters);
3. average IMDb user rating (0-10);
4. number of IMDb users voting on rating;
5. number of awards (e.g. Oscars for films, Emmys for TV) won and number nominated for;
6. the main actors (i.e., first-billed);
7. plot summary (30-50 words);
8. genre(s) (of 26; can be multiple); and
9. MPAA rating (e.g. PG-13, NC-17) if applicable.

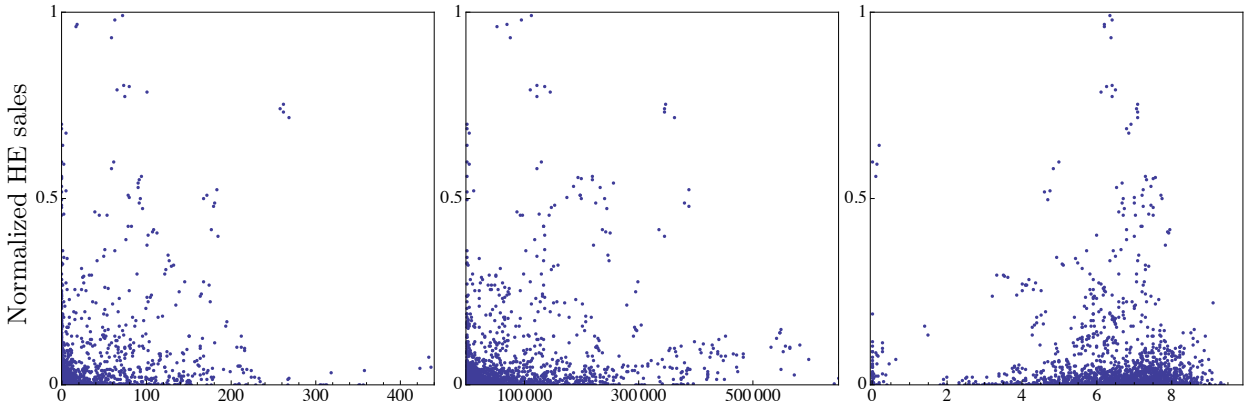
And the following information from RT:

10. professional reviewers' aggregate score;
11. RT user aggregate rating;

[3] See `https://developers.google.com/maps/documentation/geocoding` for details.

(a) IMDb (b) Rotten Tomatoes

Figure 7 Screen shots from the websites of IMDb and Rotten Tomatoes, displaying details for 2012 Bond movie Skyfall and showing such meta-data as release date, user rating, number of user rating votes, plot summary, first-billed actors, MPAA rating, and aggregate reviews.



(a) Box office gross (\$M), $\rho = 0.32$ (b) Number IMDb votes, $\rho = 0.25$ (c) Average IMDb rating, $\rho = 0.02$

Figure 8 Scatter plots of various data from IMDb and RT (horizontal axes) against total European sales during first week of HE release (vertical axes, rescaled to anonymize) and corresponding coefficients of correlation (ρ).

12. number of RT users voting on rating; and
13. if item is a film, then American box office gross when shown in theaters.

In Figure 8, we provide scatter plots of some of these attributes against sale figures in the first week of HE release. Notice that the number of users voting on the rating of a title is much more indicative of HE sales than the quality of a title as reported in the aggregate score of these votes.

Public Data: Search Engine Attention. In the above, we saw that box office gross is reasonably informative about future HE sale figures. The box office gross we are able to access, however, is for the American market and is also missing for various European titles. We therefore would like additional data to quantify the attention being given to different titles and to understand the local nature of such attention. For this we turned to Google Trends (GT; `www.google.com/trends`).[4]

[4] While GT is available publicly online, access to massive-scale querying and week-level trends data is not public. See acknowledgements.

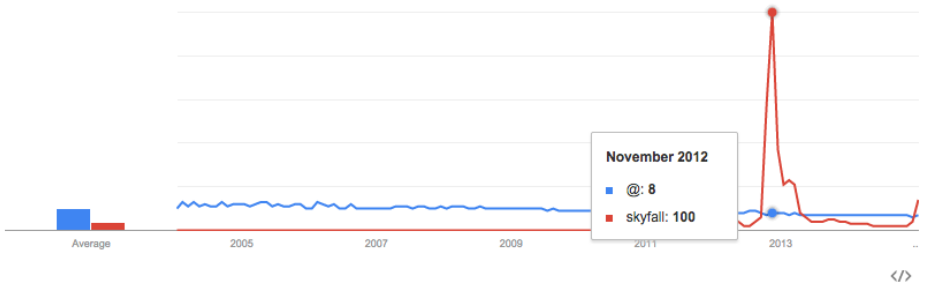
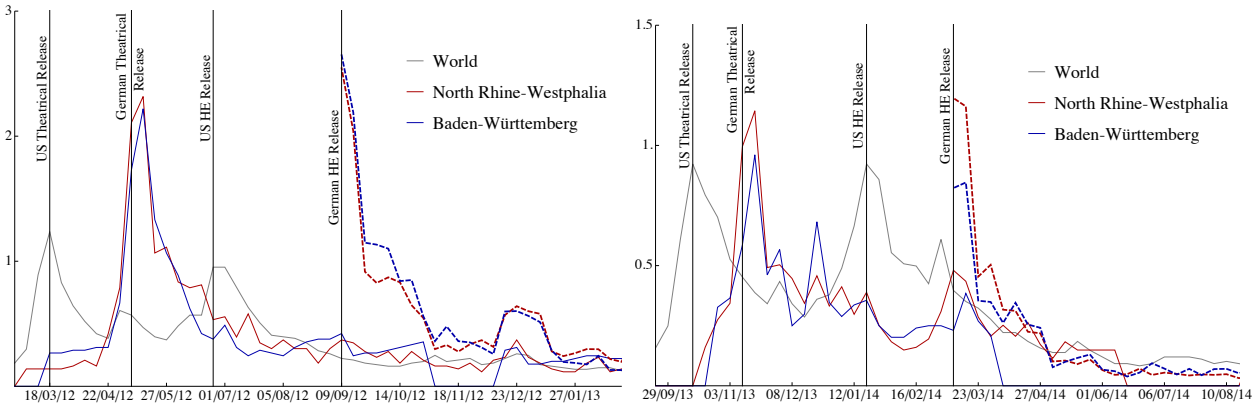


Figure 9 **Screen shot from Google Trends, comparing searches for "Skyfall" (red) and "@" (blue) in Germany.**



(a) *Unnamed film 1* (b) *Unnamed film 2*

Figure 10 **Weekly search engine attention for two unnamed films in the world and in two populous German states (solid lines) and weekly HE sales for the same films in the same states (dashed lines). Search engine attention and sales are both shown relative to corresponding overall totals in the respective region. The scales are arbitrary but common between regions and the two plots.**

GT provides data on the volume of Google searches for a given search term by time and geographic location. An example screen shot from GT is seen in Figure 9. GT does not provide absolute search volume figures, only volume time series given in terms relative to itself or to another series and in whole-number precision between 0 and 100. Therefore, to compare different such series, we establish as a baseline the search volume for the query "@", which works well because it has a fairly constant volume across the regions of Europe and because its search volume is neither too high (else the volume for another query would be drowned out by it and reported as 0) nor too low (else the volume for another query would overwhelm it and have it be reported as 0, in which case it would be useless as a reference point).

For each title, we measure the relative Google search volume for the search term equal to the original content title in each week from 2011 to 2014 (inclusive) over the whole world, in each European country, and in each country subdivision (states in Germany, cantons in Switzerland, autonomous communities in Spain, etc.). In each such region, after normalizing against the volume of our baseline query, the measurement can be interpreted as the fraction of Google searches for the

title in a given week out of all searches in the region, measured on an arbitrary but (approximately) common scale between regions.

In Figure 10, we compare this search engine attention to sales figures in Germany for two unnamed films.[5] Comparing panel (a) and (b), we first notice that the overall scale of sales correlates with the overall scale of *local* search engine attention at the time of theatrical release, whereas the global search engine attention is less meaningful (note vertical axis scales, which are common between the two figures). Looking closer at differences between regions in panel (b), we see that, while showing in cinemas, unnamed film 2 garnered more search engine attention in North Rhine-Westphalia (NW) than in Baden-Württemberg (BW) and, correspondingly, HE sales in NW in the first weeks after HE release were greater than in BW. In panel (a), unnamed film 1 garnered similar search engine attention in both NW and BW and similar HE sales as well. In panel (b), we see that the search engine attention to unnamed film 2 in NW accelerated in advance of the HE release, which was particularly successful in NW. In panel (a), we see that a slight bump in search engine attention 3 months into HE sales corresponded to a slight bump in sales. These observations suggest that local search engine attention both at the time of local theatrical release and in recent weeks may be indicative of future sales volumes.

5.4. Constructing Auxiliary Data Features and a Random Forest Prediction

For each instance (t, r) of problem (20) and for each item i we construct a vector of numeric predictive features x_{tri} that consist of backward cumulative sums of the sale volume of the item i at location r over the past 3 weeks (as available; e.g., none for new releases), backward cumulative sums of the total sale volume at location r over the past 3 weeks, the overall mean sale volume at location r over the past 1 year, the number of weeks since the original release data of the content (e.g., for a new release this is the length of time between the premier in theaters to release on DVD), an indicator vector for the country of the location r, an indicator vector for the identity of chain to which the location r belongs, the total search engine attention to the title i over the first two weeks of local theatrical release globally, in the country, and in the country-subdivision of the location r, backward cumulative sums of search engine attention to the title i over the past 3 weeks globally, in the country, and in the country-subdivision of the location r, and features capturing item information harvested from IMDb and RT as described below.

[5] These films must remain unnamed because a simple search can reveal their European distributor and hence the vendor who prefers their identity be kept confidential.

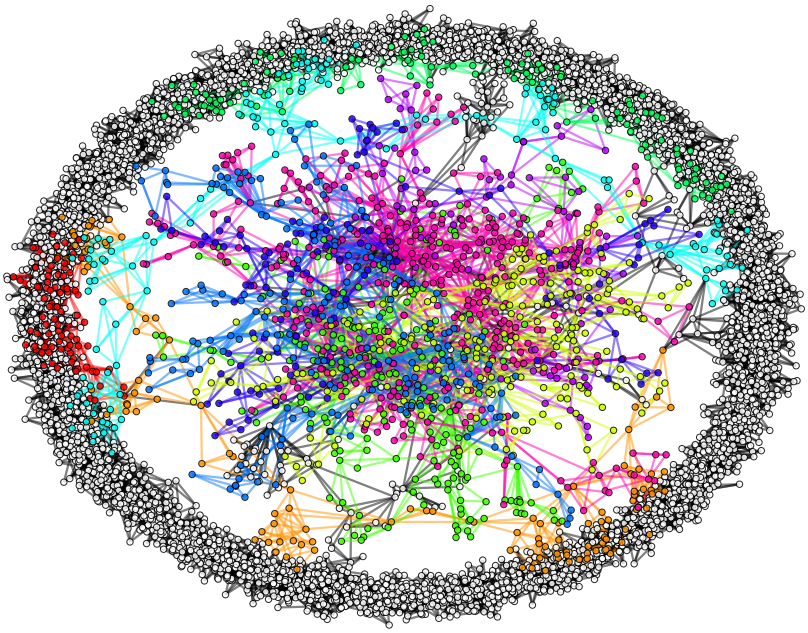


Figure 11 The graph of actors, connected via common movies where both are first-billed. Colored nodes correspond to the 10 largest communities of actors. Colored edges correspond to intra-community edges.

For some information harvested from IMDb and RT, the corresponding numeric feature is straightforward (e.g. number of awards). For other pieces of information, some distillation is necessary. For genre, we create an indicator vector. For MPAA rating, we create a single ordinal (from 1 for G to 5 for NC-17). For plot, we measure the cosine-similarity between plots,

$$\text{similarity}(P_1, P_2) = \frac{p_1^T p_2}{||p_1||\,||p_2||}$$

where p_{ki} denotes the number of times word i appears in plot text P_k and i indexes the collection of unique words appearing in plots P_1, P_2 ignoring common words like "the",

and use this as a distance measure to hierarchically cluster the plots using Ward's method (cf. Ward (1963)). This captures common themes in titles. We construct 12 clusters based solely on historical data and, for new data, include a feature vector of median cosine similarity to each of the clusters. For actors, we create a graph with titles as nodes and with edges between titles that share actors, weighted by the number of actors shared. We use the method of Blondel et al. (2008) to find communities of titles and create an actor-counter vector for memberships in the 10 largest communities (see Figure 11). This approach is motivated by the existence of such actor groups as the "Rat Pack" (Humphery Bogart and friends), "Brat Pack" (Molly Ringwald and friends), and "Frat Pack" (Owen Wilson and friends) that often co-star in titles with a similar theme, style, and target audience.

We end up with $d_x = 91$ numeric predictive features. Having summarized these numerically, we train a RF of 500 trees to predict sales. In training the RF, we normalize each the sales in each

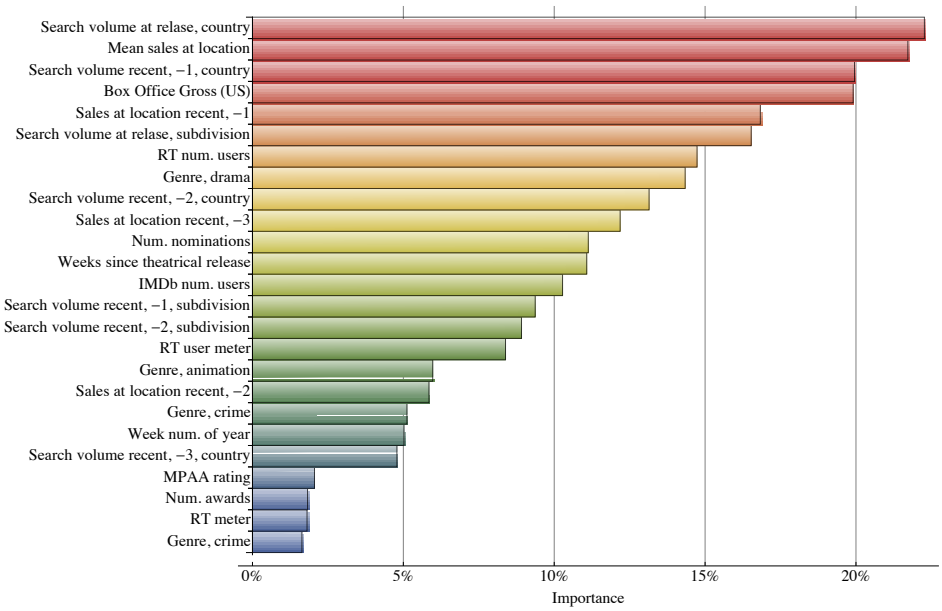


Figure 12 The 25 top x variables in predictive importance, measured as the average over forest trees of the change in mean-squared error of the tree (shown here as percentage of total variance) when the value of the variables is randomly permuted among the out-of-bag training data.

instance by the training-set average sales in the corresponding location; we de-normalize after predicting. To capture the decay in demand from time of release in stores, we train a separate RFs for sale volume on the k^{th} week on the shelf for $k = 1, \ldots, 35$ and another RF for the "steady state" weekly sale volume after 35 weeks.

For $k = 1$, we are predicting the demand for a new release, the uncertainty of which, as discussed in Section 5.1, constitutes one of the greatest difficulties of the problem to the company. In terms of predictive quality, when measuring out-of-sample performance we obtain an $R^2 = 0.67$ for predicting sale volume for new releases. The 25 most important features in this prediction are given in Figure 12. In Figure 13, we show the R^2 obtained also for predictions at later times in the product life cycle, compared to the performance of a baseline heuristic that always predicts for next week the demand of last week.

Considering the uncertainty associated with new releases, we feel that this is a positive result, but at the same time what truly matters is the performance of the prescription in the problem. We discuss this next.

5.5. Applying Our Predictive Prescriptions to the Problem

In the last section we discussed how we construct RFs to predict sales, but our problem of interest is to prescribe order quantities. To solve our problem (20), we use the trees in the forests we trained to construct weights $w_{N,i}(x)$ exactly as in (18), then we transform these as in (21), and finally we prescribe data-driven order quantities $\hat{z}_N(x)$ as in (8). Thus, we use our data to go from

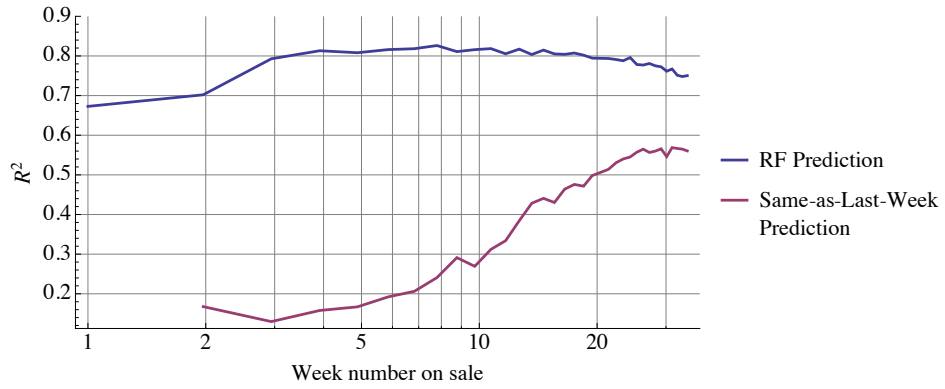


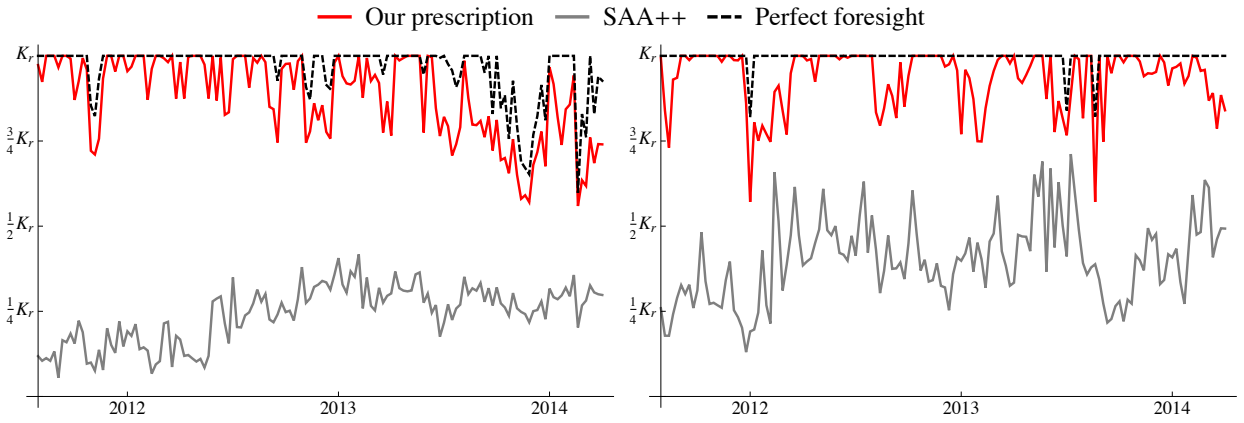
Figure 13 Out-of-sample coefficients of determination R^2 for predicting demand next week at different stages of product life cycle.

an observation $X = x$ of our varied auxiliary data directly to a replenishment decision on order quantities.

We would like to test how well our prescription does out-of-sample and as an actual live policy. To do this we consider what we would have done over the 150 weeks from December 19, 2011 to November 9, 2014 (inclusive). At each week, we consider only data from time prior to that week, train our RFs on this data, and apply our prescription to the current week. Then we observe what had actually materialized and score our performance.

There is one issue with this approach to scoring: our historical data only consists of sales, not demand. While we corrected for the adverse effect of demand censorship on our prescriptions using the transformation (21), we are still left with censored demand when scoring performance as described above. In order to have a reasonable measure of how good our method is, we therefore consider the problem (20) with capacities K_r that are a *quarter* of their nominal values. In this way, demand censorship hardly ever becomes an issue in the scoring of performance. To be clear, this correction is necessary just for a counterfactual scoring of performance; not in practice. The transformation (21) already corrects for prescriptions trained on censored observations of the quantity Y that affects true costs.

We compare the performance of our method with two other quantities. One is the performance of the perfect-forecast policy, which knows future demand exactly (no distributions). Another is the performance of a data-driven policy without access to the auxiliary data (i.e., SAA). Because the decay of demand over the lifetime of a product is significant, to make it a fair comparison we let this policy depend on the distributions of product demand based on how long its been on the market. That is, it is based on T separate datasets where each consists of the demands for a product after t weeks on the market (again, considering only past data). Due to this handicap we term it SAA++ henceforth.



(a) Munich, Germany, $P = 0.89$

(b) Paris, France, $P = 0.90$

(c) Waterloo, Belgium, $P = 0.85$

(d) The Hague, Netherlands, $P = 0.86$

Figure 14 **The performance of our prescription over time. The vertical axis is shown in terms of the location's capacity, K_r.**

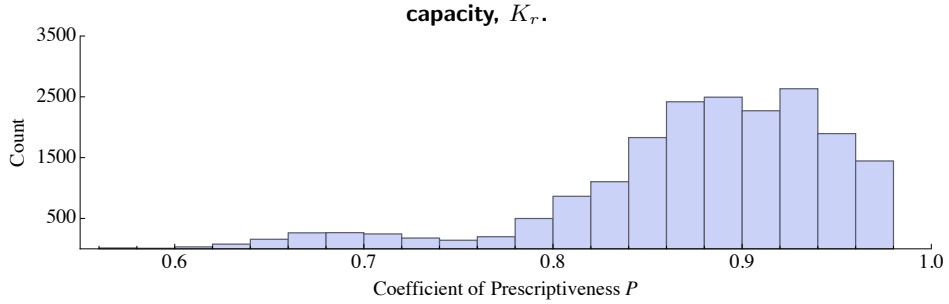


Figure 15 **The distribution of coefficients of prescriptiveness P over retail locations.**

The ratio of the difference between our performance and that of the prescient policy and the difference between the performance of SAA++ and that of the prescient policy is the coefficient of prescriptiveness P. When measured out-of-sample over the 150-week period as these policies make live decisions, we get $P = 0.88$. Said another way, in terms of our objective (sales volumes), our data X and our prescription $\hat{z}_N(x)$ gets us 88% of the way from the best data-poor decision to the impossible perfect-foresight decision. This is averaged over just under 20,000 locations. In Figure 14, we plot the performance over time at four specific locations, the city of which is noted.

In Figure 15, we plot the overall distribution of coefficients of prescriptiveness P over all retail locations in Europe.

6. Alternative Approaches

In the beginning of Section 2, we noted that the empirical distribution is insufficient for approximating the full-information problem (2). The solution was to consider local neighborhoods in approximating conditional expected costs; these were computed separately for each x. Another approach would be to develop an explicit decision rule and impose structure on it. In this section, we consider an approach to constructing a predictive prescription by selecting from a family of linear functions restricted in some norm,

$$\mathcal{F} = \left\{ z(x) = Wx \,:\, W \in \mathbb{R}^{d_z \times d_x},\, ||W|| \leq R \right\}, \tag{22}$$

so to minimize the empirical marginal expected costs as in (4),

$$\hat{z}_N(\cdot) \in \operatorname*{arg\,min}_{z(\cdot)\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} c(z(x^i); y^i).$$

The linear decision rule can be generalized by transforming X to include nonlinear terms.

We consider two examples of a norm on the matrix of linear coefficients, W. One is the row-wise p,p'-norm:

$$||W|| = \left|\left| \left(\gamma_1 \left|\left|W_1\right|\right|_p, \ldots, \gamma_d \left|\left|W_d\right|\right|_p \right) \right|\right|_{p'}.$$

Another is the Schatten p-norm:

$$||W|| = \left|\left| \left(\tau_1, \ldots, \tau_{\min\{d_z,d_x\}}\right) \right|\right|_p \quad \text{where } \tau_i \text{ are } W\text{'s singular values.}$$

For example, the Schatten 1-norm is the matrix nuclear norm. In either case, the restriction on the norm is equivalent to an appropriately-weighted regularization term incorporated into the objectives of (4).

Problem (4) corresponds to the traditional framework of empirical risk minimization in statistical learning with a general loss function. There is no great novelty in this formulation except for the potential multivariateness of Y and z. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z;y) = (z-y)^2$, problem (4) corresponds to least-squares regression. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z;y) = (y-z)(\tau - \mathbb{I}\left[y - z < 0\right])$, problem (4) corresponds to quantile regression, which estimates the conditional τ-quantile as a function of x. Rearranging terms, $c(z;y) = (y-z)(\tau - \mathbb{I}\left[y - z < 0\right]) = \max\left\{(1-\tau)(z-y),\, \tau(y-z)\right\}$ is the same as the newsvendor cost function where τ is the service level requirement. Rudin and Vahn (2014) consider this cost function and the selection of a linear decision rule with regularization on ℓ_1-norm in order to solve a newsvendor problem with auxiliary data. Quantile regression

(cf. Koenker (2005)) and ℓ_1-regularized quantile regression (cf. Belloni and Chernozhukov (2011)) are standard techniques in regression analysis. Because most OR/MS problems involve multivariate uncertainty and decisions, in this section we generalize the approach and its associated theoretical guarantees to such multivariate problems ($d_y \geq 1$, $d_z \geq 1$).

Before continuing, we note a few limitations of any approach based on (4). For general problems, there is no reason to expect that optimal solutions will have a linear structure (whereas certain distributional assumptions lead to such conclusions in least-squares and quantile regression analyses). In particular, unlike the predictive prescriptions studied in Section 2, the approach based on (4) does not enjoy the same universal guarantees of asymptotic optimality. Instead, we will only have out-of-sample guarantees that depend on our class $\mathcal{F}$ of decision rules.

Another limitation is the difficulty in restricting the decisions to a constrained feasible set $\mathcal{Z} \neq \mathbb{R}^{d_z}$. Consider, for example, the portfolio allocation problem from Section 1.1, where we must have $\sum_{i=1}^{d_x} z_i = 1$. One approach to applying (4) to this problem might be to set $c(z;y) = \infty$ for $z \notin \mathcal{Z}$ (or, equivalently, constrain $z(x^i) \in \mathcal{Z}$ $\forall i$). However, not only will this not guarantee that $z(x) \in \mathcal{Z}$ for x outside the dataset, but we would also run into a problem of infeasibility as we would have N linear equality constraints on $d_z \times d_x$ linear coefficients (a constraint such as $\sum_{i=1}^{d_x} z_i \leq 1$ that does not reduce the affine dimension will still lead to an undesirably flat linear decision rule as N grows). Another approach may be to compose $\mathcal{F}$ with a projection onto $\mathcal{Z}$, but this will generally lead to a non-convex optimization problem that is intractable to solve. Therefore, the approach is limited in its applicability to OR/MS problems.

In a few limited cases, we may be able to sensibly extend the cost function synthetically outside the feasible region while maintaining convexity. For example, in the shipment planning example of Section 1.1, we may allow negative order quantities z and extend the first-stage costs to depend only on the positive part of z, i.e. $p_1 \sum_{i=1}^{d_z} \max\{z_i, 0\}$ (but leave the second-stage costs as they are for convexity). Now, if after training $\hat{z}_N(\cdot)$, we transform any resulting decision by only taking the positive part of each order quantity, we end up with a feasible decision rule whose costs are no worse than the synthetic costs of the original rule. If we follow this approach and apply (4), either without restrictions on norms or with a diminishing Frobenius norm penalty on coefficients, we end up with results as shown in Figure 16. The results suggest that, while we are able to apply the approach to the problem, restricting to linear decision rules is inefficient in this particular problem.

In the rest of this section we consider the application of the approach (4) to problems where y and z are multivariate and $c(z;y)$ is general, but only treat unconstrained decisions $\mathcal{Z} = \mathbb{R}^{d_z}$.

6.1. Tractability

We first develop sufficient conditions for the problem (4) to be optimized in polynomial time. The proof is in the E-companion.

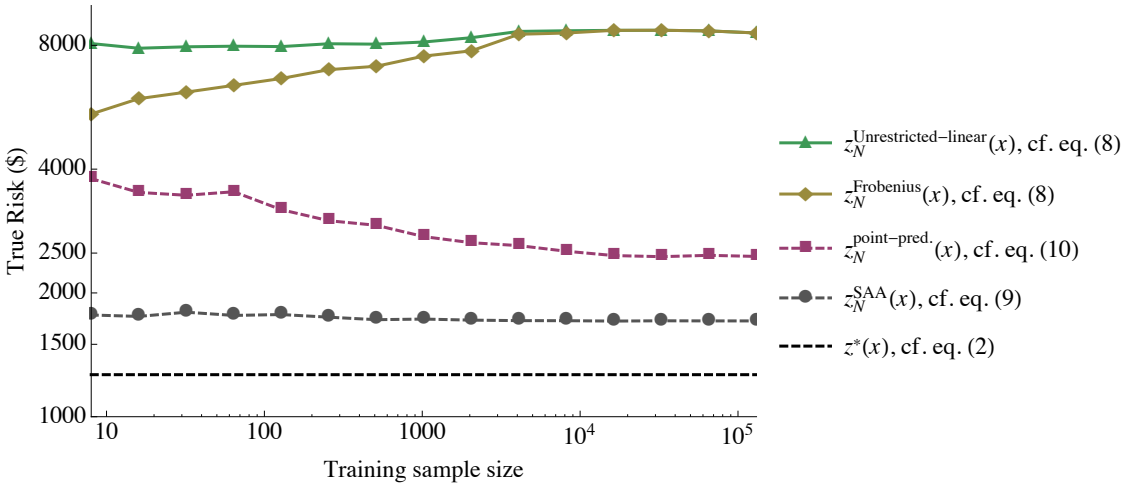


Figure 16 **Performance of various prescriptions in the shipment planning example. Note the horizontal and vertical log scales.**

THEOREM 7. *Suppose that $c(z;y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z. Then for any fixed x we can find an ϵ-optimal solution to (4) in time and oracle calls polynomial in n, d, $\log(1/\epsilon)$ for $\mathcal{F}$ as in (22).*

6.2. Out-of-Sample Guarantees

Next, we characterize the out-of-sample guarantees of a predictive prescription derived from (4). All proofs are in the E-companion. In the traditional framework of empirical risk minimization in statistical learning such guarantees are often derived using Rademacher complexity but these only apply to univariate problems (c.f. Bartlett and Mendelson (2003)). Because most OR/MS problems are multivariate, we generalize this theory appropriately. We begin by generalizing the definition of Rademacher complexity to multivariate-valued functions.

DEFINITION 2. Given a sample $S_N = \{s_1, \ldots, s_N\}$, The *empirical multivariate Rademacher complexity* of a class of functions $\mathcal{F}$ taking values in $\mathbb{R}^d$ is defined as

$$\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N) = \mathbb{E}\left[\frac{2}{N}\sup_{g\in\mathcal{F}}\sum_{i=1}^{n}\sum_{k=1}^{d}\sigma_{ik}g_k(s_i)\Big| s_1, \ldots, s_n\right]$$

where σ_{ik} are independently equiprobably $+1, -1$. The *marginal multivariate Rademacher complexity* is defined as

$$\mathfrak{R}_N(\mathcal{F}) = \mathbb{E}\left[\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N)\right]$$

over the sampling distribution of S_N.

Note that given only data S_N, the quantity $\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N)$ is observable. Note also that when $d = 1$ the above definition coincides with the common definition of Rademacher complexity.

The theorem below relates the multivariate Rademacher complexity of $\mathcal{F}$ to out-of-sample guarantees on the performance of the corresponding predictive prescription $\hat{z}_N(x)$ from (4). A generalization of the following to mixing processes is given in the supplemental Section 9. We denote by $S_N^x = \{x^1, \ldots, x^N\}$ the restriction of our sample to data on X.

THEOREM 8. *Suppose $c(z;y)$ is bounded and equi-Lipschitz in z:*

$$\sup_{z\in\mathcal{Z},\, y\in\mathcal{Y}} \quad c(z;y) \le \overline{c},$$
$$\sup_{z\neq z'\in\mathcal{Z},\, y\in\mathcal{Y}} \quad \frac{c(z;y)-c(z';y)}{||z_k - z'_k||_\infty} \le L < \infty.$$

Then, for any $\delta > 0$, we have that with probability $1-\delta$,

$$\mathbb{E}\left[c(z(X);Y)\right] \le \frac{1}{N}\sum_{i=1}^N c(z(x^i);y^i) + \overline{c}\sqrt{\log(1/\delta')/2N} + L\mathfrak{R}_N(\mathcal{F}) \qquad \forall z\in\mathcal{F}, \tag{23}$$

and that, again, with probability $1-\delta$,

$$\mathbb{E}\left[c(z(X);Y)\right] \le \frac{1}{N}\sum_{i=1}^N c(z(x^i);y^i) + 3\overline{c}\sqrt{\log(2/\delta'')/2N} + L\widehat{\mathfrak{R}}_N(\mathcal{F};S_N^x) \qquad \forall z\in\mathcal{F}. \tag{24}$$

In particular, these hold for $z = \hat{z}_N(\cdot) \in \mathcal{F}$.

Equations (23) and (24) provide a bound on the out-of-sample performance of any predictive prescription $z(\cdot)\in\mathcal{F}$. The bound is exactly what we minimize in problem (4) because the extra terms do not depend on $z(\cdot)$. That is, we minimize the empirical risk, which, with additional confidence terms, bounds the true out-of-sample costs of the resulting predictive prescription $\hat{z}_N(\cdot)$.

These confidence terms involve the multivariate Rademacher complexity of our class $\mathcal{F}$ of decision rules. In the next lemmas, we compute appropriate bounds on the complexity of our examples of classes $\mathcal{F}$. The theory, however, applies beyond linear rules.

LEMMA 1. *Consider $\mathcal{F}$ as in (22) with row-wise p,p' norm for $p\in[2,\infty)$ and $p'\in[1,\infty]$. Let q be the conjugate exponent of p ($1/p+1/q=1$) and suppose that $||x||_q \le M$ for all $x\in\mathcal{X}$. Then*

$$\mathfrak{R}_N(\mathcal{F}) \le 2MR\sqrt{\frac{p-1}{N}}\sum_{k=1}^{d_z}\frac{1}{\gamma_k}.$$

LEMMA 2. *Consider $\mathcal{F}$ as in (22) with Schatten p-norm. Let $r = \max\{1-1/p, 1/2\}$. Then*

$$\widehat{\mathfrak{R}}_N(\mathcal{F};S_N^x) \le 2Rd_z^r\sqrt{\frac{1}{N}}\sqrt{\widehat{\mathbb{E}}_{S_N^x}||X||_2^2}$$
$$\mathfrak{R}_N(\mathcal{F}) \le 2Rd_z^r\sqrt{\frac{1}{N}}\sqrt{\mathbb{E}||X||_2^2},$$

where $\widehat{\mathbb{E}}_{S_N^x}$ denotes empirical average over the sample S_N^x.

The above results indicate that the confidence terms in equations (23) and (24) shrink to 0 as $N\to\infty$ even if we slowly relax norm restrictions. Hence, we can approach the optimal out-of-sample performance over the class $\mathcal{F}$ without restrictions on norms.

7. Concluding Remarks

In this paper, we combine ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems that leverage auxiliary observations. We motivate our methods based on existing predictive methodology from ML, but, in the OR/MS tradition, focus on the making of a decision and on the effect on costs, revenues, and risk. Our approach is

a) generally applicable,

b) tractable,

c) asymptotically optimal,

d) and leads to substantive and measurable improvements in a real-world context.

We feel that the above qualities, together with the growing availability of data and in particular auxiliary data in OR/MS applications, afford our proposed approach a potential for substantial impact in the practice of OR/MS.

Acknowledgments

We would like to thank Ross Anderson and David Nachum of Google for assistance in obtaining access to Google Trends data. This paper is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

References

Altman, Naomi S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**(3) 175–185.

Arya, Sunil, David Mount, Nathan Netanyahu, Ruth Silverman, Angela Wu. 1998. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM* **45**(6) 891–923.

Asur, Sitaram, Bernardo Huberman. 2010. Predicting the future with social media. *WI-IAT*. 492–499.

Bartlett, Peter, Shahar Mendelson. 2003. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482.

Belloni, Alexandre, Victor Chernozhukov. 2011. ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39**(1) 82–130.

Ben-Tal, Aharon, Laurent El Ghaoui, Arkadi Nemirovski. 2009. *Robust optimization*. Princeton University Press.

Bentley, Jon. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**(9) 509–517.

Beran, Rudolf. 1981. Nonparametric regression with randomly censored survival data. Tech. rep., Technical Report, Univ. California, Berkeley.

Berger, James O. 1985. *Statistical decision theory and Bayesian analysis*. Springer.

Bertsekas, Dimitri, Angelia Nedić, Asuman Ozdaglar. 2003. *Convex analysis and optimization*. Athena Scientific, Belmont.

Bertsekas, Dimitri P. 1995. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA.

Bertsekas, Dimitri P. 1999. *Nonlinear programming*. Athena Scientific, Belmont.

Bertsimas, Dimitris, David B Brown, Constantine Caramanis. 2011. Theory and applications of robust optimization. *SIAM review* **53**(3) 464–501.

Bertsimas, Dimitris, Vishal Gupta, Nathan Kallus. 2013. Data-driven robust optimization Submitted to *Operations Research*.

Bertsimas, Dimitris, Vishal Gupta, Nathan Kallus. 2014. Robust saa Submitted to *Mathematical Programming*.

Bertsimas, Dimitris, Nathan Kallus. 2015. From predictive to prescriptive analytics in multi-period decision problems .

Besbes, Omar, Assaf Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* **57**(6) 1407–1420.

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Birge, John R, Francois Louveaux. 2011. *Introduction to stochastic programming*. Springer.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10) P10008.

Bradley, Richard. 1986. Basic properties of strong mixing conditions. *Dependence in Probability and Statistics*. Birkhausser, 165–192.

Bradley, Richard. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2**(107-44) 37.

Breiman, Leo. 1996. Bagging predictors. *Machine learning* **24**(2) 123–140.

Breiman, Leo. 2001. Random forests. *Machine learning* **45**(1) 5–32.

Breiman, Leo, Jerome Friedman, Charles Stone, Richard Olshen. 1984. *Classification and regression trees*. CRC press.

Breiman, Leo, et al. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* **16**(3) 199–231.

Calafiore, Giuseppe, Marco C Campi. 2005. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming* **102**(1) 25–46.

Calafiore, Giuseppe Carlo, Laurent El Ghaoui. 2006. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications* **130**(1) 1–22.

Cameron, A Colin, Pravin K Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge university press.

Carrasco, Marine, Xiaohong Chen. 2002. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory* **18**(1) 17–39.

Choi, Hyunyoung, Hal Varian. 2012. Predicting the present with google trends. *Econ. Rec.* **88**(s1) 2–9.

Cleveland, William S, Susan J Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**(403) 596–610.

Da, Zhi, Joseph Engelberg, Pengjie Gao. 2011. In search of attention. *J. Finance* **66**(5) 1461–1499.

Delage, E, Y Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **55**(3) 98–112.

Devroye, Luc P, TJ Wagner. 1980. On the l 1 convergence of kernel estimators of regression functions with applications in discrimination. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **51**(1) 15–25.

Doukhan, Paul. 1994. *Mixing: Properties and Examples*. Springer.

Dudley, Richard M. 2002. *Real analysis and probability*, vol. 74. Cambridge University Press, Cambridge.

Fan, Jianqing. 1993. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics* 196–216.

Geurts, Pierre, Damien Ernst, Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* **63**(1) 3–42.

Goel, Sharad, Jake Hofman, Sébastien Lahaie, David Pennock, Duncan Watts. 2010. Predicting consumer behavior with web search. *PNAS* **107**(41) 17486–17490.

Grotschel, M, L Lovasz, A Schrijver. 1993. *Geometric algorithms and combinatorial optimization*. Springer, New York.

Gruhl, Daniel, Laurent Chavet, David Gibson, Jörg Meyer, Pradhan Pattanayak, Andrew Tomkins, J Zien. 2004. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Syst. J.* **43**(1) 64–77.

Gruhl, Daniel, Ramanathan Guha, Ravi Kumar, Jasmine Novak, Andrew Tomkins. 2005. The predictive power of online chatter. *SIGKDD*. 78–87.

Hanasusanto, Grani Adiwena, Daniel Kuhn. 2013. Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems*. 827–835.

Hannah, Lauren, Warren Powell, David M Blei. 2010. Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems*. 820–828.

Hansen, Bruce E. 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24**(03) 726–748.

Ho, Tin Kam. 1998. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**(8) 832–844.

Huh, Woonghee Tim, Retsef Levi, Paat Rusmevichientong, James B Orlin. 2011. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research* **59**(4) 929–941.

Kakade, Sham, Karthik Sridharan, Ambuj Tewari. 2008. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *NIPS*. 793–800.

Kallus, Nathan. 2014. Predicting crowd behavior with big public data. *WWW*. 23, 625630.

Kao, Yi-hao, Benjamin V Roy, Xiang Yan. 2009. Directed regression. *Advances in Neural Information Processing Systems*. 889–897.

Kaplan, Edward L, Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**(282) 457–481.

Kleywegt, Anton, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* **12**(2) 479–502.

Koenker, Roger. 2005. *Quantile regression*. 38, Cambridge university press.

Lai, Tze Leung, Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6**(1) 4–22.

Ledoux, Michel, Michel Talagrand. 1991. *Probability in Banach Spaces: isoperimetry and processes*. Springer.

Lehmann, Erich Leo, George Casella. 1998. *Theory of point estimation*, vol. 31. Springer.

Mcdonald, Daniel, Cosma Shalizi, Mark Schervish. 2011. Estimating beta-mixing coefficients. *AISTATS*. 516–524.

Mohri, Mehryar, Afshin Rostamizadeh. 2008. Rademacher complexity bounds for non-iid processes. *NIPS*. 1097–1104.

Mohri, Mehryar, Afshin Rostamizadeh, Ameet Talwalkar. 2012. *Foundations of machine learning*. MIT press.

Mokkadem, Abdelkader. 1988. Mixing properties of arma processes. *Stochastic Process. Appl.* **29**(2) 309–315.

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

Nemirovski, Arkadi, Anatoli Juditsky, Guanghui Lan, Alexander Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574–1609.

Parzen, Emanuel. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 1065–1076.

Robbins, Herbert. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58** 527–535.

Robbins, Herbert, Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Rockafellar, R Tyrrell. 1997. *Convex analysis*. Princeton university press.

Rockafellar, R Tyrrell, Roger J-B Wets. 1998. *Variational analysis*. Springer.

Rockafellar, Tyrrell, Stanislav Uryasev. 2000. Optimization of conditional value-at-risk. *J. Risk* **2** 21–42.

Roussas, George G. 1992. Exact rates of almost sure convergence of a recursive kernel estimate of a probability densiy function: Application to regression and hazard rate estimation. *Journal of Nonparametric Statistics* **1**(3) 171–195.

Rudin, Cynthia, Gah-Yi Vahn. 2014. The big data newsvendor: Practical insights from machine learning .

Shapiro, Alexander. 2003. Monte carlo sampling methods. *Handbooks Oper. Res. Management Sci.* **10** 353–425.

Shapiro, Alexander, Arkadi Nemirovski. 2005. On complexity of stochastic programming problems. *Continuous optimization*. Springer, 111–146.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Trevor, Hastie, Tibshirani Robert, Jerome Friedman. 2001. *The Elements of Statistical Learning*, vol. 1. Springer.

Vapnik, Vladimir. 1992. Principles of risk minimization for learning theory. *Advances in neural information processing systems*. 831–838.

Vapnik, Vladimir. 2000. *The nature of statistical learning theory*. springer.

Wald, Abraham. 1949. Statistical decision functions. *The Annals of Mathematical Statistics* 165–205.

Walk, Harro. 2010. Strong laws of large numbers and nonparametric estimation. *Recent Developments in Applied Probability and Statistics*. Springer, 183–214.

Ward, Joe. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**(301) 236–244.

Watson, Geoffrey. 1964. Smooth regression analysis. *Sankhyā A* 359–372.

Supplement

8. Extensions of Asymptotic Optimality to Mixing Processes and Proofs

In this supplemental section, we generalize the asymptotic results to mixing process and provide the omitted proofs.

8.1. Mixing Processes

We begin by defining stationary and mixing processes.

DEFINITION 3. A sequence of random variables $V_1, V_2, \dots$ is called *stationary* if joint distributions of finitely many consecutive variables are invariant to shifting. That is,

$$\mu_{V_t,\dots,V_{t+k}} = \mu_{V_s,\dots,V_{s+k}} \quad \forall s,t \in \mathbb{N},\, k \geq 0.$$

In particular, if a sequence is stationary then the variables have identical marginal distributions, but they may not be independent and the sequence may not be exchangeable. Instead of independence, mixing is the property that if standing at particular point in the sequence we look far enough ahead, the head and the tail look nearly independent, where "nearly" is defined by different metrics for different definitions of mixing.

DEFINITION 4. Given a stationary sequence $\{V_t\}_{t\in\mathbb{N}}$, denote by $\mathcal{A}^t = \sigma(V_1,\dots,V_t)$ the sigma-algebra generated by the first t variables and by $\mathcal{A}_t = \sigma(V_t, V_{t+1},\dots)$ the sigma-algebra generated by the subsequence starting at t. Define the *mixing coefficients at lag* k

$$\begin{aligned}
\alpha(k) &= \sup_{t\in\mathbb{N},\, A\in\mathcal{A}^t,\, B\in\mathcal{A}_{t+k}} |\mu(A\cap B) - \mu(A)\mu(B)| \\
\beta(k) &= \sup_{t\in\mathbb{N}} \left|\left| \mu_{\{V_s\}_{s\leq t}} \otimes \mu_{\{V_s\}_{s\geq t+k}} - \mu_{\{V_s\}_{s\leq t \vee s\geq t+k}} \right|\right|_{\mathrm{TV}} \\
\rho(k) &= \sup_{t\in\mathbb{N},\, Q\in L_2(\mathcal{A}^t),\, R\in L_2(\mathcal{A}_{t+k})} |\mathrm{Corr}(Q,R)|
\end{aligned}$$

where $||\cdot||_{\mathrm{TV}}$ is the total variance and $L_2(\mathcal{A})$ is the set of $\mathcal{A}$-measurable square-integrable real-valued random variables.

$\{V_t\}$ is said to be α-mixing if $\alpha(k) \overset{k\to\infty}{\longrightarrow} 0$, β-mixing if $\beta(k) \overset{k\to\infty}{\longrightarrow} 0$, and ρ-mixing if $\rho(k) \overset{k\to\infty}{\longrightarrow} 0$.

Notice that an iid sequence has $\alpha(k) = \beta(k) = \rho(k) = 0$. Bradley (1986) establishes that $2\alpha(k) \leq \beta(k)$ and $4\alpha(k) \leq \rho(k)$ so that either β- or ρ-mixing implies α-mixing.

Many processes satisfy mixing conditions under mild assumptions: auto-egressive moving-average (ARMA) processes (cf. Mokkadem (1988)), generalized autoregressive conditional heteroskedasticity (GARCH) processes (cf. Carrasco and Chen (2002)), and certain Markov chains. For a thorough discussion and more examples see Doukhan (1994) and Bradley (2005). Mixing rates are often given explicitly by model parameters but they can also be estimated from data (cf. Mcdonald

et al. (2011)). Sampling from such processes models many real-life sampling situations where observations are taken from an evolving system such as, for example, the stock market, inter-dependent product demands, or aggregates of doubly stochastic arrival processes as in the posts on social media.

8.2. Asymptotic Optimality

Let us now restate the results of Section 4.2 in more general terms, encompassing both iid and mixing conditions on S_N. We will also establish that our cost estimates converge, i.e.,

$$\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x)c(z;y^i) \to v^*(x), \tag{25}$$

for μ_X-almost-everywhere $x \in X$ (henceforth, μ_X-a.e.x) almost surely (henceforth, a.s.).

THEOREM 9 (**kNN**). *Suppose Assumptions 1, 2, and 3 hold and that S_N is generated by iid sampling. Let $w_{N,i}(x)$ be as in (12) with $k = \min\left\{\lceil CN^\delta\rceil, N-1\right\}$ for some $C>0$, $0<\delta<1$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and (25) holds.*

THEOREM 10 (**Kernel Methods**). *Suppose Assumptions 1, 2, and 3 hold and that $\mathbb{E}\left[|c(z;Y)|\max\left\{\log|c(z;Y)|,0\right\}\right]<\infty$ for each z. Let $w_{N,i}(x)$ be as in (13) with K being any of the kernels in (14) and $h=CN^{-\delta}$ for $C,\delta>0$. Let $\hat{z}_N(x)$ be as in (3). If S_N comes from*

1. *an iid process and $\delta<1/d_x$, or*
2. *a ρ-mixing process with $\rho(k)=O(k^{-\gamma})$ ($\gamma>0$) and $\delta<2\gamma/(d_x+2d_x\gamma)$, or*
3. *an α-mixing process with $\alpha(k)=O(k^{-\gamma})$ ($\gamma>1$) and $\delta<2(\gamma-1)/(3d_x+2d_x\gamma)$,*

then $\hat{z}_N(x)$ is asymptotically optimal and (25) holds.

THEOREM 11 (**Recursive Kernel Methods**). *Suppose Assumptions 1, 2, and 3 hold and that S_N comes from a ρ-mixing process with $\sum_{k=1}^{\infty}\rho(k)<\infty$ (or iid). Let $w_{N,i}(x)$ be as in (15) with K being the naïve kernel and with $h_i=Ci^{-\delta}$ for some $C>0$, $0<\delta<1/(2d_x)$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and (25) holds.*

THEOREM 12 (**Local Linear Methods**). *Suppose Assumptions 1, 2, and 3 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X, and that costs are bounded over y for each z (i.e., $|c(z;y)|\le g(z)$). Let $w_{N,i}(x)$ be as in (16) with K being any of the kernels in (14) and with $h_N=CN^{-\delta}$ for some $C,\delta>0$. Let $\hat{z}_N(x)$ be as in (3). If S_N comes from*

1. *an iid process and $\delta<1/d_x$, or*
2. *an α-mixing process with $\alpha(k)=O(k^{-\gamma})$, $\gamma>d_x+3$, and $\delta<(\gamma-d_x-3)/(d_x(\gamma-d_x+3))$,*

then $\hat{z}_N(x)$ is asymptotically optimal and (25) holds.

8.3. Proofs of Asymptotic Results for Local Predictive Prescriptions

First, we establish some preliminary results. In what follows, let

$$\begin{aligned}
C(z|x) &= \mathbb{E}\left[c(z;Y)\middle|X=x\right],\\
\widehat{C}_N(z|x) &= \sum_{i=1}^N w_{N,i}(x)c(z;y^i),\\
\mu_{Y|x}(A) &= \mathbb{E}\left[\mathbb{I}\left[Y\in A\right]\middle|X=x\right],\\
\hat{\mu}_{Y|x,N}(A) &= \sum_{i=1}^N w_{N,i}(x)\mathbb{I}\left[x_i\in A\right].
\end{aligned}$$

LEMMA 3. *If $\{(x^i,y^i)\}_{i\in\mathbb{N}}$ is stationary and $f:\mathbb{R}^{m_Y}\to\mathbb{R}$ is measurable then $\{(x^i,f(y^i))\}_{i\in\mathbb{N}}$ is also stationary and has mixing coefficients no larger than those of $\{(x^i,y^i)\}_{i\in\mathbb{N}}$.*

Proof This is simply because a transform can only make the generated sigma-algebra coarser. For a single time point, if f is measurable and $B\in\mathcal{B}(\mathbb{R})$ then by definition $f^{-1}(B)\in\mathcal{B}(\mathbb{R})$ and, therefore, $\{Y^{-1}(f^{-1}(B)):B\in\mathcal{B}(\mathbb{R})\}\subset\{Y^{-1}(B):B\in\mathcal{B}(\mathbb{R}^{m_Y})\}$. Here the transform is applied independently across time so the result holds ($f\times\cdots\times f$ remains measurable). □

LEMMA 4. *Suppose Assumptions 1 and 2 hold. Fix $x\in\mathcal{X}$ and a sample path of data such that, for every $z\in\mathcal{Z}$, $\widehat{C}_N(z|x)\to C(z|x)$. Then $\widehat{C}_N(z|x)\to C(z|x)$ uniformly in z over any compact subset of $\mathcal{Z}$.*

Proof Let any convergent sequence $z_N\to z$ and $\epsilon>0$ be given. By equicontinuity and $z_N\to z$, $\exists N_1$ such that $|c(z_N;y)-c(z;y)|\le\epsilon/2\ \forall N\ge N_1$. Then $\left|\widehat{C}_N(z_N|x)-\widehat{C}_N(z|x)\right|\le\mathbb{E}_{\hat{\mu}_{Y|x,N}}|c(z_N;y)-c(z;y)|\le\epsilon/2\ \forall N\ge N_1$. By assumption $\widehat{C}_N(z|x)\to C(z|x)$ and hence $\exists N_2$ such that $\left|\widehat{C}_N(z|x)-C(z|x)\right|\le\epsilon/2$. Therefore, for $N\ge\max\{N_1,N_2\}$,

$$\left|\widehat{C}_N(z_N|x)-C(z|x)\right|\le\left|\widehat{C}_N(z_N|x)-\widehat{C}_N(z|x)\right|+\left|\widehat{C}_N(z|x)-C(z|x)\right|\le\epsilon.$$

Hence $\widehat{C}_N(z_N|x)\to C(z|x)$ for any convergent sequence $z_N\to z$.

Now fix $E\subset\mathcal{Z}$ compact and suppose for contradiction that $\sup_{z\in E}\left|\widehat{C}_N(z|x)-C(z|x)\right|\not\to 0$. Then $\exists\epsilon>0$ and $z_N\in E$ such that $\left|\widehat{C}_N(z_N|x)-C(z_N|x)\right|\ge\epsilon$ infinitely often. Restricting first to a subsequence where this always happens and then using the compactness of E, there exists a convergent subsequence $z_{N_k}\to z\in E$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z_{N_k}|x)\right|\ge\epsilon$ for every k. Then,

$$0<\epsilon\le\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z_{N_k}|x)\right|\le\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z|x)\right|+\left|C(z|x)-C(z_{N_k}|x)\right|.$$

Since $z_{N_k}\to z$, we have shown before that $\exists k_1$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x)-C(z|x)\right|\le\epsilon/2\ \forall k\ge k_1$. By equicontinuity and $z_{N_k}\to z$, $\exists k_2$ such that $\left|c(z_{N_k};y)-c(z;y)\right|\le\epsilon/4\ \forall k\ge k_2$. Hence, also $\left|C(z|x)-C(z_{N_k}|x)\right|\le\mathbb{E}\left[\left|c(z_{N_k};y)-c(z;y)\right|\middle|X=x\right]\le\epsilon/4\ \forall k\ge k_2$. Considering $k=\max\{k_1,k_2\}$ we get the contradiction that $0<\epsilon\le\epsilon/2$. □

LEMMA 5. *Suppose Assumptions 1, 2, and 3 hold. Fix $x \in \mathcal{X}$ and a sample path of data such that $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly and, for every $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to C(z|x)$. Then $\lim_{N\to\infty}\left(\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)\right) = v^*(x)$ and every sequence $z_N \in \arg\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)$ satisfies $\lim_{N\to\infty} C(z_N|x) = v^*(x)$ and all of its limit points are contained in $\arg\min_{z\in\mathcal{Z}} C(z|x)$.*

Proof Suppose that case 1 or 3 of Assumption 3 holds (i.e. boundedness or infinite limit). First, we show $\widehat{C}_N(z|x)$ and $C(z|x)$ are continuous and eventually coercive. Let $\epsilon > 0$ be given. By equicontinuity, $\exists \delta > 0$ such that $|c(z;y) - c(z';y)| \ \forall y \in \mathcal{Y}$ whenever $||z - z'|| \le \delta$. Hence, whenever $||z - z'|| \le \delta$, we have $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \le \mathbb{E}_{\hat{\mu}_{Y|x,N}} |c(z;y) - c(z';y)| \le \epsilon$ and $|C(z|x) - C(z'|x)| \le \mathbb{E}\left[|c(z;y) - c(z';y)| \,\middle|\, X = x\right] \le \epsilon$. This gives continuity. Coerciveness is trivial if $\mathcal{Z}$ is bounded. Suppose it is not. Without loss of generality D_x is compact, otherwise we can take any compact subset of it that has positive probability on it. Then by assumption of weak convergence $\exists N_0$ such that $\hat{\mu}_{Y|x,N}(D_x) \ge \mu_{Y|x}(D_x)/2 > 0$ for all $N \ge N_0$. Now let $z_k \in \mathcal{Z}$ be any sequence such that $||z_k|| \to \infty$. Let $M > 0$ be given. Let $\lambda' = \liminf_{||k||\to\infty} \inf_{y\notin D_x} c(z_k;y)$ and $\lambda = \max\{\lambda', 0\}$. By assumption $\lambda' > -\infty$. Hence $\exists k_0$ such that $\inf_{y\notin D_x} c(z_k;y) \ge \lambda' \ \forall k \ge k_0$. By D_x-uniform coerciveness and $||z_k|| \to \infty$, $\exists k_1 \ge k_0$ such that $c(z_k;y) \ge (2M - 2\lambda)/\mu_{Y|x}(D_x) \ \forall k \ge k_1$ and $y \in D_x$. Hence, $\forall k \ge k_1$ and $N \ge N_0$,

$$C(z|x) \ge \mu_{Y|x}(D) \times (2M - 2\lambda)/\mu_{Y|x}(D_x) + (1 - \mu_{Y|x}(D))\lambda' \ge 2M - 2\lambda + \lambda \ge M,$$
$$\widehat{C}_N(z|x) \ge \hat{\mu}_{Y|x,N}(D) \times (2M - 2\lambda)/\hat{\mu}_{Y|x,N}(D_x) + (1 - \hat{\mu}_{Y|x,N}(D))\lambda' \ge M - \lambda + \lambda = M,$$

since $\alpha\lambda' \ge \lambda$ if $\alpha \ge 0$. This gives coerciveness eventually. By the usual extreme value theorem (c.f. Bertsekas (1999), pg. 669), $\widehat{\mathcal{Z}}_N(x) = \arg\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)$ and $\mathcal{Z}^*(x) = \arg\min_{z\in\mathcal{Z}} C(z|x)$ exist, are nonempty, and are compact.

Now we show there exists $\mathcal{Z}^*_\infty(x)$ compact such that $\mathcal{Z}^*(x) \subset \mathcal{Z}^*_\infty(x)$ and $\widehat{\mathcal{Z}}_N(x) \subset \mathcal{Z}^*_\infty(x)$ eventually. If $\mathcal{Z}$ is bounded this is trivial. So suppose otherwise (and again, without loss of generality D_x is compact). Fix any $z^* \in \mathcal{Z}^*(x)$. Then by Lemma 4 we have $\widehat{C}_N(z^*|x) \to C(z^*|x)$. Since $\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \le \widehat{C}_N(z^*|x)$, we have $\limsup_{N\to\infty} \min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \le C(z^*|x) = \min_{z\in\mathcal{Z}} C(z|x) = v^*$. Now suppose for contradiction no such $\mathcal{Z}^*_\infty(x)$ exists. Then there must be a subsequence $z_{N_k} \in \widehat{\mathcal{Z}}_{N_k}$ such that $\left|\left|z_{N_k}\right|\right| \to \infty$. By D_x-uniform coerciveness and $\left|\left|z_{N_k}\right|\right| \to \infty$, $\exists k_1 \ge k_0$ such that $c(z_{N_k};y) \ge 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x) \ \forall k \ge k_1$ and $y \in D_x$. Hence, $\forall k \ge k_1$ and $N \ge N_0$,

$$\widehat{C}_N(z_{N_k}|x) \ge \hat{\mu}_{Y|x,N}(D) \times 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x) + (1 - \hat{\mu}_{Y|x,N}(D)) \ge v^* + 1.$$

This yields a contradiction $v^* + 1 \le v^*$. So $\mathcal{Z}^*_\infty(x)$ exists.

Applying Lemma 4,

$$\tau_N = \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| \to 0.$$

The first result follows from

$$\delta_N = \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \leq \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| = \tau_N \to 0.$$

Now consider any sequence $z_N \in \widehat{\mathcal{Z}}_N(x)$. The second result follows from

$$\left|C\left(\hat{z}_N|x\right) - \min_{z\in\mathcal{Z}} C(z|x)\right| \leq \left|\widehat{C}_N(\hat{z}_N(x)|x) - C\left(\hat{z}_N|x\right)\right| + \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \leq \tau_N + \delta_N \to 0.$$

Since $\widehat{\mathcal{Z}}_N(x) \subset \mathcal{Z}^*_\infty(x)$ and $\mathcal{Z}^*_\infty(x)$ is compact, z_N has at least one convergence subsequence. Let $z_{N_k} \to z$ be any convergent subsequence. Suppose for contradiction that $z \notin \mathcal{Z}^*(x)$, i.e., $\epsilon = C(z|x) - v^* > 0$. Since $z_{N_k} \to z$ and by equicontinuinty, $\exists k_2$ such that $\left|c(z_{N_k};y) - c(z;y)\right| \leq \epsilon/2 \ \forall y \in \mathcal{Y} \ \forall k \geq k_2$. Then, $\left|C(z_{N_k}|x) - C(z|x)\right| \leq \mathbb{E}\left[\left|c(z_{N_k};y) - c(z;y)\right| \middle| X = x\right] \leq \epsilon/4 \ \forall k \geq k_2$. In addition, there exists k_3 such that $\delta_{N_k} \leq \epsilon/4 \ \forall k \geq k_3$. Then, $\forall k \geq \max\{k_2, k_3\}$, we have

$$\min_{z\in\mathcal{Z}} \widehat{C}_{N_k}(z|x) = \widehat{C}_{N_k}(z_{N_k}|x) \geq C(z_{N_k}|x) - \epsilon/4 \geq C(z|x) - \epsilon/2 \geq v^* + \epsilon/2.$$

Taking limits, we derive a contradiction, yielding the third result.

Now suppose that case 2 of Assumption 3 holds (i.e. convexity). By Lemma 4, $\widehat{C}_N(z|x) \to C(z|x)$ uniformly in z over any compact subset of $\mathcal{Z}$. By Theorem 6.2 of Rockafellar (1997), closed convex $\mathcal{Z}$ has a non-empty relative interior. Let us restrict to its affine hull where it has a non-empty interior. We have already shown that Assumption 2 implies that $\widehat{C}_N(z|x)$ and $C(z|x)$ are continuous. Hence they are lower semi-continuous. Therefore, by Theorem 7.17 of Rockafellar and Wets (1998), $\widehat{C}_N(z|x)$ epi-converges to $C(z|x)$ on $\mathcal{Z}$. Consider any $z^* \in \mathcal{Z}^*(x) \neq \varnothing$. Then clearly $\min_{z\in\{z^*\}} \widehat{C}_N(z|x) = \widehat{C}_N(z^*|x) \to C(z^*|x) = \min_{z\in\mathcal{Z}} C(z|x)$ and $\{z^*\}$ is compact. By Theorem 7.31 of Rockafellar and Wets (1998) we have precisely the results desired. □

Lemma 6. *Suppose $c(z;y)$ is equicontinuous in z. Suppose moreover that for each fixed $z \in \mathcal{Z} \subset \mathbb{R}^d$ we have that $\widehat{C}_N(z|x) \to C(z|x)$ a.s. for μ_X-a.e.x and that for each fixed measurable $D \subset \mathcal{Y}$ we have that $\hat{\mu}_{Y|x,N}(D) \to \mu_{Y|x}(D)$ a.s. for μ_X-a.e.x. Then, a.s. for μ_X-a.e.x, $\widehat{C}_N(z|x) \to C(z|x)$ for all $z \in \mathcal{Z}$ and $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly.*

Proof Since Euclidean space is separable, $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly a.s. for μ_X-a.e.x (c.f. Theorem 11.4.1 of Dudley (2002)). Consider the set $\mathcal{Z}' = \mathcal{Z} \cap \mathbb{Q}^d \cup \{\text{the isolated points of } \mathcal{Z}\}$. Then $\mathcal{Z}'$ is countable and dense in $\mathcal{Z}$. Since $\mathcal{Z}'$ is countable, by continuity of measure, a.s. for μ_X-a.e.x, $\widehat{C}_N(z'|x) \to C(z'|x)$ for all $z' \in \mathcal{Z}'$. Restrict to a sample path and x where this event occurs. Consider any $z \in \mathcal{Z}$ and $\epsilon > 0$. By equicontinuity $\exists \delta > 0$ such that $|c(z;y) - c(z';y)| \leq \epsilon/2$ whenever $||z - z'|| \leq \delta$. By density there exists such $z' \in \mathcal{Z}'$. Then, $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \leq \mathbb{E}_{\hat{\mu}_{Y|x,N}}\left[|c(z;y) - c(z';y)|\right] \leq \epsilon/2$ and $|C(z|x) - C(z'|x)| \leq \mathbb{E}\left[|c(z;y) - c(z';y)| \middle| X = x\right] \leq \epsilon/2$. Therefore, $0 \leq \left|\widehat{C}_N(z|x) - C(z|x)\right| \leq \left|\widehat{C}_N(z'|x) - C(z'|x)\right| + \epsilon \to \epsilon$. Since true for each ϵ, the result follows for all $z \in \mathcal{Z}$. The choice of particular sample path and x constitute a measure-1 event by assumption. □

Now, we prove the general form of the asymptotic results from Section 8.2.

Proof of Theorem 9 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 1, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 5 of Walk (2010) to Y'. By iid sampling and choice of k, we have that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0,1]$. Therefore applying Theorem 5 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma 6 we obtain that assumptions for Lemma 5 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem 10 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 1, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 3 of Walk (2010) to Y'. By assumption in theorem statement, we also have that $\mathbb{E}\{|Y'| \max\{\log|Y'|, 0\}\} < \infty$. Moreover each of the kernels in (14) can be rewritten $K(x) = H(||x||)$ such that $H(0) > 0$ and $\lim_{t\to\infty} t^{d_X} H(t) \to 0$.

Consider the case of iid sampling. Then our data on (X, Y') is ρ-mixing with $\rho(k) = 0$. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Consider the case of ρ-mixing or α-mixing. By Lemma 3, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and $\mathbb{E}\{|Y'| \max\{\log|Y'|, 0\}\} \le 1 < \infty$. Therefore applying Theorem 3 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma 6 we obtain that assumptions for Lemma 5 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem 11 Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 1, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 4 of Walk (2010) to Y'. Note that the naïve kernel satisfies the necessary conditions.

Since our data on (X, Y) is ρ-mixing by assumption, we have that by Lemma 3, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choice of the naïve kernel and h_N, Theorem 4 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability. Therefore applying Theorem 4 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma 6 we obtain that assumptions for Lemma 5 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem 12 Fix $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Set $Y' = c(z;Y)$. By Assumption 1, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 11 of Hansen (2008) to Y' and use the notation thereof. Fix the neighborhood of consideration to the point x (i.e., set $c_N = 0$) since uniformity in x is not of interest. All of the kernels in (14) are bounded above and square integrable and therefore satisfy Assumption 1 of Hansen (2008). Let f be the density of X. By assumption $0 < \delta \le f(x) \le B_0 < \infty$ for all $x \in \mathcal{X}$. Moreover, our choice of h_N satisfies $h_N \to 0$.

Consider first the iid case. Then we have $\alpha(k) = 0 = O(k^{-\gamma})$ for $\gamma = \infty$ (β in Hansen (2008)). Combined with boundedness conditions of Y' and f ($|Y'| \le g(z) < \infty$ and $\delta < f < B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $\gamma = \infty$, $s = \infty$ in (17) of Hansen (2008) we get $\theta = 1$. Therefore, since $h = O(N^{-1/d_x})$ we have

$$\frac{(\log\log N)^4 (\log N)^2}{N^\theta h_N^{d_x}} \to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ a.s.

Now consider the α-mixing case. If the mixing conditions hold for X, Y then by Lemma 3, equal or lower mixing coefficients hold for X, Y'. By letting $s = \infty$ we have $\gamma > d_x + 3 > 2$. Combined with boundedness conditions of Y' and f ($|Y'| \le g(z) < \infty$ and $\delta < f < B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $q = \infty$, $s = \infty$ in (16) and (17) of Hansen (2008) we get $\theta = \frac{\gamma - d_x - 3}{\gamma - d_x + 3}$. Therefore, since $h_N = O(N^{-\theta/d_x})$ we have

$$\frac{(\log\log N)^4 (\log N)^2}{N^\theta h_N^{d_x}} \to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have again that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ a.s.

Since $x \in \mathcal{X}$ was arbitrary we have convergence for μ_X-a.e.x a.s.

Now fix D measurable. Consider a response variable $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0,1]$. In addition, by Lemma 3, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Therefore applying Theorem 11 of Hansen (2008) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma 6 we obtain that assumptions for Lemma 5 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem 6 By assumption of Y and V sharing no atoms, $\delta \overset{a.s.}{=} \tilde{\delta} = \mathbb{I}[Y \le V]$ is observable so let us replace δ^i by $\tilde{\delta}^i$ in (21). Let

$$F(y|x) = \mathbb{E}\left[\mathbb{I}[Y > y] \,\middle|\, X = x\right]$$

$$\hat{F}_N(y|x) = \sum_{i=1}^{N} \mathbb{I}\left[u^i > y\right] w_{N,i}^{\text{Kaplan-Meier}}(x),$$
$$H_1(y|x) = \mathbb{E}\left[\mathbb{I}\left[U > y, \tilde{\delta} = 1\right] \middle| X = x\right]$$
$$\hat{H}_{1,N}(y|x) = \sum_{i=1}^{N} \mathbb{I}\left[u^i > y,\, \tilde{\delta}^i = 1\right] w_{N,i}(x),$$
$$H_2(y|x) = \mathbb{E}\left[\mathbb{I}\left[U > y\right] \middle| X = x\right]$$
$$\hat{H}_{2,N}(y|x) = \sum_{i=1}^{N} \mathbb{I}\left[u^i > y\right] w_{N,i}(x).$$

By assumption on conditional supports of Y and V, $\sup\{y : F(y:x) > 0\} \leq \sup\{y : H_2(y:x) > 0\}$. By the same arguments as in Theorem 2, 3, 4, or 5, we have that, for all y, $\hat{H}_{1,N}(y|x) \to H_1(y|x)$, $\hat{H}_{2,N}(y|x) \to H_2(y|x)$ a.s. for μ_X-a.e.x. By assumption of conditional independence and by the main result of Beran (1981), we have that, for all y, $F_N(y|x) \to F(y|x)$ a.s. for μ_X-a.e.x. Since $\mathcal{Y}$ is a separable space we can bring the "for all y" inside the statement, i.e., we have weak convergence (c.f. Theorem 11.4.1 of Dudley (2002)): $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ a.s. for μ_X-a.e.x where $\hat{\mu}_{Y|x,N}$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Since costs are bounded, the portmanteau lemma (see Theorem 2.1 of Billingsley (1999)) gives that for each $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to \mathbb{E}[c(z;Y)|X = x]$ where $\widehat{C}_N(z|x)$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Applying Lemma 6 we obtain that assumptions for Lemma 5 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

9. Extensions of Out-of-Sample Guarantees to Mixing Processes and Proofs

First we establish a comparison lemma that is an extension of Theorem 4.12 of Ledoux and Talagrand (1991) to our multivariate case.

LEMMA 7. *Suppose that c is L-Lipschitz uniformly over y with respect to ∞-norm:*

$$\sup_{z \neq z' \in \mathcal{Z},\, y \in \mathcal{Y}} \frac{c(z;y) - c(z';y)}{\max_{k=1,\ldots,d} |z_k - z'_k|} \leq L < \infty.$$

Let $\mathcal{G} = \{(x,y) \mapsto c\,(f(x);y) : f \in \mathcal{F}\}$. Then we have that $\widehat{\mathfrak{R}}_n(\mathcal{G}; S_N) \leq L\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N^x)$ and therefore also that $\mathfrak{R}_n(\mathcal{G}) \leq L\mathfrak{R}_n(\mathcal{F})$. (Notice that one is a univariate complexity and one multivariate and that the complexity of $\mathcal{F}$ involves only the sampling of x.)

Proof Write $\phi_i(z) = c(z;y^i)/L$. Then by Lipschitz assumption and by part 2 of Proposition 2.2.1 from Bertsekas et al. (2003), for each i, ϕ_i is 1-Lipchitz. We now would like to show the inequality in

$$\widehat{\mathfrak{R}}_n(\mathcal{G}; S_N) = L\mathbb{E}\left[\frac{2}{n} \sup_{z \in \mathcal{F}} \sum_{i=1}^{n} \sigma_{i0} \phi_i(z(x^i)) \,\middle|\, S_N\right]$$

$$\leq L\mathbb{E}\left[\frac{2}{n}\sup_{z\in\mathcal{F}}\sum_{i=1}^{n}\sum_{k=1}^{d}\sigma_{ik}z_k(x^i)\,\middle|\,S_N^x\right]$$
$$= L\widehat{\mathfrak{R}}_n(\mathcal{F};S_N^x).$$

By conditioning and iterating, it suffices to show that for any $T\subset\mathbb{R}\times\mathcal{Z}$ and 1-Lipchitz ϕ,

$$\mathbb{E}\left[\sup_{t,z\in T}\left(t+\sigma_0\phi(z)\right)\right]\leq\mathbb{E}\left[\sup_{t,z\in T}\left(t+\sum_{k=1}^{d}\sigma_k z_k\right)\right]. \tag{26}$$

The expectation on the left-hand-side is over two values ($\sigma_0=\pm1$) so there are two choices of (t,z), one for each scenario. Let any $(t^{(+1)},z^{(+1)}),(t^{(-1)},z^{(-1)})\in T$ be given. Let k^* and $s^*=\pm1$ be such that

$$\max_{k=1,\ldots,d}\left|z_k^{(+1)}-z_k^{(-1)}\right|=s^*\left(z_{k^*}^{(+1)}-z_{k^*}^{(-1)}\right).$$

Fix $(\tilde{t}^{(\pm1)},\tilde{z}^{(\pm1)})=(t^{(\pm s^*)},z^{(\pm s^*)})$. Then, since these are feasible choices in the inner supremum, choosing $(t,z)(\sigma)=(\tilde{t}^{(\sigma_{k^*})},\tilde{z}^{(\sigma_{k^*})})$, we see that the right-hand-side of (26) has

$$\begin{aligned}\text{RHS (26)}\geq{}&\frac{1}{2}\mathbb{E}\left[\tilde{t}^{(+1)}+\tilde{z}_{k^*}^{(+1)}+\sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(+1)}\right]\\&+\frac{1}{2}\mathbb{E}\left[\tilde{t}^{(-1)}-\tilde{z}_{k^*}^{(-1)}+\sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(-1)}\right]\\={}&\frac{1}{2}\left(t^{(+1)}+t^{(-1)}+\max_{k=1,\ldots,d}\left|z_k^{(+1)}-z_k^{(-1)}\right|\right)\\\geq{}&\frac{1}{2}\left(t^{(+1)}+\phi\left(z^{(+1)}\right)\right)+\frac{1}{2}\left(t^{(-1)}-\phi\left(z^{(-1)}\right)\right)\end{aligned}$$

where the last inequality is due to the Lipschitz condition. Since true for any $(t^{(\pm1)},z^{(\pm1)})$ given, taking suprema over the left-hand-side completes the proof. □

Next, the theorem below is a combination and restatement of the main results of Bartlett and Mendelson (2003) (for iid) and Mohri and Rostamizadeh (2008) (for mixing) about univariate Rademacher complexities. These are mostly direct result of McDiarmid's inequality.

THEOREM 13. *Consider a class $\mathcal{G}$ of functions $\mathcal{U}\to\mathbb{R}$ that are bounded: $|g(u)|\leq\overline{g}\ \forall g\in\mathcal{G}$, $u\in\mathcal{U}$. Consider a sample $S_n=(u^1,\ldots,u^N)$ of some random variable $T\in\mathcal{T}$. Fix $\delta>0$. If S_N is generated by IID sampling, let $\delta'=\delta''=\delta$ and $\nu=N$. If S_N comes from a β-mixing process, fix some t,ν such that $2t\nu=N$, let $\delta'=\delta/2-(\nu-1)\beta(t)$ and $\delta''=\delta/2-2(\nu-1)\beta(t)$. Then (only for $\delta'>0$ or $\delta''>0$ where they appear), we have that with probability $1-\delta$,*

$$\mathbb{E}\left[g(T)\right]\leq\frac{1}{N}\sum_{i=1}^{N}g(u^i)+\overline{g}\sqrt{\log(1/\delta')/2\nu}+\mathfrak{R}_\nu(\mathcal{G})\qquad\forall g\in\mathcal{G}, \tag{27}$$

and that, again, with probability $1-\delta$,

$$\mathbb{E}\left[g(T)\right]\leq\frac{1}{N}\sum_{i=1}^{N}g(u^i)+3\overline{g}\sqrt{\log(2/\delta')/2\nu}+\widehat{\mathfrak{R}}_\nu(\mathcal{G})\qquad\forall g\in\mathcal{G}. \tag{28}$$

Now, we can prove Theorem 8 and extend it to the case of data generated by a mixing process.

Proof of Theorem 8 Apply Theorem 13 to the random variable $U=(X,Y)$ and function class $\mathcal{G}=\{(x,y)\mapsto c\left(f(x);y\right) \,:\, f\in\mathcal{F}\}$. Note that by assumption we have boundedness of functions in $\mathcal{G}$ by the constant $\bar{c}$. Bound the complexity of $\mathcal{G}$ by that of $\mathcal{F}$ using Lemma 7 and the assumption of $c(z;y)$ being L-Lipschitz. Equations (27) and (28) hold for every $g\in\mathcal{G}$ and hence for every $f\in\mathcal{F}$ and $g(x,y)=c\left(f(x);y\right)$, of which the expectation is the expected costs of the decision rule f. □

Next, we prove our bounds on the complexities of the function classes we consider.

Proof of Lemma 1 Consider $\mathcal{F}_k=\{z_k(\cdot):z\in\mathcal{F}\}=\{z_k(x)=w^Tx:||w||_p\leq\frac{R}{\gamma_k}\}$, the projection of $\mathcal{F}$ onto the k^{th} coordinate. Then $\mathcal{F}\subset\mathcal{F}_1\times\cdots\times\mathcal{F}_{d_z}$ and $\mathfrak{R}_N(\mathcal{F})\leq\sum_{k=1}^{d_z}\mathfrak{R}_N(\mathcal{F}_k)$. The latter right-hand-side complexities are the common univariate Rademacher complexities. Applying Theorem 1 of Kakade et al. (2008) to each component we get $\mathfrak{R}_N(\mathcal{F}_k)\leq 2M\sqrt{\frac{p-1}{N}}\frac{R}{\gamma_k}$. □

Proof of Lemma 2 Let q be p's conjugate exponent ($1/p+1/q=1$). In terms of vector norms on $v\in\mathbb{R}^d$, if $q\geq 2$ then $||v||_p\leq||v||_2$ and if $q\leq 2$ then $||b||_p\leq d^{1/2-1/p}||v||_2$. Let F be the matrix $F_{ji}=x_j^i$. Note that $F\sigma\in\mathbb{R}^{d_x\times d_z}$. By Jensen's inequality and since Schatten norms are vector norms on singular values,

$$\begin{aligned}
\widehat{\mathfrak{R}}_N^2(\mathcal{F};S_N^x) &\leq \frac{4}{N^2}\mathbb{E}\left[\sup_{||W||_p\leq R}\text{Trace}\left(WF\sigma\right)^2\,\middle|\,S_N^x\right]\\
&= \frac{4R^2}{N^2}\mathbb{E}\left[||F\sigma||_q^2\middle|S_N^x\right]\\
&\leq \frac{4R^2}{N^2}\max\left\{\min\left\{d_z,d_x\right\}^{1-2/p},1\right\}\mathbb{E}\left[||F\sigma||_2^2\middle|S_N^x\right]\\
&\leq \frac{4R^2}{N^2}\max\left\{d_z^{1-2/p},1\right\}\mathbb{E}\left[||F\sigma||_2^2\middle|S_N^x\right].
\end{aligned}$$

The first result follows because

$$\begin{aligned}
\frac{1}{N}\mathbb{E}\left[||F\sigma||_2^2\,\middle|S_N^x\right] &= \frac{1}{n}\sum_{k=1}^{d_z}\sum_{j=1}^{d_x}\sum_{i,i'=1}^{N}x_j^ix_j^{i'}\mathbb{E}\left[\sigma_{ik}\sigma_{i'k}\right]\\
&= \frac{d_z}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_x}(x_j^i)^2 = d_z\,\widehat{\mathbb{E}}_N\,||x||_2^2
\end{aligned}$$

The second result follows by applying Jensen's inequality again to pass the expectation over S_n into the square. □

10. Proofs of Tractability Results

Proof of Theorem 1 Let $I=\{i:w_{N,i}(x)>0\}$, $w=(w_{N,i}(x))_{i\in I}$. Rewrite (3) as $\min w^T\theta$ over $(z,\theta)\in\mathbb{R}^{d\times n_0}$ subject to $z\in\mathcal{Z}$ and $\theta_i\geq c(z;y^i)\ \forall i\in I$. Weak optimization of a linear objective over a closed convex body is reducible to weak separation via the ellipsoid algorithm (see Grotschel et al. (1993)). A weak separation oracle for $\mathcal{Z}$ is assumed given. To separate over the i^{th} cost constraint

at fixed z', θ_i' call the evaluation oracle to check violation and if violated call the subgradient oracle to get $s \in \partial_z c(z'; y^i)$ with $||s||_\infty \leq 1$ and produce the cut $\theta_i \geq c(z'; y^i) + s^T(z - z')$. □

Proof of Theorem 7 In the case of (22), $z(x^i) = Wx^i$. By computing the norm of W we have a trivial weak membership algorithm for the norm constraint and hence by Theorems 4.3.2 and 4.4.4 of Grotschel et al. (1993) we have a weak separation algorithm. By adding affine constraints $\zeta_{ij} = z_j(x^i)$, all that is left is to separate over constraints of the form $\theta_i \geq c(\zeta_i; y^i)$, which can be done as in the proof of Theorem 1. □

11. Omitted Details from Section 1.1

11.1. Portfolio Allocation Example

In our portfolio allocation example, we consider constructing a portfolio with $d_y = d_z = 12$ securities. We simulate the observation of $d_x = 3$ market factors X that, instead of iid, evolve as a 3-dimensioanl ARMA(2,2) process:

$$X(t) - \Phi_1 X(t-1) - \Phi_2 X(t-2) = U(t) + \Theta_1 U(t-1) + \Theta_2 U(t-2)$$

where $U \sim \mathcal{N}(0, \Sigma_U)$ are innovations and

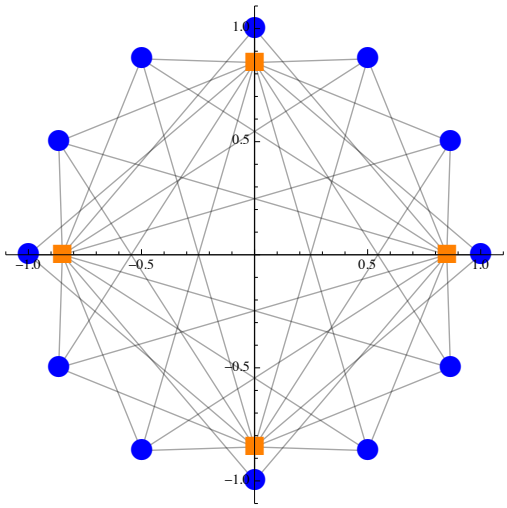
$$(\Sigma_U)_{ij} = \left(\mathbb{I}[i=j]\frac{8}{7} - (-1)^{i+j}\frac{1}{7}\right) 0.05,$$

$$\Phi_1 = \begin{pmatrix} 0.5 & -0.9 & 0 \\ 1.1 & -0.7 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0. & -0.5 & 0 \\ -0.5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\Theta_1 = \begin{pmatrix} 0.4 & 0.8 & 0 \\ -1.1 & -0.3 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Theta_2 = \begin{pmatrix} 0 & -0.8 & 0 \\ -1.1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We suppose the returns are generated according to a factor model $Y_i = A_i^T (X + \delta_i/4) + (B_i^T X)\epsilon_i$, where A_i is the mean-dependence of the i^{th} security on these factors with some idiosyncratic noise, B_i the variance-dependence, and ϵ_i and δ_i are independent standard Gaussian idiosyncratic contributions. For A and B we use

$$A = 2.5\% \times \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \quad B = 7.5\% \times \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$



(a) The network of warehouses (orange) and locations (blue).

$$D^T = \begin{pmatrix} 0.15 & 1.3124 & 1.85 & 1.3124 \\ 0.50026 & 0.93408 & 1.7874 & 1.6039 \\ 0.93408 & 0.50026 & 1.6039 & 1.7874 \\ 1.3124 & 0.15 & 1.3124 & 1.85 \\ 1.6039 & 0.50026 & 0.93408 & 1.7874 \\ 1.7874 & 0.93408 & 0.50026 & 1.6039 \\ 1.85 & 1.3124 & 0.15 & 1.3124 \\ 1.7874 & 1.6039 & 0.50026 & 0.93408 \\ 1.6039 & 1.7874 & 0.93408 & 0.50026 \\ 1.3124 & 1.85 & 1.3124 & 0.15 \\ 0.93408 & 1.7874 & 1.6039 & 0.50026 \\ 0.50026 & 1.6039 & 1.7874 & 0.93408 \end{pmatrix}$$

(b) The distance matrix.

Figure 17 Network data for shipment planning example.

Marginally, all of the returns have mean 0% and standard deviations 20∼30%.

In our objective, we use $\lambda = 0$ and $\epsilon = 0.15$, i.e., we minimize the conditional value at risk at level 15%.

11.2. Shipment Planning Example

In our shipment planning example, we consider stocking $d_z = 4$ warehouses to serve $d_y = 12$ locations. We take locations spaced evenly on the 2-dimensional unit circle and warehouses spaced evenly on the circle of radius 0.85. The resulting network and its associated distance matrix are shown in Figure 17. We suppose shipping costs from warehouse i to location j are $c_{ij} = \$10 D_{ij}$ and that production costs are \$5 per unit when done in advance and \$100 per unit when done last minute.

We consider observing $d_x = 3$ demand-predictive features X. We simulate X in the same manner as in the portfolio allocation example. We simulate demands as

$$Y_i = 100 \max\{0,\, A_i^T (X + \delta_i/4) + \left(B_i^T X\right) \epsilon_i\}$$

with A, B, δ, ϵ as in the portfolio allocation example.