·专题一:双清论坛“大规模商务场景的统计管理论”·

预测驱动最优化:不确定性、统计理论与管理应用*

王曙明[1**] 毛宇晨[1] 汪寿阳[1,2,3,4**]

1. 中国科学院大学 经济与管理学院,北京 100190
2. 中国科学院 数学与系统科学研究院,北京 100190
3. 中国科学院 预测科学研究中心,北京 100190
4. 上海科技大学 创业与管理学院,上海 201210

[摘 要] 现代管理决策面临着错综复杂的不确定性。随着大数据应用的不断深化,最优化技术与算力的持续提升,以及统计学与机器学习的蓬勃发展,预测驱动最优化正在成为应对复杂不确定决策问题的有力工具。预测驱动最优化通过整合统计预测建模与决策最优化,实现对不确定性与决策效能的联合统计管理,从而形成统计一致、有效的数据驱动型决策范式。本文聚焦于不确定环境下的统计预测建模与管理决策最优化,分别探讨分布已知(随机型)与分布未知(分布鲁棒型)条件下的预测驱动最优化模型框架。在此基础上,介绍预测驱动最优化在运营管理领域的前沿应用研究,并总结若干重要的未来研究方向与挑战。

[关键词] 预测驱动最优化;不确定性;统计理论;管理应用

随着全球化进程的不断加快、技术与市场的不断迭代更新,以及环境与气候变化影响的日益显著,现代供应链、产业链网络以及社会网络日趋错综复杂,这些因素促成了现代企业与组织越发复杂与多样化的不确定性决策环境[1-3]。

与此同时,随着数据及技术环境的不断进阶演化,现代管理决策正面临以下深刻挑战与创新机遇。(1) 大数据应用的不断深化。随着移动互联技术的广泛普及、数据采集、存储以及数字化技术的飞速发展,具有超大规模、超高维度、多源异构、时空关联、流式产生特点的大数据日益可测可获[4]。在大数据背景下,经济社会、政府决策以及商业运营等管理活动呈现出高频实时、深度定制化、全周期沉浸式交互、跨组织数据整合、多主体协同等新特征[5]。基于大数据的管理决策正逐渐成为科学研究与产业应用的主流决策范式[6]。(2) 最优化技术与算力的持续提升。复杂的决策问题往往可以建模为数学优化问题,而现代最优化技术已经能够高效地求解线性规划、二阶锥规划、半定规划以及整数规划等典型优化问题。用于求解大规模复杂优化问题的数值优化算法也迎来突破性进展,包括内点法、信赖域算法、增广拉格朗日函数法、零阶优化、稀疏优化等[7,8]。同时,并行计算、分布式计算以及云计算等技术的发

汪寿阳 中国科学院特聘研究员,教育部长江学者奖励计划特聘教授,国家杰出青年科学基金获得者。现担任中国科学院数学与系统科学研究院研究员、中国科学院预测科学研究中心主任,兼任国际系统与控制科学院副院长。担任包括国际知名期刊 *Energy Economics* 等在内的 12 种学术期刊的主编、执行主编、副主编或编委。

王曙明 中国科学院大学经济与管理学院教授,博士生导师。主要从事随机鲁棒优化、统计学习以及模型不确定性的理论、方法与应用研究。承担国家自然科学基金优秀青年科学基金等多项科研项目。目前担任运筹学著名期刊 *Computers & Operations Research* 领域主编以及决策科学旗舰期刊 *Decision Sciences Journal* 副主编。

收稿日期:2023-12-29;修回日期:2024-05-17
* 本文根据国家自然科学基金委员会第 344 期“双清论坛”讨论的内容整理。
** 通信作者,Email: sywang@amss.ac.cn; wangshuming@ucas.edu.cn
本文受到国家自然科学基金项目(71922020, 72171221, 71988101)和中央高校基本科研业务费专项资金(UCAS-E2ET0808X2)的资助。

展，大幅提升了处理大规模数据与计算任务的效率[9]。进一步，各类优化建模方法，如随机优化、分布鲁棒优化与在线优化[10-12]等也全面扩展了最优化方法的管理应用场景。特别地，分布鲁棒优化已成为处理不确定环境下决策与最优化问题最重要的建模范式之一[11, 13]。最优化技术的不断发展也催生了众多高性能最优化求解器的涌现与迭代，如：Gurobi、CPLEX、COPT、CMIP 等，为现代管理决策提供了有力的软件系统支持。与此同时，以大模型为主的智算爆发引领了新一波算力发展浪潮，截至 2022 年底，全球算力总规模同比增长 24.8%[14]，算力赋能产业发展，互联网、大数据、人工智能与实体经济融合发展的新业态、新模式正加速涌现。最优化技术与算力已经成为大数据决策的基础支撑。(3) 统计学与机器学习的蓬勃发展。统计学的诸多重要理论与思想，如：反事实因果推断、超参数化模型与正则化以及贝叶斯多层模型等，都为机器学习等人工智能的发展奠定了坚实理论基础[15]。而在大数据与现代计算技术的推动下，统计学也在不断催生出新视角与新方法。另一方面，机器学习作为数据科学的重要领域，发展了一系列引领潮流的先进模型，如随机森林、生成对抗网络（Generative Adversarial Network，GAN）、深度残差网络（Residual Network，ResNet），以及大模型（如 Transformer、GPT-3、ChatGPT）等[16, 17]，并在智慧医疗、金融科技、自动驾驶、市场营销等领域取得了惊人的进展。特别地，现代计量经济学方法与机器学习也正在不断渗透融合[18]，并显现出独特优势。例如，改进传统计量经济学方法，加强因果推断[18]；从海量、复杂、高维、非结构化、多模态数据中提取有价值信息，改善变量测量[19]；直接基于数据灵活选择函数形式并构建算法，提高样本外事件预测的准确性[20]；以及发现稳健且具有解释性的模式，促进理论构建[21]等。统计与机器学习的进阶发展正在深刻影响并改变着预测与决策科学的研究范式与方法[18]。

大数据、最优化技术及算力、统计学与机器学习的发展为管理决策提供了要素、理论与技术支持，推动了现代管理决策向数据驱动范式不断演进，以应对日益复杂的决策环境。当前数据驱动型商业分析的重要组成部分是指导性分析（Prescriptive Analytics），其致力于将混合数据、预测模型、优化技术及业务规则进行协同，进而提供决策支持。指导性分析的本质就是预测驱动最优化，即决策者利用辅助信息（如协变量信息）来推断不确定参数的统计信息，进而进行决策。例如，股票回报可能依赖于历史价格以及社交媒体上反映市场情绪的文本[22]，基于这些信息，决策者能够更精准地预测未来股票回报并建立收益更高、风险更低的投资组合；零售商面对不确定的产品需求，可以根据预测的未来天气推断出需求高低，从而制定订货策略[23]；此外，天气状况和时刻可以帮助预测道路网络拥堵的不确定性，有助于找到具有最短时间的配送路径。随着全面跨入大数据时代，预测驱动最优化已经在交通、医疗保健、流程优化、供应链、电网与能源管理[24-26]等领域崭露头角，为应对各类复杂运营场景提供了灵活而有效的方法框架。

预测驱动最优化并不是预测模型在优化模型中的简单嵌入，而是两者的有机融合。构建预测驱动最优化模型的关键在于确保其具有良好的统计性质和高效的优化结构，这是一个复杂而系统性的挑战。实现这一目标既需要借鉴统计学、数据科学、决策学科与最优化方法的成熟理论，也需要在实际问题中持续探索，深刻理解在具体决策问题下预测驱动最优化模型的结构。为此，本文将重点基于不确定性的统计建模方式以及预测-决策协同机制探讨各类预测驱动型决策优化框架，并对不同框架的理论模型、统计性能保证以及运营管理影响进行分析。具体而言，本文首先介绍不确定性决策问题下主要的预测驱动最优化框架及其统计理论，进而介绍预测驱动最优化在具体运营管理领域如交通与物流管理、库存管理以及收益管理中的应用研究，最后展望预测驱动最优化的未来研究方向与挑战。

1　不确定性决策：预测驱动最优化框架与统计理论

考虑一类一般的不确定环境下的决策问题：决策者制定决策 z 以最小化具有不确定变量 y 的运营成本 $c(z,y)$。在制定决策 z 时，不确定变量 y 是未知的。决策者可以观测到不确定变量 y 以及与不确定变量 y 相关的协变量 x 的历史数据 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^n$。基于这些信息，决策者可以预测未来某种情景下不确定变量 y 的条件分布，从而针对未来可能情景做出更为精准的决策。具体地，给定一个由协变量 x 所描述的情景，决策者致力于最小化此情景下的条件期望运营成本：

$$\min_{z\in Z}\mathbb{E}_{y\sim\mathbb{P}_{y|x}}[c(z,y)], \tag{1}$$

其中 $\mathbb{P}_{y|x}$ 表示给定协变量 x 下不确定变量 y 的条件

分布。该类问题也被称为情景随机优化问题(Contextual Stochastic Optimization Problem)。在该问题中,协变量 x 可以是外生变量,也可以是不确定变量 y 自身的滞后项,甚至是截至决策时决策者所收集的所有信息。传统的无条件随机优化模型忽略了协变量信息,基于不确定变量的无条件分布进行决策[10]。这可能导致决策在样本外是次优(Suboptimal)甚至不可行(Infeasible)[27]。

由于预测误差以及未来环境变化等因素,决策者通常无法准确预测条件分布 $\mathbb{P}_{y|x}$,而是仅具有 $\mathbb{P}_{y|x}$ 的部分信息。为了对抗该分布不确定性(Ambiguity),给定协变量 x,决策者构造可能包含条件分布$\mathbb{P}_{y|x}$的集合$\mathcal{A}(x)$,并求解以下 min-max 问题,以最小化可能发生的最坏情况分布下的条件期望成本:

$$\inf_{z\in Z}\sup_{\mathbf{P}_{y|x}\in\mathcal{A}(x)}\mathbb{E}_{y\sim\mathbf{P}_{y|x}}[c(z,y)]. \tag{2}$$

该问题也被称为情景分布鲁棒优化(Contextual Distributionally Robust Optimization)问题。

1.1 随机预测驱动决策框架

本小节介绍近似求解情景随机优化问题[28,29](1)的三个典型的决策框架:“先预测后优化”(Estimate-then-optimize)、“联合预测优化”(Integrated-estimation-and-optimization)以及“端到端优化”(End-to-end optimization)。三者都由数据所驱动,“先预测后优化”框架先构建预测模型,再进行决策优化,预测模型不受决策模型影响;“联合预测优化”框架在构建预测模型时考虑决策模型的影响;“端到端优化”不对不确定变量直接进行预测,而是学习协变量 x 映射到决策 z 的决策规则。通过对比这三种框架,我们可以更清晰地了解它们各自的优势和局限性,为实际问题的求解提供更有针对性的指导。

在“先预测后优化”决策框架中,我们首先利用历史数据预测给定某一协变量 x 下的不确定变量 y,然后基于此预测进行优化决策。经典的统计与机器学习方法通常致力于估计不确定变量的条件期望 $\mathbb{E}[y|x]$。由于成本函数往往是高度非线性的,即有 $c(z,\mathbb{E}[y|x])\neq\mathbb{E}[c(z,y)\mid x]$。因此,直接将条件期望的点估计值代入成本函数再进行优化是不恰当的。更合适的方法是先构建条件分布 $\mathbb{P}_{y|x}$ 的预测分布 $\hat{\mathbb{P}}(x)$,再优化求解以下问题:

$$\min_{z\in Z}\mathbb{E}_{y\sim\hat{\mathbf{P}}(x)}[c(z,y)]. \tag{3}$$

如果 $\hat{\mathbb{P}}(x)$能够较准确地预测真实条件分布$\mathbb{P}_{y|x}$,上述问题(3)将生成目标问题(1)的近似最优策略。因此,构建预测分布 $\hat{\mathbb{P}}(x)$是先预测后优化决策框架的核心。

针对高维不确定变量 y 的预测分布 $\hat{\mathbb{P}}(x)$的构建,研究者们提出了两种主要的方法:基于残差的方法和基于权重的方法。这两类方法都使用离散分布来建模 $\hat{\mathbb{P}}(x)$,原因在于具有连续分布的随机优化问题常受困于“维度灾难”而难以求解[10]。第一种方法基于预测模型的残差来构建条件分布[30]。基于历史数据 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^n$,通过最小化预测损失,可以得到不确定变量 y 的条件期望的点估计函数 $\hat{f}(\cdot)$。相应地,样本 i 的残差为 $\hat{\varepsilon}_i=y_i-\hat{f}(x_i)$。基于残差可形成预测分布

$$\hat{\mathbb{P}}^{\text{res}}(x)=\frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{f}(x)+\hat{\varepsilon}_i},$$

其中 δ_{A} 为在 A 处概率为 1 的狄拉克(Dirac)分布。该方法的优势在于可与任何预测模型结合,且当预测模型具有一致性时,由该方法所得到的最优决策具有渐近最优性。另一种方法则为历史数据分配不同的权重来构造预测分布[31],即

$$\hat{\mathbb{P}}^{\text{wtd}}(x)=\sum_{i=1}^{n}w_i(x)\delta_{y_i}.$$

权重选择的核心思想是:对于新情景 x,越接近其的历史情景 x_i 所对应的历史不确定变量 y_i 应获得更大的权重。Bertsimas 和 Kallus[31]证明了使用 k 近邻、NW 核密度估计、局部线性回归、回归树以及随机森林等机器学习方法确定权重,能够保证所得最优决策的渐近最优性。在基于离散预测分布的方法中,预测模型的复杂性并不会增加决策问题求解的难度。这意味着,我们可以根据具体数据特征调整预测模型,以改善所得最优决策的统计性质(例如,加快收敛速率或减小方差),同时不影响优化性质。

在“先预测后优化”框架中,预测模型通过最小化预测损失函数获得,并不涉及后续的决策优化问题。然而,在某些情况下,误差较小的预测模型可能产生次优的决策。这是因为真实最优决策基于真实模型指定的方向进行优化;若预测模型所指引的优化方向偏离了正确方向,就可能导致次优决策。而预测模型的损失函数往往不对预测模型偏离真实模型的方向进行区分。

为了缓解上述问题,“联合预测优化”决策框架通过在训练预测模型时融入决策优化信息,以促使预测模型指定的方向与真实模型指定的方向趋于一致。该框架的提出可追溯至 Bengio 等[32]在投资组

合管理中的早期研究，研究发现在处理金融时间序列数据时，相较于最小化预测损失（例如均方误差），通过最大化金融准则（如总利润）预测股票收益能够带来更高的超额收益。Donti 等[33]将这一思想拓展至一般情景优化问题。他们使用连续参数分布（如指数分布或正态分布）$\mathbb{P}_\theta(x)$来建模不确定变量 y 的条件分布 $\mathbb{P}_{y|x}$。给定参数 θ，代入 $\mathbb{P}_\theta(x)$ 求解问题(1)可以得到决策 $z^*(x;\theta)\in\underset{z\in Z}{\operatorname{argmin}}\,\mathbb{E}_{y\sim\mathbb{P}_\theta(x)}[c(z,y)]$。此时，最优参数可通过最小化 $z^*(x;\theta)$ 在经验分布 $\mathbb{Q}_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}$ 上的平均成本得到，即：

$$L(\theta)=\mathbb{E}_{(x,y)\sim\mathbb{Q}_n}[c(z^*(x;\theta),y)]. \tag{4}$$

进一步，假设不确定变量 y 的支撑集有限，Grigas 等[34]在损失函数(4)的基础上学习条件分布 $\mathbb{P}_{y|x}$，并证明了该框架具有渐近最优性，且样本内最优值收敛于真实最优值的速率为 $\mathcal{O}(n^{-1/2})$。Elmachtoub 和 Grigas[35]从“遗憾”（Regret）的角度衡量由预测导致的决策偏差，并最小化以下 SPO（Smart “Predict, then Optimize”）损失函数来构建不确定变量的参数估计 $\hat{f}_\theta(x)$：

$$L_{\mathrm{SPO}}(\theta)=\mathbb{E}_{(x,y)\sim\mathbb{Q}_n}[c(z^*(\hat{f}_\theta(x)),y)-c(z^*(y),y)],$$

其中，$z^*(\cdot)\in\underset{z\in Z}{\operatorname{argmin}}c(z,\cdot)$。该损失函数度量了使用参数估计 $\hat{f}_\theta(x)$ 代替不确定变量 y 进行决策所造成的偏差。由于 $\mathbb{E}_{(x,y)\sim\mathbb{Q}_n}[c(z^*(y),y)]$ 为常数，SPO 损失函数 $L_{\mathrm{SPO}}(\theta)$ 等价于损失函数(4)。SPO 损失可以理解为使用成本函数度量决策 $z^*(\hat{f}_\theta(x))$ 与 $z^*(y)$ 之间的差异。基于相似性的思想，Kong 等[36]直接最小化二者之间的经验距离。此外，还有一些方法在传统的预测损失上加入决策误差正则项，以保证在单一优化问题中同时考虑预测精度和决策质量，如 Loke 等[37]。

然而，在很多情况下，此类复杂的损失函数是非凸、不可微甚至不连续的，因此其优化求解是一个严峻的挑战。一个可行的思路是给出这些损失函数的凸近似。例如，Elmachtoub 和 Grigas[35]提出使用 SPO 损失的凸包络作为近似损失，其次梯度具有解析表达式。后续研究[35,38]探讨了该凸包络损失相对于 SPO 损失的 Fisher 一致性、风险界限等统计性质。对于复杂的运营管理问题，这些损失函数及其近似函数的实际效果仍需大量的实践验证。

不同于以上两个框架需要构建不确定变量 y 的预测，“端到端”决策框架寻找一个由协变量直接映射到决策的决策规则函数，以最小化协变量与不确定变量的联合分布下的期望成本。该框架的优势在于，一旦获得决策规则函数的解，对于未来可能发生情景，只需将相应的协变量代入决策规则函数即可迅速获得相应的决策，而无需再求解新的优化问题。这一特性使得该框架特别适用于需要快速做出决策的场景。

然而，该框架的应用也存在一些潜在的挑战。首先，真实的协变量与不确定变量的联合分布是未知的，我们通常只能通过经验风险最小化（Empirical Risk Minimization，ERM）来求解经验分布下的最优策略函数。此时如何为历史数据中未出现的协变量选择恰当的决策成为一个关键问题。其次，为了限制搜索空间，通常假设决策规则函数属于某个由参数化函数（例如线性函数或决策树）构成的集合中，这个函数集合通常被称为假设类（Hypothesis Class）。假设类的指定对该框架至关重要。当假设类过于简单时，如线性策略，可能无法准确捕捉真实的决策与协变量之间的关系；而当假设类过于复杂时，如神经网络，可能会导致优化过程变得极其困难，并且大大降低决策的解释力。

针对线性决策规则，Ban 和 Rudin[27]对报童问题中线性订货策略的性能进行了研究。他们指出，在线性需求模型的情况下，ERM 的泛化误差为 $\mathcal{O}(\dim(x)\cdot n^{-1/2})$，其中 $\dim(x)$ 为协变量的维度。Bertsimas 和 Kallus[31]进一步指出，对于一般的优化问题，线性决策策略并非渐近最优。为了解决这一问题，Bertsimas 和 Koduri[39]将假设类指定为再生核希尔伯特空间（Reproducing Kernel Hilbert Space，RKHS）并证明了一般问题下 RKHS 决策策略的渐近最优性。在非线性决策规则方面，深度神经网络（Deep Neural Network，DNN）被应用于报童问题中学习订货策略[40]。Rychener 和 Sutter[41]表明，由随机梯度下降法训练得到的基于 DNN 的决策规则近似最小化贝叶斯后验损失。Zhang 等[42]引入了分段线性决策规则，并给出了无约束问题的有限样本风险界限以及有约束问题的渐近最优性。

前文所提及的情景也可以指决策者可利用的所有信息，例如历史数据。将历史数据看作随机变量，进而将从中学习到的决策规则视为一个统计量。基于此思想，Feng 和 Shanthikumar[43]提出了运营统计量（Operational Statistic）的概念。他们进一步阐明：应当基于成本函数的结构信息以及不确定变量分布的统计信息构建决策规则 $z(\cdot)$（运营统计量的具体函数形式）所属的假设类（记为 $\mathcal{Z}_{\mathrm{stat}}$），并优化

以下问题构建运营统计量：

$$z^*(\cdot)\in\underset{z(\cdot)\in\mathcal{Z}_{\text{stat}}}{\operatorname{argmin}}\mathbb{E}\left[c(z(Y),y)\right],$$

其中，Y 表示不确定变量 y 的历史数据。他们展示了在一维不确定变量的情况下，运营统计量相较于传统的最小二乘等统计量，具有更优的决策性能。

1.2 分布鲁棒预测驱动决策框架

虽然随机预测驱动决策方法具有渐近一致性等优良性质，但当协变量向量的维度较大时，这些方法会产生过于乐观的决策[44]。此外，随机预测驱动决策方法高度依赖于数据生成过程的“稳定”性而缺乏应对高度不确定环境的鲁棒性。为了获得更具鲁棒性的决策，学者们考虑将情景随机优化问题正则化[45]或将其拓展为分布鲁棒优化问题(2)。Kannan 等[46]考虑了以基于残差的预测分布为中心，分布间距离度量为 Wasserstein 距离 $d_W(\cdot)$ 的分布不确定集：

$$\mathcal{A}_\gamma^{\text{res}}(x)=\{\mathbb{P}\in\mathcal{P}(\mathbb{R}^{\dim(y)})\mid d_W(\mathbb{P},\hat{\mathbb{P}}^{\text{res}}(x))\leqslant\gamma\},$$

其中，$\mathcal{P}(\cdot)$ 表示定义在该支撑集上的所有概率分布。当数据生成过程以及成本函数满足某些条件时，他们证明了最小二乘等回归方法所构建的预测分布能够确保问题(2)具有渐近最优性。此外，他们展示了当不确定分布集的半径 γ 被设置为 $\mathcal{O}((n\cdot\log\alpha)^{-1/\dim(y)})$ 时，样本外成本超过样本内最坏情况分布下期望成本的概率，即“样本外失望度”(Out-of-sample Disappointment)，将小于等于任意预设的显著性水平 α，即有：

$$\mathbb{P}\{\mathbb{E}_{y\sim\mathbb{P}_{y|x}}[c(z^*,y)]>\sup_{\mathbf{P}\in\mathcal{A}_\gamma^{\text{wtd}}(x)}\mathbb{E}_{y|x\sim\mathbf{P}}[c(z^*,y)]\}\leqslant\alpha,$$

其中，$\mathbb{P}_{y|x}$ 为真实条件分布，z^* 为样本内最优决策：$z^*\in\inf_{z\in Z}\sup_{\mathbf{P}\in\mathcal{A}_\gamma^{\text{wtd}}(x)}\mathbb{E}_{y|x\sim\mathbf{P}}[c(z^*,y)]$。类似地，Bertsimas 和 Van Parys[44]考虑了以基于权重的预测分布为中心，分布间距离度量为相对熵(Relative Entropy)$d_E(\cdot)$ 的分布不确定集：

$$\mathcal{A}_\gamma^{\text{wtd}}(x)=\{\mathbb{P}\in\mathcal{P}(\mathbb{R}^{\dim(y)})\mid d_E(\mathbb{P},\hat{\mathbb{P}}^{\text{wtd}}(x))\leqslant\gamma\}.$$

鉴于测试数据可能非常稀少，他们利用自举法(Bootstrap)从训练集中生成数据，并提供了以自举数据作为代理测试数据的样本外失望性能保证。

针对时间序列数据，Hu 等[47]提出了一种嵌套分布不确定集(Nested Ambiguity Set)，其结构性地嵌入了一般多元时间序列预测模型，以保证鲁棒性的同时提升预测准确性。具体而言，他们构造了以下基于历史观测 $y_{[t-1]}=\{y_1,\cdots,y_{t-1}\}$ 条件下第 t 期不确定变量 y_t 的分布不确定集：

$$\mathcal{A}_\gamma^{t|y_{[t-1]}}=\left\{\mathbb{Q}_t\in\mathcal{P}(\mathbb{R}_+^{\dim(y)})\;\middle|\;\begin{array}{l}y_t\sim\mathbb{Q}_t,\ \varepsilon_t\sim\mathbb{P}_t\\ y_t=A_t^0+A_t^1y_1+\cdots+A_t^{t-1}y_{t-1}+\varepsilon_t\\ d_W(\mathbb{P}_t,\hat{\mathbb{P}}_t)\leqslant\gamma\end{array}\right\},$$

其中，不确定变量 y_t 与其滞后项 $y_{[t-1]}$ 的关系由一类一般多元时间序列模型所刻画，误差 ε_t 的分布位于以残差经验分布为中心、半径为 γ 的 Wasserstein 球内。进一步地，他们嵌套地构建了整个时间序列 $(y_1,\cdots,y_T)$ 的分布不确定集：

$$\mathcal{A}_\gamma^{\text{nstd}}=\left\{\mathbb{Q}\in\mathcal{P}(\mathbb{R}_+^{\dim(y)\times T})\;\middle|\;\begin{array}{ll}(y_1,\cdots,y_T)\sim\mathbb{Q}&\\ y_{t|y_{[t-1]}}\sim\mathbb{Q}_t&\forall t=1,\cdots,T\\ \mathbb{Q}_t\in\mathcal{A}_\gamma^{t|y_{[t-1]}})&\forall t=1,\cdots,T\end{array}\right\}.$$

该集合在模型设定正确且满足一定的正则条件的情况下，能够保证决策具有渐近最优性。此外，他们还探讨了在有限样本情况下，如何合适地设定参数 γ 以使得发生样本外失望的可能性较小。

Mao 等[48]采用似不相关回归(Seemingly Unrelated Regression，SUR)来捕捉高维不确定变量之间相关性(异方差性)以及决策依赖性。通过使用可行广义最小二乘估计该计量模型，他们进一步构建一个基于计量模型、经验协变量和残差的分布不确定集：

$$\mathcal{A}_\gamma^{\text{SUR}}(z)=\left\{\mathbb{Q}\in\mathcal{P}(R_+^{\dim(y)}\times R^{\dim(x)}\times R^{\dim(\varepsilon)})\;\middle|\;\begin{array}{l}(y,x,\varepsilon)\sim\mathbb{Q};\ (x,\varepsilon)\sim\mathbb{P}_{(x,\varepsilon)}^n\\ \mathbb{E}_{\mathbb{Q}}[\|y-\hat{f}^{\text{SUR}}(z,x,\varepsilon)\|]\leqslant\gamma\end{array}\right\},$$

其中，z 作为决策影响不确定变量 y，$\mathbb{P}_{(x,\varepsilon)}^n=\frac{1}{n}\sum_{i=1}^n\delta_{(x_i,\hat{\varepsilon}_i)}$ 为由样本内协变量与残差构成的经验分布，$\hat{f}^{\text{SUR}}$ 为估计的需求预测模型。约束条件 $\mathbb{E}_{\mathbb{Q}}[\|y-\hat{f}(z,x,\varepsilon)\|]\leqslant\gamma$ 实际上刻画了预测模型 $\hat{f}^{\text{SUR}}$ 的期望预测误差不超过容忍水平 γ。在他们所研究的问题中，Mao 等证明了最坏情况分布

下期望成本等于基于预测经验分布 $\hat{\mathbb{P}}^{\mathrm{SUR}}(z)=\frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{f}^{\mathrm{SUR}}(z,x_i,\hat{\varepsilon}_i)}$ 的期望成本与某一正则化函数 $\phi(z)$ 之和，即：

$$\sup_{\mathbf{P}\in\mathcal{A}_{\gamma}^{\mathrm{SUR}}(z)}\mathbb{E}_{y|z\sim\mathbf{P}}[c(z,y)]=\mathbb{E}_{y|z\sim\hat{\mathbf{P}}^{\mathrm{SUR}}(z)}[c(z,y)]+\gamma\phi(z).$$

基于此正则化结构，他们证明了所提出的分布鲁棒优化模型具有渐近最优性以及样本外失望的有限样本性能保证。

区别于最小化最坏情况分布下的期望成本，Sim 等[49]提出决策者也可以在达到一定目标成本的情况下最大化模型的鲁棒性。他们将此框架中集成预测分布 $\hat{\mathbb{P}}(x)$，并构建了如下的优化问题以考虑预测分布的分布不确定性：

$$\begin{aligned}\min\quad&\kappa\\ \text{s.t.}\quad&\mathbb{E}_{y\sim\mathbf{P}}[c(z,y)]-\tau\leqslant\kappa\cdot d(\mathbb{P},\hat{\mathbb{P}}(x)),\quad\forall\,\mathbb{P}\in\mathcal{P}(\mathcal{Y})\\ &z\in\mathcal{Z},\kappa\geqslant 0,\end{aligned}\tag{5}$$

其中 τ 为决策者制定的目标成本，κ 表示系统的脆弱性，$d(\mathbb{P},\hat{\mathbb{P}}(x))$ 为预测分布与经验分布之间的距离度量。该模型还可以拓展考虑模型（系数）的不确定性。在鞍目标函数、两阶段线性优化问题以及决策依赖预测等场景，问题(5)具有可求解重构。

在情景分布鲁棒优化问题(2)中，若条件分布 $\mathbb{P}_{y|x}$ 落在了分布不确定集 $\mathcal{A}(x)$ 之内，则可以使用分布鲁棒优化方法来对冲概率分布的不确定性；反之，若真实分布落在给定的分布不确定集之外，即发生了模型误设（Model Misspecification），则分布鲁棒优化模型在真实概率分布下的决策性能可能无法得到保证。为了解决该问题，Liu 等[50]提出全局分布鲁棒优化模型（Globalized Distributionally Robust Optimization）：

$$\begin{aligned}\min\quad&\tau\\ \text{s.t.}\quad&\mathbb{E}_{y\sim\mathbf{P}}[c(z,y)]-\tau\leqslant\gamma\cdot\min_{\mathbb{Q}\in\mathcal{A}(x)}(\mathbb{P},\mathbb{Q}),\quad\forall\,\mathbb{P}\in\mathcal{P}(\mathcal{Y})\\ &z\in\mathcal{Z},\quad\tau\geqslant 0,\end{aligned}$$

该模型通过考虑不确定概率分布的支撑集之外的分布信息，在对冲分布不确定性的同时，又能有效缓解模型错误所带来的决策下行风险。全局鲁棒优化模型可与主流的不确定性优化模型无缝集成，且不会产生任何额外的计算成本。

大多数关于决策规则学习的文献都假设了决策规则函数的参数形式，但 Zhang 等[51]在 Wasserstein 分布不确定集下研究了分布鲁棒条件报童问题，而没有对假设类设置一个明确的结构：

$$z^{*}(\cdot)\in\operatorname*{arginf}_{z(\cdot)\in\mathcal{Z}}\sup_{d_W(\mathbb{Q},\mathbb{Q}_0)\leqslant r}\mathbb{E}_{(x,y)\sim\mathbb{Q}}[c(z(x),y)].$$

其中，假设类 Z 不包含任何结构信息。他们通过“Shapley”插值拓展得到最优订货决策规则。类似的思路被扩展到分布间距离为因果传输距离（Causal Transport Distance）——一类在 Wasserstein 距离的基础上增加条件分布约束的分布距离度量。

2　预测驱动最优化在运营管理中的应用

本节将介绍预测驱动最优化方法在运营管理领域的应用研究，重点关注交通与物流管理、库存管理以及收益管理。

2.1　预测驱动交通与物流管理

在交通与物流管理领域，“先预测后优化”方法已得到广泛应用。Glaeser 等[52]研究了在线零售商的时空选址问题。他们利用人口统计、经济数据、商业地点数据以及零售商的历史销售与运营数据，采用固定效应回归模型估计时空竞食效应（Cannibalization Effects），并运用随机森林预测需求。基于上述预测结果，他们进一步求解优化问题，以确定最优选址与排程决策。Hu 等[47]研究了在商品需求分布不确定性下的多阶段分布鲁棒枢纽选址问题。他们通过融合时间序列预测模型，构建了基于 Wasserstein 距离的嵌套分布不确定集，并进一步发展了预算驱动的多阶段枢纽选址模型。Mao 等[48]研究了两阶段联合生产与服务规划问题，该问题中存在需求不确定性且服务决策是需求的一个关键协变量。为了解决服务依赖性、跨产品需求的相关性以及异方差性问题，他们使用了似不相关回归模型和可行广义最小二乘法来开发需求预测计量模型，并基于此构建了相应的预测驱动型分布不确定集。

近年来，“联合预测优化”方法也成为了研究与应用的热点。Elmachtoub 等[53]利用 SPO 损失函数

构建了决策树模型。他们的方法得益于决策树的可解释性，通过将协变量空间划分为不同部分，每部分都对应着一个最优决策。他们证明了叶节点成本向量的平均值可以最小化该叶节点的 SPO 损失，这极大地简化了在 SPO 损失下训练决策树的优化问题。他们利用真实数据对具有不确定旅行时间的最短路径问题进行了研究。实验结果表明，与分类回归树（CART）等侧重于最小化预测误差的机器学习方法相比，他们的方法能够提供更高质量的决策，并且模型复杂度显著降低。Demirovic 等[54, 55]将 SPO 损失函数应用于排序优化和动态规划中的特定问题。基于 SPO 损失函数构建的预测模型方法，同样可推广应用于更广泛的物流管理相关问题，如最后一英里配送问题、海上运输问题以及船舶检修问题[56, 57]。

2.2 库存管理

有效利用协变量信息能够显著提高管理决策的效能，这在众多“先预测后优化”的实践研究中得到了体现。Ban 等[58]研究了需求不确定的新产品动态采购问题。企业虽无法预知新产品的需求，却掌握过往同类产品销售的相关数据，如历史需求和产品特性等协变量信息。他们利用协变量关联相似产品的历史需求数据，构建了新产品的需求预测模型，并据此发展了情景树与多阶段随机规划模型。研究证明，随着样本量的增加，所提出的方法趋于渐近最优。通过运用全球时尚零售商 Zara 提供的真实数据，研究揭示了忽视协变量信息可能导致最优决策出现系统性偏差，并可能引起总成本上升 6% 至 15%。Bertsimas 和 Kallus[31]应用了基于权重的条件分布预测方法，解决了一家大型媒体公司的库存管理问题，该公司需对不同零售点的各类产品库存进行管理。他们采用机器学习模型，将需求视作店铺位置、电影属性（如类型、评分、票房等）及其他大规模公共数据（如电影的本地化谷歌搜索查询）的函数，进行建模分析。研究表明，该方法能显著缩小不考虑上下文数据的朴素方法与具有完美预见能力的方法（即在需求实现前已知需求）之间的成本差异，达到 88% 的成本缩减。Lin 等[45]探讨了在一个具有未知新产品需求的报童模型中，如何在利润风险约束下最大化期望利润。他们同样采用机器学习方法，根据协变量对新产品与先前产品间的相似度进行加权。然后，该权重被用于构建期望利润目标函数以及利润风险约束的近似，从而得到风险规避报童模型的数据驱动近似解。他们发现，在数据驱动型决策中，平均利润可能会随着风险承受能力的增加而增加，这与理论上风险规避的报童模型所预期的结果相反。这一现象的出现，主要原因是利润风险约束在缓解数据驱动决策中抽样误差和模型误设方面发挥了有效的规范作用。关于新产品需求预测的其他研究还包括 Baardman 等[59]。

Ban 和 Rudin[27]提出了两种联合预测优化方法求解报童问题中的最优订货量。第一种方法基于经验风险最小化原则，通过求解单一问题确定订购量，其中决策变量涉及将产品特征映射至订购量的决策规则，目标在于最小化基于样本数据的成本估计。第二种方法采用核回归模型对条件需求分布进行建模，并运用排序算法确定最优订购量。论文推导了这两种方法的样本外表现的理论界限，并将这些方法应用于英国某大型教学医院的急诊室护士人员配备问题。研究结果表明，所提出的方法在样本外成本方面相较于现行实践基准有高达 24% 的提升。Qin 等[60]在数据驱动的背景下，对经典的多期联合定价与库存控制问题进行了研究。在该研究中，需求分布未知，而决策者仅能获得由一系列候选函数所构成的需求假设集。进一步假设真实的需求函数可以表示为需求假设集中候选函数的非负组合，并提出了一种数据驱动型近似算法。研究证明了算法的样本复杂度界限：为确保接近最优利润，每一周期所需的样本数量需满足 $\mathcal{O}(\varepsilon^{-2}T^{6}\log T)$，其中 T 代表周期数，而 ε 表示数据驱动策略的期望利润与最优期望利润之间的绝对差值。数值研究部分阐述了如何基于数据构建需求假设集，并验证了所提出的数据驱动算法能有效地解决动态问题，同时在与基准算法的比较中显著缩小了最优利润差值。

端到端决策框架，尤其是基于深度学习的决策策略，已在众多库存管理研究中得到了应用[61-63]。Qi 等[62]的研究聚焦于多周期库存补货问题，该问题涉及不确定的需求与供应商提前期。该框架运用深度学习模型，直接根据输入特征预测建议的补货量，无需对不确定变量的分布做出任何先验假设。作者在京东开展的实地实验结果显示，与京东之前的做法相比，新算法显著降低了持有成本、缺货成本、总库存成本以及周转率。此外，该算法还缩短了决策周期，并提供了具备通用性与扩展性的自动化库存管理方案。Tian 和 Zhang[63]同样在报童问题中应用深度学习模型以确定订货量，通过利用文本在线评论数据，他们证明了该方法能将订购决策成本降低 28.7%。Neghab 等[64]专注于处理具有复杂相关

需求和不可观察系统状态的报童模型，提出了一种集成优化算法，该算法基于神经网络与隐马尔可夫模型。最后，对数据驱动模型在库存管理理论与实践方面影响感兴趣的读者，可进一步参阅 Erkip[65] 的综述。

2.3　数据驱动收益管理

收益管理的核心目标是通过优化选品、定价、促销等策略，在资源有限及市场波动的环境下，最大化企业的总收益。该领域的研究问题多源自企业的实际管理需求，并且需充分考虑具体商业实践情境，例如在预测产品需求时须兼顾产品生命周期、类别等因素[66]。这催生了一系列具有实际商业需求的研究。Baardman 等[67] 提出了一种需求预测模型，该模型能够有效捕捉价格、市场环境与顾客行为之间的相互作用效应。该模型采用的估计方法融合了正则化有界变量最小二乘法、工具变量法等计量经济学与因果推断领域的常规技术，进而为客户提供定制化的价格促销策略。以快时尚企业 Oracle 的实际数据为例，该需求模型成功降低了 11% 的预测误差，并使得相应的最优策略带来了 3% 至 11% 的利润增长。此外，相关研究还涵盖了运用竞争对手价格信息进行需求预测、全新产品需求预测以及多销售渠道中异质性顾客购买行为的预测[68, 69]。

在线学习框架整合了预测与决策制定过程，其运作流程与联合预测优化相似。在初始阶段，决策者对模型（如市场价格弹性）及数据（如历史价格与需求）认识不足。随着时间的推移，企业持续进行决策（如价格设定），并同步收集数据以学习模型参数，逐步逼近最优决策。此过程中，数据收集、模型推理与决策制定相互依存，并循环动态地推进。在线学习方法在个性化定价领域的应用尤为广泛，即企业依据产品特性为各别产品定制价格，并兼顾客户对产品价值随时间变化的评估。Qiang 和 Bayati[70] 首次引入协变量至需求函数中，其中需求函数对价格及协变量（包括客户特征与市场环境）呈线性关系。决策者须同步确定产品价格并学习协变量的线性系数。在特定条件下，贪婪算法能够实现 $\log(T)$-遗憾，其中 T 代表学习阶段的持续时间。后续研究进一步探讨了协变量的稀疏性、非参数需求函数、低销量产品定价策略以及非独立同分布的协变量特性[71, 72]。

此外，还有研究[73, 74] 探讨了在需求函数未知条件下，卖方如何通过定价策略实现收益最大化的问题。由于决策者仅基于未知分布中抽取的样本，直接将样本映射至价格决策，而非估计需求函数，因此该方法实质上属于端到端优化的范畴。在这些研究中，Chen 等[74] 针对多产品定价问题进行了深入探讨。作者未使用交易数据来拟合离散选择模型（Discrete Choice Model），而是基于数据将顾客潜在估值的不确定性集合建模为一个多面体集合。研究假设新顾客的估值遵循由多面体结构隐含的经验分布，卖方据此通过最大化极值概率分布下的收入来设定价格。基于真实数据，验证了在历史数据规模受限或模型设定存在偏差的情况下，该定价策略具有显著优势。

3　未来研究方向与挑战

本节旨在探讨预测驱动最优化领域未来研究的重要方向。

（1）分布不确定性。在已有的预测驱动最优化理论与应用研究中，通常假定不确定变量的分布在样本外保持不变。因此，多采用随机规划模型进行处理，而考虑分布不确定性的研究相对稀缺，且主要集中在先预测后优化的框架内。在实际数据生成过程中，不确定变量的分布往往呈现出参数性、甚至结构性的变化，这些变化超出了假定已知分布的随机规划模型的处理范畴。此外，分布不确定性模型也不一定相对保守。例如，Van Parys 等[75] 借助大偏差理论阐明，基于相对熵距离的分布鲁棒优化问题，等价于在对样本外失望进行约束的情况下寻找最不保守的预测分布与决策。因此，未来研究的一个重要方向在于如何扩展现行的分布鲁棒优化模型，赋予其更为灵活的统计与优化建模能力，从而使其更加“智慧”地求解运营管理决策问题。

（2）可解释性。在自动化决策系统中，决策实体在进行决策时需要提供“有意义的逻辑信息”（即具备“可解释性”）。尽管在机器学习领域，可解释性问题已受到广泛关注，然而在预测驱动最优化的研究中，这一议题尚未得到充分探讨。当前，探索如何借助多种方法提升预测驱动最优化模型的可解释性，从而使决策者能够更加直观地理解和信赖模型输出，正成为亟待深入研究的重要课题。

（3）因果推断与内生不确定性。虽然决策依赖不确定性问题——即不确定变量受决策影响的情形——已取得一定研究进展，但在考虑协变量信息的背景下，相关文献仍相对匮乏[31, 39, 48]。在众多实际问题中，常常存在不确定变量与决策之间具有未知的因果关联。例如，在设施选址问题中，设施的布

局可能引起所在区域需求的变动；同样，在报童问题的定价决策中，需求量以及影响需求量的其它协变量可能受价格影响。由于决策是基于不确定变量的历史数据做出的，当尝试使用回归模型建立它们之间的关系时，往往会面临内生性问题，从而导致回归系数乃至后续决策出现偏差。因此，在处理具有决策依赖不确定性问题时，需要运用因果推断方法，以应对内生不确定性问题。

（4）不可观测效应。现代大规模数据集通常由具有不同结构的数据子集构成。例如，总需求数据往往是由基于人口统计学特征划分的多个需求子数据集汇总而成。一方面，仅依靠最小化总数据集分布的平均目标损失，并不足以确保各个基于子数据集分布的目标损失均得到有效控制。决策者通常仅能观测到总数据集的概率分布，而难以获取子数据集的具体信息，即对应的隐变量信息。另一方面，预测模型往往难以涵盖所有对因变量有影响的协变量。以时空需求预测为例，城市区位、文化背景、经济趋势周期等因子显然会对需求水平产生影响，但这些因素往往难以直接度量。这些不可观测因素可以被视为隐变量（Latent Variables）。因此，如何处理不可观测效应也是预测驱动最优化发展的重要方向之一。

（5）模型误设。模型误设（Model Misspecification）一直是统计学领域的基础性问题。早期研究中，学者们致力于深入剖析模型设定偏差对统计推断及参数估计的具体影响，并在此基础之上逐步发展出模型选择、模型验证、偏差诊断及检验等研究方向。近年来，模型误设问题还在一系列热门领域中被广泛研究，包括机器学习、贝叶斯推断、因果推断、金融建模以及动态最优控制[76-78]等。近期，Cerreia-Vioglio等[79]与 Liu 等[50]分别从决策理论与优化建模的视角，将分布不确定性（Ambiguity）与模型误设整合入不确定性决策模型与分布鲁棒优化模型中。特别地，Liu 等[50]提出了全局分布鲁棒优化模型，该模型规定鲁棒性约束必须严格适用于分布不确定性集中的概率分布；而对于分布在该不确定性集之外的情形（即出现模型设定偏差的情况），则允许鲁棒性约束在一定范围内被放宽，其放宽程度受模型设定偏差容忍度的制约。此外，运营管理领域中，运用模型平均方法以规避模型误设的潜在研究方向亦值得关注[80]。

4 总结与展望

预测驱动最优化现已成为提升决策效率的关键工具，它是预测模型与优化模型的深度融合。构建具备优良统计特性及高效优化架构的预测驱动最优化模型是当前一项复杂且至关重要的挑战。实现此目标不仅需要汲取统计学、数据科学、经济学等领域的成熟理论，并且需要在实际问题中不断探索，挖掘预测驱动最优化模型在特定决策场景下的独特结构。将预测模型，不确定性建模，最优化及机器学习整合至决策过程中，为复杂管理决策问题提供创新的分析视角与管理洞见，正在成为意义深远的统计管理研究。

参 考 文 献

[1] Sreedevi R, Saranga H. Uncertainty and supply chain risk: the moderating role of supply chain flexibility in risk mitigation. International Journal of Production Economics, 2017, 193: 332—342.

[2] Guan DB, Wang DP, Hallegatte S, et al. Global supply-chain effects of COVID-19 control measures. Nature Human Behaviour, 2020, 4(6): 577—587.

[3] Sun YD, Zhu SP, Wang DP, et al. Global supply chains amplify economic costs of future extreme heat risk. Nature, 2024, 627(8005): 797—804.

[4] 陈松蹊，毛晓军，王聪. 大数据情境下的数据完备化：挑战与对策. 管理世界，2022，38(1)：196—206.

[5] 徐宗本，冯芷艳，郭迅华，等. 大数据驱动的管理与决策前沿课题. 管理世界，2014(11)：158—163.

[6] 陈国青，曾大军，卫强，等. 大数据环境下的决策范式转变与使能创新. 管理世界，2020，36(2)：95—105，220.

[7] Ye YY. Interior point algorithms: theory and analysis. New Jersey: Wiley, 2011.

[8] Yuan YX. Recent advances in trust region algorithms. Mathematical Programming, 2015, 151(1): 249—281.

[9] Marinescu DC. Cloud computing: theory and practice. Morgan Kaufmann, 2022.

[10] Shapiro A, Dentcheva D, Ruszczynski A. Lectures on stochastic programming: modeling and theory, Third Edition. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2021.

[11] Delage E, Ye YY. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research, 2010, 58(3): 595—612.

[12] Li XC, Ye YY. Online linear programming: dual convergence, new algorithms, and regret bounds. Operations Research, 2022, 70(5): 2948—2966.

[13] Wiesemann W, Kuhn D, Sim M. Distributionally robust convex optimization. Operations Research, 2014, 62(6): 1358—1376.

[14] 刘睿，李宁东，于美泽，等. 复杂国际形势下算力产业的发展研究. 信息通信技术与政策，2024，50(2)：63—67.

[15] Gelman A，Vehtari A. What are the most important statistical ideas of the past 50 years? Journal of the American Statistical Association，2021，116(536)：2087—2097.
[16] Brown TB，Mann B，Ryder N，et al. Language models are few-shot learners// Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver：ACM，2020：1877—1901.
[17] Wu TY，He SZ，Liu JP，et al. A brief overview of ChatGPT：the history，status quo and potential future development. IEEE/CAA Journal of Automatica Sinica，2023，10(5)：1122—1136.
[18] 洪永淼，汪寿阳. 大数据如何改变经济学研究范式. 管理世界，2021，37(10)：40—55，72.
[19] 洪永淼，刘俸奇，薛涧坡. 政府与市场心理因素的经济影响及其测度. 管理世界，2023，39(3)：30—51.
[20] Athey S，Imbens GW. Machine learning methods that economists should know about. Annual Review of Economics，2019，11：685—725.
[21] Choudhury P，Allen RT，Endres MG. Machine learning for pattern discovery in management research. Strategic Management Journal，2021，42(1)：30—57.
[22] Xu YM，Cohen SB. Stock Movement Prediction from Tweets and Historical Prices// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Melbourne：Association for Computational Linguistics，2018：1970—1979.
[23] Martínez-de-Albéniz V，Belkaid A. Here comes the Sun：fashion goods retailing under weather fluctuations. European Journal of Operational Research，2021，294(3)：820—830.
[24] Jacquillat A. Predictive and prescriptive analytics toward passenger-centric ground delay programs. Transportation Science，2022，56(2)：265—298.
[25] Lopes J，Guimarães T，Santos MF. Predictive and prescriptive analytics in healthcare：a survey. Procedia Computer Science，2020，170：1029—1034.
[26] Nguyen T，Zhou L，Spiegler V，et al. Big data analytics in supply chain management：a state-of-the-art literature review. Computers & Operations Research，2018，98：254—264.
[27] Ban GY，Rudin C. The big data newsvendor：practical insights from machine learning. Operations Research，2019，67(1)：90—108.
[28] Chen NY，Hu M. Frontiers in service science：data-driven revenue management：the interplay of data，model，and decisions. Service Science，2023，15(2)：79—91.
[29] Sadana U，Chenreddy A，Delage E，et al. A survey of contextual optimization methods for decision-making under uncertainty. European Journal of Operational Research，2024.
[30] Kannan R，Bayraksan G，Luedtke JR. Data-driven sample average approximation with covariate information.(2022-07-27)/[2024-06-26]. https://doi.org/10.48550/arXiv.2207.13554.
[31] Bertsimas D，Kallus N. From predictive to prescriptive analytics. Management Science，2020，66(3)：1025—1044.
[32] Bengio Y. Using a financial training criterion rather than a prediction criterion. International Journal of Neural Systems，1997，8(4)：433—443.
[33] Donti PL，Amos B，Kolter JZ. Task-based end-to-end model learning in stochastic optimization// Proceedings of the 31st International Conference on Neural Information Processing Systems. California. ACM，2017：5490—5500.
[34] Grigas P，Qi M，Shen ZJ M. Integrated conditional estimation-optimization. 2021，doi：arxiv.org/abs/2110.12351.
[35] Elmachtoub AN，Grigas P. Smart "predict，then optimize". Management Science，2022，68(1)：9—26.
[36] Kong L，Cui J，Zhuang Y，et al. End-to-end stochastic optimization with energy-based model. Advances in Neural Information Processing Systems 2022，35：11341—11354.
[37] Loke GG，Tang Q，Xiao Y. Decision-driven regularization：a blended model for predict-then-optimize.[2024-06-28]. https://api.semanticscholar.org/CorpusID:235435571.
[38] Ho-Nguyen N，Kılınç-Karzan F. Risk guarantees for end-to-end prediction and optimization processes. Management Science，2022，68(12)：8680—8698.
[39] Bertsimas D，Koduri N. Data-driven optimization：a reproducing kernel Hilbert space approach. Operations Research，2022，70(1)：454—471.
[40] Oroojlooyjadid A，Snyder LV，Takáč M. Applying deep learning to the newsvendor problem. IISE Transactions，2020，52(4)：444—463.
[41] Rychener Y，Kuhn D，Sutter T. End-to-end learning for stochastic optimization：a bayesian perspective. International Conference on Machine Learning. PMLR，2023：29455—29472.
[42] Zhang Y，Liu J，Zhao X. Data-driven piecewise affine decision rules for stochastic programming with covariate information. 2023，doi：/arxiv.org/abs/2304.13646.
[43] Feng Q，Shanthikumar JG. The framework of parametric and nonparametric operational data analytics. Production and Operations Management，2023，32(9)：2685—2703.
[44] Bertsimas D，Van Parys B. Bootstrap robust prescriptive analytics. Mathematical Programming，2022，195(1)：39—78.
[45] Lin SC，Chen YF，Li YZ，et al. Data-driven newsvendor problems regularized by a profit risk constraint. Production and Operations Management，2022，31(4)：1630—1644.
[46] Kannan R，Bayraksan G，Luedtke JR. Residuals-based distributionally robust optimization with covariate information. Mathematical Programming 2023：1—57.
[47] Hu J，Chen Z，Wang SM. Budget-driven multi-period hub location：a robust time series approach. SSRN Electronic Journal，2022：4221971.
[48] Mao YC，Saldanha-da-Gama F，Wang SM，et al. Predictive Production-and-Service Planning：Ambiguity Aversion with Performance Guarantees. Working paper，2023.

[49] Sim M, Tang QS, Zhou ML, et al. The Analytics of Robust Satisficing: Predict, Optimize, Satisfice, then Fortify. SSRN Electronic Journal, 2021: 3829562.

[50] Liu F, Chen Z, Wang SM. Globalized distributionally robust counterpart. INFORMS Journal on Computing, 2023.

[51] Zhang LH, Yang JC, Gao R. Optimal robust policy for feature-based newsvendor. Management Science, 2024, 70(4): 2315—2329.

[52] Glaeser CK, Fisher M, Su XM. Optimal retail location: empirical methodology and application to practice. Manufacturing & Service Operations Management, 2019, 21(1): 86—102.

[53] Elmachtoub AN, Liang JCN, McNellis R. Decision trees for decision-making under the predict-then-optimize framework. // International Conference on Machine Learning. PMLR, 2020: 2858—2867.

[54] Demirovic E, Stuckey PJ, Bailey J, et al. Predict+Optimise with ranking objectives: exhaustively learning linear functions. IJCAI, 2019: 1078—1085.

[55] Demirovic E, Stuckey PJ, Guns T, et al. Dynamic programming for Predict + Optimise. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(2): 1444—1451.

[56] Chu HR, Zhang WS, Bai PF, et al. Data-driven optimization for last-mile delivery. Complex & Intelligent Systems, 2023, 9(3): 2271—2284.

[57] Tian XC, Yan R, Liu Y, et al. A smart predict-then-optimize method for targeted and cost-effective maritime transportation. Transportation Research Part B: Methodological, 2023, 172: 32—52.

[58] Ban GY, Gallien J, Mersereau AJ. Dynamic procurement of new products with covariate information: the residual tree method. Manufacturing & Service Operations Management, 2019, 21(4): 798—815.

[59] Baardman L, Levin I, Perakis G, et al. Leveraging comparables for new product sales forecasting. Production and Operations Management, 2018, 27(12): 2340—2343.

[60] Qin HZ, Simchi-Levi D, Wang L. Data-driven approximation schemes for joint pricing and inventory control models. Management Science, 2022, 68(9): 6591—6609.

[61] Madeka D, Torkkola K, Eisenach C, et al. Deep inventory management. 2022, doi: arxiv.org/abs/2210.03137.

[62] Qi M, Shi YY, Qi YZ, et al. A practical end-to-end inventory management model with deep learning. Management Science, 2023, 69(2): 759—773.

[63] Tian YX, Zhang C. An end-to-end deep learning model for solving data-driven newsvendor problem with accessibility to textual review data. International Journal of Production Economics, 2023, 265: 109016.

[64] Pirayesh Neghab D, Khayyati S, Karaesmen F. An integrated data-driven method using deep learning for a newsvendor problem with unobservable features. European Journal of Operational Research, 2022, 302(2): 482—496.

[65] Erkip NK. Can accessing much data reshape the theory? Inventory theory under the challenge of data-driven systems. European Journal of Operational Research, 2023, 308(3): 949—959.

[66] Gallien J, Mersereau AJ, Garro A, et al. Initial shipment decisions for new products at Zara. Operations Research, 2015, 63(2): 269—286.

[67] Baardman L, Boroujeni SB, Cohen-Hillel T, et al. Detecting customer trends for optimal promotion targeting. Manufacturing & Service Operations Management, 2023, 25(2): 448—467.

[68] Ferreira KJ, Lee BHA, Simchi-Levi D. Analytics for an online retailer: demand forecasting and price optimization. Manufacturing & Service Operations Management, 2016, 18(1): 69—88.

[69] Arslan HA, Easley RF, Wang RX, et al. Data-driven sports ticket pricing for multiple sales channels with heterogeneous customers. Manufacturing & Service Operations Management, 2022, 24(2): 1241—1260.

[70] Qiang S, Bayati M. Dynamic pricing with demand covariates. SSRN Electronic Journal, 2016.

[71] Ban GY, Keskin NB. Personalized dynamic pricing with machine learning: high-dimensional features and heterogeneous elasticity. Management Science, 2021, 67(9): 5549—5568.

[72] Chen NY, Gallego G. Nonparametric pricing analytics with customer covariates. Operations Research, 2021, 69(3): 974—984.

[73] Allouah A, Bahamou A, Besbes O. Pricing with samples. Operations Research, 2022, 70(2): 1088—1104.

[74] Chen NY, Cire AA, Hu M, et al. Model-free assortment pricing with transaction data. Management Science, 2023, 69(10): 5830—5847.

[75] Van Parys BPG, Esfahani PM, Kuhn D. From data to decisions: distributionally robust optimization is optimal. Management Science, 2021, 67(6): 3387—3402.

[76] Wang Y, Blei D. Variational Bayes under model misspecification. Advances in Neural Information Processing Systems 2019: 32.

[77] Uppal R, Wang T. Model misspecification and underdiversification. The Journal of Finance, 2003, 58(6): 2465—2486.

[78] Hansen LP, Sargent TJ, Turmuhambetova G, et al. Robust control and model misspecification. Journal of Economic Theory, 2006, 128(1): 45—90.

[79] Cerreia-Vioglio S, Hansen LP, Maccheroni F, et al. Making decisions under model misspecification. SSRN Electronic Journal, 2020.

[80] He BH, Ma SG, Zhang XY, et al. Rank-based greedy model averaging for high-dimensional survival data. Journal of the American Statistical Association, 2023, 118(544): 2658—2670.

Predictive Optimization: Uncertainty, Statistical Theory, and Managerial Application

Shuming Wang[1*]　Yuchen Mao[1]　Shouyang Wang[1, 2, 3, 4*]

1. *School of Economics and Management, University of Chinese Academy of Sciences, Beijing* 100190
2. *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190
3. *Center for Forecasting Science, Chinese Academy of Sciences, Beijing* 100190
4. *School of Entrepreneurship and Management, ShanghaiTech University, Shanghai* 201210

Abstract　Modern decision-making in management problems is confronted with intricate uncertainties. With the extensive applications of big data, persistent enhancement of optimization techniques and computing power, as well as flourishing development of statistics and machine learning, predictive optimization is emerging as a potent tool to address complex decision-making problems under uncertainty. By integrating statistical (predictive) modeling with decision optimization, predictive optimization achieves a joint statistical management of uncertainty and decision efficacy, thereby forming a statistically consistent and efficient data-driven decision-making paradigm. In this study, we focus on statistical predictive modeling and management decision optimization under uncertain environments, exploring the frameworks of predictive optimization models with known (i. e., stochastic optimization) and unknown (i. e., distributionally robust optimization) distributions. Furthermore, we introduce the state-of-the-art applications of predictive optimization in operations management. Finally, we summarize key future research directions and challenges.

Keywords　predictive optimization; uncertainty; statistical theory; managerial application

（责任编辑　张　强）

* Corresponding Authors, Email: sywang@amss.ac.cn; wangshuming@ucas.edu.cn