From Data to Decision: A Data-Driven Approach to the Newsvendor Problem

—— 课程文献阅读与复现报告

汇报人：[您的姓名]

课程：高级应用统计

2025 年 12 月 17 日

摘要

本文深度解读了 Huber 等人发表于 *EJOR* (2019) 的论文。文章针对报童问题中需求分布未知的核心挑战，提出了基于大数据的三层决策框架。我们详细阐述了从需求预测到库存优化的统计方法论，并基于 Kaggle 的 French Bakery 数据集进行了实证复现，验证了非参数方法在库存决策中的有效性。

1 The Basic Research Problem

1.1 商业背景与权衡

本文聚焦于零售管理中经典的**报童问题**。零售商（如连锁面包店）需在销售季节前决定易腐产品的订货量 q。核心权衡在于最小化期望总成本：

$$\min_q \mathbb{E}[C(q, D)] = \mathbb{E}[c_u(D-q)^+ + c_o(q-D)^+] \tag{1}$$

其中 D 为随机需求，c_u 为缺货成本，c_o 为超储成本。

1.2 统计学挑战

在理论最优解中，订货量 q^* 取决于需求累积分布函数 F 的分位数：$q^* = F^{-1}(\frac{c_u}{c_u+c_o})$。然而，现实中的**根本难题**在于：

- **分布未知**：真实的需求分布 F 往往无法获知，且可能随时间变化。
- **特征利用不足**：传统方法往往忽略了天气、节假日、促销等外部特征向量 X 对需求分布的影响。

因此，本文的研究问题是：**如何利用历史数据 (D_t, X_t)，在不预设分布形式的前提下，构建数据驱动的模型以实现成本最小化？**

2 The Idea and Methodology

本文提出了一个**三层递进的数据驱动框架**，涵盖了从参数估计到非参数优化的完整路径。

2.1 Level 1: Demand Estimation (点预测)

利用统计学习方法估计给定特征 x 下的需求期望 $\hat{y}(x) = \mathbb{E}[d|x]$。

- **传统方法**：ARIMA, ETS (指数平滑)。
- **机器学习**：文章构建了单隐层多层感知机。

$$\hat{y}(x) = f(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}) \tag{2}$$

通过引入特征工程（如滞后销量 Lag_1, Lag_7 和日历特征），捕捉非线性模式。

2.2 Level 2: 库存优化

基于预测结果 $\hat{y}$ 和预测误差 $\epsilon = d - \hat{y}$ 进行决策。

- **参数化方法**：假设误差服从正态分布 $\epsilon \sim \mathcal{N}(0, \hat{\sigma}^2)$。

$$q(x) = \hat{y}(x) + \Phi^{-1}(\text{target ratio}) \cdot \hat{\sigma}$$

- **非参数方法：样本均值逼近**。不假设分布，直接使用历史误差样本的经验分布寻找分位数。 **优势**：避免了模型误设风险，更具鲁棒性。

2.3 Level 3: 集成优化

跳过点预测，直接建立特征 x 到最优订货量 q^* 的映射。这等价于**分位数回归**。我们将报童损失函数直接作为神经网络的训练目标：

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\left[c_u(d_i - q(x_i))^+ + c_o(q(x_i) - d_i)^+\right] \tag{3}$$

该方法能自动适应**异方差性**（即需求的波动随特征变化）。

3 数据来源及处理

3.1 原论文数据处理

原研究采用了德国某大型连锁面包店的销售数据，涵盖 5 家门店的 11 种核心产品，时间跨度为 88 周。针对零售数据中普遍存在的**需求截断**问题——即因库存耗尽导致观

测到的销量低于真实需求，原作者创新性地利用日内销售模式对缺货时段进行了插值还原，从而获得了对真实历史需求的估计。此外，为了捕捉复杂的消费行为模式，原研究构建了包含天气（温度、云层）、地理位置（学校、商圈）及详细日历特征的丰富外生变量集。

3.2 复现数据构建与预处理

鉴于原研究所用的企业私有数据未公开，本研究选取了业务模式高度相似的 **Kaggle "French Bakery Daily Sales"** 公开数据集作为替代。为在现有数据条件下最大程度复现论文的方法论，我们执行了系统性的数据处理流程。

首先进行**数据重构与清洗**。原始数据为交易级记录，我们将其聚合至**日粒度**，按 `date` 和 `article` 汇总销量与收入。受限于缺乏日内库存记录，本复现采用经典假设，即观测销量近似于真实需求（销量 ≈ 需求），并对非营业日进行了补零处理。在此基础上，建立了严格的数据质量控制规则：对价格字段进行标准化解析（处理欧元符号与逗号小数），并剔除 `sales>0` 但价格缺失或非正数的异常记录，以确保数据的准确性。

随后进行**样本筛选与特征工程**。为提升模型的代表性并减少稀疏噪音，我们依据累计销量筛选了 **前 10 核心产品**，剔除长尾低频商品。为了模拟原论文的特征体系，本研究构建了高维特征空间，具体涵盖：(1) **日历特征**，包括 `weekday`, `month` 及基于法国法定节假日库生成的 `is_public_holiday`；(2) **时序特征**，构建 `lag_1`（短期依赖）和 `lag_7`（周度周期性）以捕捉自相关性。最后，数据集严格按时间序列顺序划分为训练集与测试集（按 80/20 划分），以避免数据泄漏。

3.3 论文原实证结果

基于德国某大型连锁面包店 88 周的真实运营数据（涵盖 5 家门店与 11 个 SKU），Huber 等人 (2019) 进行了系统的实证评估。实验采用滚动窗口机制进行严格的样本外测试，旨在从数据规模与特征组合两个维度，全面评估不同决策模型的绩效表现。

3.3.1 数据特征与实验设置

针对零售数据中普遍存在的需求删失问题，研究采用了日内销售模式法对历史缺货数据进行了插值还原，从而构建了无偏的需求估计基础。在此基础上，实验设计了从 0.1 到 1.0 不同比例的训练集规模，以量化探究模型性能对数据丰富度的依赖关系，并验证各模型在不同数据环境下的稳健性。

3.3.2 Level 1: 点预测精度分析

在需求估计层面，研究对比了以指数平滑和 ARIMA 为代表的传统时间序列方法与以多层感知机和梯度提升树为代表的机器学习方法。如原文 **Table 3**（图 1）所示，

当机器学习模型仅基于单变量时间序列进行独立训练时，其相较于传统方法的优势并不显著。

然而，当采用跨序列池化训练策略并引入高维外生特征时，机器学习模型展现出显著的性能优势。具体而言，该模型能够有效捕捉周度季节性与天气因素之间的非线性交互效应，从而大幅降低均方根误差。图 1 展示了各模型精度的具体对比，结果显示跨序列池化训练的机器学习方法在各项误差指标上均显著优于传统单变量方法。

Table 3
Forecast performance of the point predictions (sample size: 1.0). The best performance for each metric is underlined. Results that do not differ from the one of the best method at a significance level of 5% for each metric are printed in bold face.

Method	MPE	SMAPE	MAPE	MASE	RMSE	MAE	RAE
Median	−22.34	29.71	39.43	1.01	39.89	15.70	1.72
S-Median	−21.45	24.74	33.73	0.82	28.42	11.99	1.31
S-Naïve	**−11.84**	28.71	34.86	0.92	27.80	12.56	1.37
S-MA	−14.61	23.32	30.15	0.75	22.27	10.14	1.11
ETS	−12.47	22.19	28.47	0.71	21.83	9.66	1.06
S-ARIMA	−14.35	22.88	29.71	0.73	21.40	9.87	1.08
Linear	−18.73	23.75	32.07	0.77	23.43	10.54	1.15
DT-LGBM	−18.80	22.88	31.13	0.73	21.98	9.92	1.08
ANN-MLP	−14.73	22.63	29.59	0.72	21.28	9.75	1.07
Linear (all)	−14.33	22.14	29.18	0.71	21.23	9.63	1.05
DT-LGBM (all)	−13.44	21.51	28.34	**0.68**	**20.06**	**9.15**	**1.00**
ANN-MLP (all)	**−12.62**	**21.42**	**27.87**	**0.68**	**20.09**	**9.16**	**1.00**

图 1: 原文 Table 3：不同预测方法的点预测精度对比

3.3.3 Level 2: 库存绩效与尾部风险

在将预测结果转化为库存决策的过程中，研究揭示了预测精度与最终运营成本之间存在显著的正相关性，且不同优化方法的表现呈现出明显的非对称效应。

如原文 **Table 4**（图 2）所示，在目标服务水平不高于 0.9 的区间内，非参数的样本均值逼近方法普遍优于参数化的正态分布假设。这表明 SAA 方法能够更有效地利用残差分布信息，克服真实需求分布的有偏性。然而，当服务水平提升至 0.95 时，受限于尾部样本的稀疏性，SAA 方法的估计方差显著增大；此时，正态分布假设凭借其参数化的正则特性，反而能提供更为稳定的成本控制表现。图 2 清晰地展示了 SAA 方法在中低服务水平区间具有更低的相对成本。

3.3.4 Level 3 与样本量敏感性分析

针对端到端的集成优化及样本量的边际效应，原文 **Figure 5**（图 3）提供了直观的趋势分析。研究发现，基于分位数回归的集成优化策略仅在低服务水平且数据量极其

Table 4
Inventory performance analysis: Average cost increase relative to the best approach and average service level (SL) for various target service levels (TSLs) and a sample size of 1.0. Methods denoted with *all* are trained on data across all products and stores. The best approach for each target service level is underlined. Results that do not differ from the one of the best method at a significance level of 5% for each service level are printed in bold face.

	Method		TSL = 0.5		TSL = 0.6		TSL = 0.7		TSL = 0.8		TSL = 0.9		TSL = 0.95	
	Estimation	Optimization	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL
Benchmarks	Median	Norm	72.5	0.61	83.9	0.72	93.2	0.79	99.4	0.86	97.8	0.92	92.0	0.95
		SAA	72.5	0.61	79.4	0.70	87.9	0.79	99.6	0.87	109.5	0.94	101.9	0.98
	S-Median	Norm	31.8	0.64	33.5	0.74	34.7	0.83	34.9	0.89	32.2	0.95	29.6	0.97
		SAA	31.8	0.64	30.3	0.72	30.4	0.80	29.5	0.88	27.4	0.95	31.2	0.98
	S-Naive	Norm	38.0	0.51	37.5	0.63	37.0	0.75	37.3	0.85	37.2	0.93	37.2	0.96
		SAA	38.4	0.51	37.6	0.61	35.6	0.71	34.3	0.81	32.2	0.91	33.3	0.96
	S-MA	Norm	11.5	0.56	13.6	0.68	16.0	0.78	17.6	0.86	18.3	0.94	16.6	0.97
		SAA	10.5	0.52	11.0	0.62	11.4	0.73	11.7	0.82	12.2	0.92	13.9	0.96
	ETS	Norm	6.1	0.53	6.7	0.64	7.0	0.74	7.1	0.83	5.6	0.91	5.7	0.95
		SAA	6.2	0.50	6.5	0.61	6.7	0.71	6.7	0.80	5.6	0.90	5.9	0.95
	S-ARIMA	Norm	8.5	0.55	8.9	0.65	8.8	0.75	8.3	0.84	7.5	0.92	7.2	0.95
		SAA	8.0	0.52	8.1	0.62	8.0	0.71	7.7	0.81	6.5	0.91	7.2	0.95
ML single time series	Linear	Norm	15.8	0.58	17.7	0.69	19.6	0.78	20.8	0.85	20.2	0.93	20.9	0.95
		SAA	15.6	0.56	17.2	0.66	18.9	0.75	20.1	0.84	20.7	0.93	21.8	0.96
		QR	10.6	0.54	10.7	0.64	11.4	0.73	11.2	0.82	11.8	0.91	18.9	0.96
	DT-LGBM	Norm	9.0	0.60	8.6	0.68	8.5	0.76	8.8	0.83	10.2	0.89	**15.2**	0.93
		SAA	7.9	0.57	7.7	0.65	7.8	0.73	8.4	0.81	10.0	0.89	14.4	0.94
		QR	11.1	0.59	10.8	0.68	12.1	0.78	15.3	0.85	20.8	0.93	29.0	0.96
	ANN-MLP	Norm	7.2	0.55	8.4	0.66	9.0	0.75	9.6	0.83	9.4	0.91	10.5	0.95
		SAA	6.6	0.52	7.6	0.63	8.2	0.72	8.6	0.82	8.6	0.91	10.2	0.95
		QR	7.5	0.53	7.9	0.64	8.6	0.73	9.8	0.82	13.0	0.91	18.1	0.95
ML pooled time series + features	Linear (all)	Norm	5.9	0.53	5.5	0.64	5.6	0.75	6.1	0.84	4.9	0.91	4.0	0.95
		SAA	5.4	0.51	5.3	0.62	5.0	0.72	5.3	0.82	5.2	0.91	4.9	0.95
		QR	5.1	0.52	4.5	0.62	5.2	0.72	7.2	0.81	10.0	0.90	12.8	0.95
	DT-LGBM (all)	Norm	**0.6**	0.53	**0.4**	0.62	0.0	0.71	0.1	0.80	0.4	0.87	2.1	0.92
		SAA	0.9	0.51	**0.4**	0.61	**0.0**	0.69	**0.0**	0.79	0.2	0.88	1.7	0.92
		QR	1.6	0.52	1.7	0.61	1.6	0.71	3.1	0.80	6.4	0.90	11.4	0.94
	ANN-MLP (all)	Norm	0.7	0.52	0.7	0.63	0.7	0.73	0.7	0.82	**0.0**	0.90	**0.0**	0.95
		SAA	**0.3**	0.51	**0.2**	0.61	0.3	0.72	0.4	0.81	0.4	0.90	1.5	0.95
		QR	**0.0**	0.50	**0.0**	0.61	0.9	0.72	3.3	0.82	6.8	0.91	11.2	0.95

图 2: 原文 Table 4：不同目标服务水平下的平均库存成本增加比例（相对于最佳方法）

充足的条件下才具有竞争力；在数据稀疏区域，其泛化能力不如“预测加优化”的分离式策略。

此外，样本量敏感性分析表明，机器学习模型是唯一随着样本量增加而持续降低成本的方法，而朴素预测等传统方法的性能并未随数据规模扩大而显著改善。图 3 展示了随着数据量增加，机器学习模型与 SAA 方法组合的成本优势逐渐扩大的趋势，呈现出显著的规模收益。

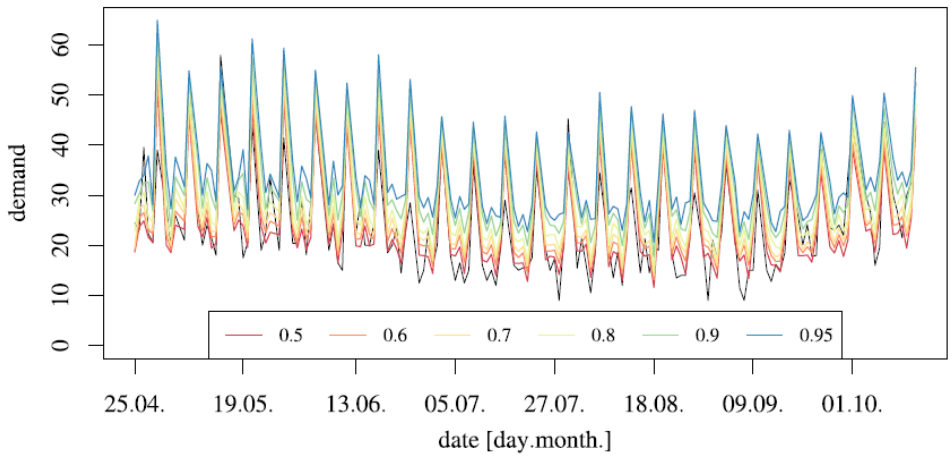

图 3: 原文 Figure 5：不同样本量大小对库存成本的影响 ($SL = 0.7$)

3.4 小组复现结果

为了验证原论文提出的数据驱动框架在不同数据集上的泛化能力与有效性，本研究采用 **Kaggle "French Bakery Daily Sales"** 数据集，筛选销量排名前 10 的核心产品（样本量 $N \approx 6300$），对需求估计、库存优化及集成优化三个层级进行了系统性的复现与实证分析。

3.4.1 Level 1: 需求预测与特征工程

在需求估计层面，本研究构建了包含短期与周期性滞后项（Lag_1, Lag_7）及日历特征（`weekday`, `month`）的高维特征空间，并分别训练了作为基准的线性回归模型与作为核心研究对象的随机森林模型。

实证结果如表 1 所示，机器学习模型展现出显著优越的拟合优度。具体而言，Random Forest 模型的 R^2 Score 达到 **92.89%**，显著高于线性基准模型的 90.10%；同时，其均方根误差从 27.54 降至 23.35，相对改善幅度达 **15.2%**。

这一结果表明，简单的线性模型难以充分挖掘数据中的特征价值。而通过引入非线性模型（随机森林），能够有效捕捉面包销售数据中存在的“周度季节性”与“短期自相关”之间的复杂非线性交互关系，从而大幅降低预测残差，为后续的库存决策提供更精准的均值估计。

表 1: Level 1 需求预测模型性能对比

Model	**RMSE**	**MAE**	R^2 **Score**
Linear Regression	27.54	14.82	90.10%
Random Forest	**23.35**	**12.71**	**92.89%**
Improvement	*-15.2%*	*-14.2%*	*+2.79 %*

3.4.2 Level 2: 统计检验与库存决策优化

在库存决策阶段，本研究首先对 Random Forest 模型的预测残差进行了统计诊断，随后对比了参数化与非参数化方法的成本表现。

1. 残差分布诊断： Shapiro-Wilk 正态性检验结果显示，p-value 远小于显著性水平 0.05 ($p = 1.94 \times 10^{-70}$)，从而在统计上以极高的置信度**拒绝了残差服从正态分布的原假设**。这一显著的非正态特征表明传统参数模型存在模型误设风险，为采用基于样本均值逼近的非参数方法提供了坚实的统计学依据。

2. 决策敏感性与尾部效应：我们测试了不同目标服务水平下各方法的平均日成本，具体结果如表 2 所示。通过对比分析，我们观察到显著的**“尾部效应”：**

在中低服务水平（$SL \leq 0.7$）下，正态参数化方法表现稳健，成本与 SAA 方法持平甚至略优，这符合中心极限定理在分布中心区域的适用性。然而，随着服务水平的提

升（$SL \geq 0.8$），数据驱动的 SAA 方法展现出显著优势；特别是在 $SL = 0.95$ 的极端分位数下，SAA 的平均成本较正态假设降低了约 **10%**。该结果证实，正态假设倾向于低估极端需求的概率（即忽视了“肥尾”现象），而 SAA 方法在处理高服务水平要求的库存决策时具有更强的鲁棒性。

表 2: 不同服务水平下的平均日成本对比

Service Level 目标	**Level 2** 正态	**SAA**	**Level 3** 分位数回归
0.50	**9.52**	9.55	10.56
0.70	**9.32**	9.34	9.55
0.80	9.65	**9.27**	9.81
0.90	10.43	**9.45**	11.95
0.95	11.24	**10.15**	13.19

3. 决策可视化解读：图 4 直观展示了某一产品在测试集期间的决策细节。图中绿色阴影区域表示超储带来的浪费成本，红色阴影区域表示缺货带来的机会成本。可以看出，SAA 方法计算出的订货量（绿线）并非对预测均值（蓝线）的简单线性平移，而是根据局部残差分布特征进行动态调整，从而在需求波峰处有效控制了缺货风险（红色区域面积）。

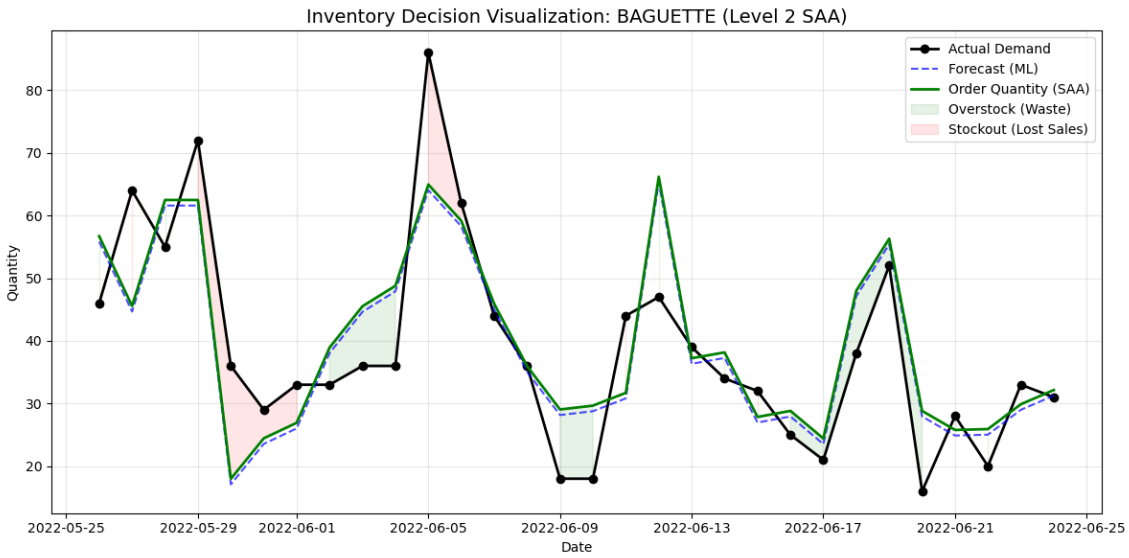

图 4: Level 2 决策可视化：真实需求、ML 预测值与基于 SAA 的最优订货量

3.4.3 Level 3: 基于分位数回归的集成优化

作为本研究的进阶探究，我们复现了基于**分位数回归**的端到端决策模型，旨在验证“集成估计与优化”策略的有效性。

1. 模型构建与超参数调优：具体实施中，本研究采用 `GradientBoostingRegressor` 作为基学习器，将损失函数设定为分位数损失，直接针对特定服务水平 $\alpha = \frac{c_u}{c_u+c_o}$ 优化最优订货量 $q(x)$。为了防止过拟合，我们采用了 3 折时间序列交叉验证，在包含 `n_estimators`、`max_depth` 及 `learning_rate` 的参数网格中进行搜索。

表 3 展示了不同服务水平下的最优模型配置。可以看出，随着目标分位数的提高（即服务水平要求变严），模型倾向于选择更复杂的参数组合（如更高的 `n_estimators`），以捕捉尾部极端值的非线性模式。

表 3: Level 3 集成优化模型最优超参数配置

Target SL	**Test Cost**	**Best Parameters (n_estimators, max_depth, lr)**
0.50	10.56	(400, 2, 0.05)
0.70	**9.55**	(200, 3, 0.05)
0.80	9.81	(100, 3, 0.10)
0.90	11.95	(100, 2, 0.10)
0.95	13.19	(400, 4, 0.03)

2. 综合成本分析与局限性讨论：图 5 汇总了 Normal、SAA 及 Integrated 三种方法在多服务水平下的成本演变趋势。总体而言，随着目标服务水平的提升，由于缺货惩罚权重的增加，所有模型的预期总成本均呈现上升态势。

在具体方法对比中，Integrated 方法（绿线）在 $SL = 0.70$ 处取得了全场最低的平均日成本（9.55），展现了端到端学习策略在特定区间的优化潜力；然而，当进入 $SL \geq 0.90$ 的高服务水平区间时，集成方法并未表现出优于 SAA 方法（橙线）的显著优势，部分指标甚至略有逊色。这一现象与原论文的实证结论高度一致，主要归因于端到端模型试图在有限样本（本研究约 6300 条）下直接学习输入特征与极端分位数（如 95% 分位点）之间的复杂非线性映射，其对数据规模的高敏感性导致了在数据稀疏区域的泛化能力不如结构更为简约的 SAA 方法稳定。

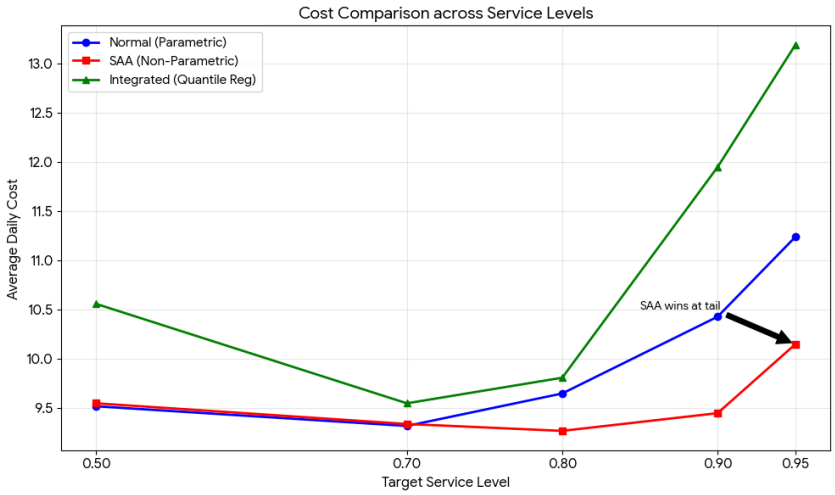

图 5: 多服务水平下的平均成本对比曲线

4 Understanding, Comments and Thinking

4.1 统计学视角的洞察

- **Prediction $\neq$ Decision**：统计上的“高预测精度”（低 MSE）并不完全等同于商业上的“低成本”。针对特定损失函数进行优化是应用统计的高级方向。
- **非参数的胜利**：本研究再次印证了在“大数据”时代，基于经验分布的 SAA 方法往往比强依赖假设的参数模型更安全、更有效。

4.2 局限与改进

- **删失数据**：当前的复现假设 $Sales \approx Demand$，忽略了缺货导致的截断。未来可引入 **Tobit 模型** 或 Survival Analysis 中的 Kaplan-Meier 估计来还原真实需求。
- **算法扩展**：考虑到实际数据量可能有限，未来可对比 **Random Forest** 或 **XGBoost**，这些树模型通常在表格数据上比简单的 ANN 表现更稳健。