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 (Advanced Applied Statistics)

2025 年 12 月 13 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 (Newsvendor Problem)**。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 (Underage cost)，c_o 为超储成本 (Overage cost)。

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 (Unknown Distribution)**：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 (ANN)：** 文章构建了单隐层多层感知机 (MLP)。

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

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: Inventory Optimization (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。

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

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

- **Non-parametric (SAA)：样本均值逼近 (Sample Average Approximation)**。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 (Misspecification) 风险，更具鲁棒性。

2.3 Level 3: Integrated Estimation (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 (Quantile Regression)**。我们将报童损失函数直接作为神经网络的训练目标 (Loss Function)：

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

该方法能自动适应**异方差性**（即需求的波动随特征变化）。

3 Data and Preparation

本研究基于 **French Bakery Daily Sales** 数据，原始为交易级记录。为复现论文方法，我们进行了数据重构与清洗：

- 聚合到**日粒度**：按 `date`，`article` 汇总销量与收入，并对缺失日期进行**补零**，保留零销量日。
- 价格解析与约束：解析含欧元符号与逗号小数的单价，计算 `avg_price`；当 `sales>0` 且 `avg_price` 缺失或非正数时**删除该行**。
- 异常过滤：依据零销量比与非零天数剔除信息稀缺的产品，同时**保留销量 Top 10** 提升代表性。
- 特征工程：构造 `weekday`，`month`，`is_public_holiday`，`lag_1`，`lag_7`；按时间 **80/20** 划分训练与测试集。

4 The Results

4.1 论文原实证结果 (Key Findings from Paper)

基于德国大型面包连锁店的数据，论文得出以下结论：

1. **预测精度**：机器学习模型（特别是利用跨序列信息的 ANN）在 RMSE 指标上显著优于传统时间序列模型。
2. **成本表现**：
   - **SAA vs Normal**：在大部分服务水平下，数据驱动的 SAA 方法成本低于正态分布假设，证明了非参数方法的价值。
   - **Integrated vs Separate**：集成方法在低服务水平下表现优异，但在高服务水平下对数据量要求极高，容易过拟合。

4.2 小组复现结果 (Our Replication Study)

我们基于 **Kaggle "French Bakery Daily Sales"** 数据集（Top 10 产品，约 6300 条样本）进行了完整复现。

- **Level 1 预测表现**：
  - **模型拟合**：Random Forest 取得了 **92.89%** 的 R^2，显著优于 Linear Regression 的 90.10%。RMSE 降低了约 15%，验证了非线性模型在捕捉周期性需求上的优势。

- **统计检验：**
  - **Shapiro-Wilk Test：**预测残差的 p-value 为 1.94×10^{-70}，显著拒绝正态分布假设。这为采用非参数方法 (SAA) 提供了坚实的统计学基础。
- **Level 2 决策敏感性分析 (关键发现)：**我们测试了不同目标服务水平 (SL) 下的平均成本表现，观察到了显著的**“尾部效应”(Tail Effect)：**
  - **在中低服务水平 ($SL \leq 0.7$)：**Parametric (Normal) 方法表现稳健，成本略低于 SAA。这表明在分布中心区域，正态近似依然有效。
  - **在高服务水平 ($SL \geq 0.8$)：**数据驱动的 SAA 方法开始显著跑赢。特别是在 $SL = 0.95$ 时，SAA 的平均成本比 Normal 假设低约 **10%**。
  - **结论：**这复现并深化了原论文的观点——非参数方法 (SAA) 的核心价值在于处理**尾部风险 (Tail Risk)**。当零售商追求高服务水平（即极少缺货）时，依赖正态假设会带来巨大的潜在损失，而 SAA 展现了极强的鲁棒性。
- **Level 3 集成优化（分位数回归）：**
  - **模型与方法：**在服务水平 $\alpha = \frac{c_u}{c_u + c_o}$ 下，直接以分位损失拟合最优订货量 $q(x)$；实现采用 `GradientBoostingRegressor`（`loss=quantile`, `alpha=SL`），并对预测进行**非负裁剪**以符合业务约束。
  - **调参与验证：**使用**时间序列交叉验证**（3 折）在网格 `n_estimators` $\in \{100, 200, 400\}$、`max_depth` $\in \{2, 3, 4\}$、`learning_rate` $\in \{0.03, 0.05, 0.1\}$ 上搜索最优组合。
  - **代表性结果：**例如在 $SL = 0.70$ 时，最优参数为 `(200, 3, 0.05)`，测试集平均成本约 **9.5468**；整体上在**高服务水平**下 Integrated 未稳定优于 SAA（见后文综合曲线），提示需进一步扩展外生变量与调优网格。

Level 2 决策可视化示例　选取某一产品在一段时间内展示**真实需求**、**ML 预测**与 **SAA 订货量**。绿色与红色区域分别表示**超储**与**缺货**成本，有助于直观理解 Level 2 的库存决策效果。

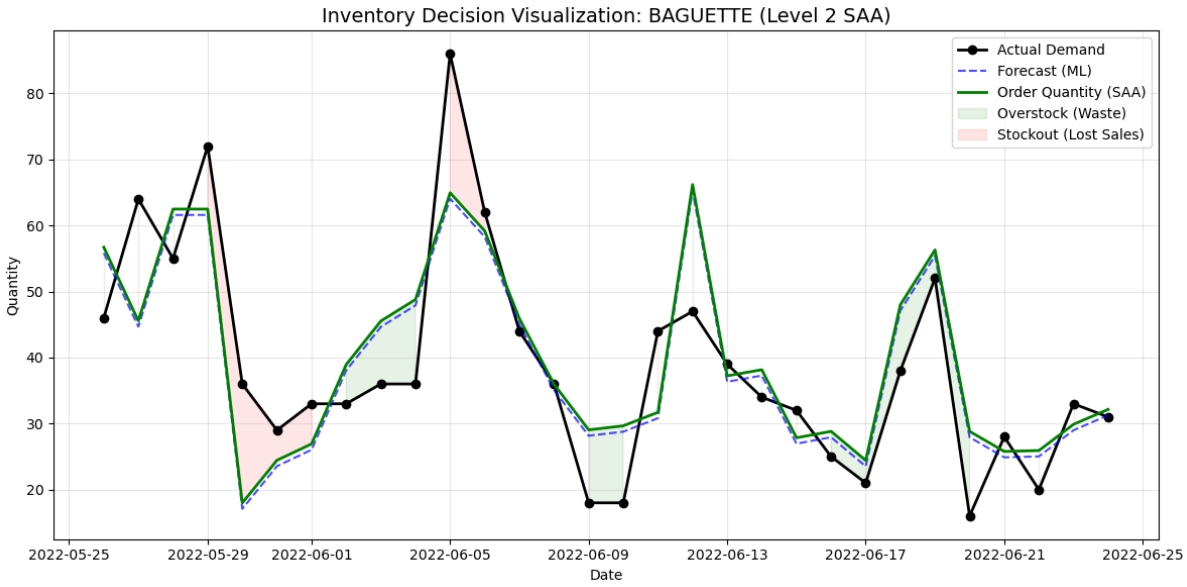

图 1: Level 2 决策可视化：真实需求、预测与 SAA 订货量

Level 3 简要说明与综合成本曲线 Level 3 采用**分位数回归**直接学习订货量 $q(x)$。下图展示 **Normal/SAA/Integrated** 在不同服务水平下的平均成本对比，其中**绿色曲线**为 Level 3 的 Integrated。结合当前特征与数据规模，Integrated 在高服务水平未稳定优于 SAA，提示需进一步扩展外生变量与调优网格。

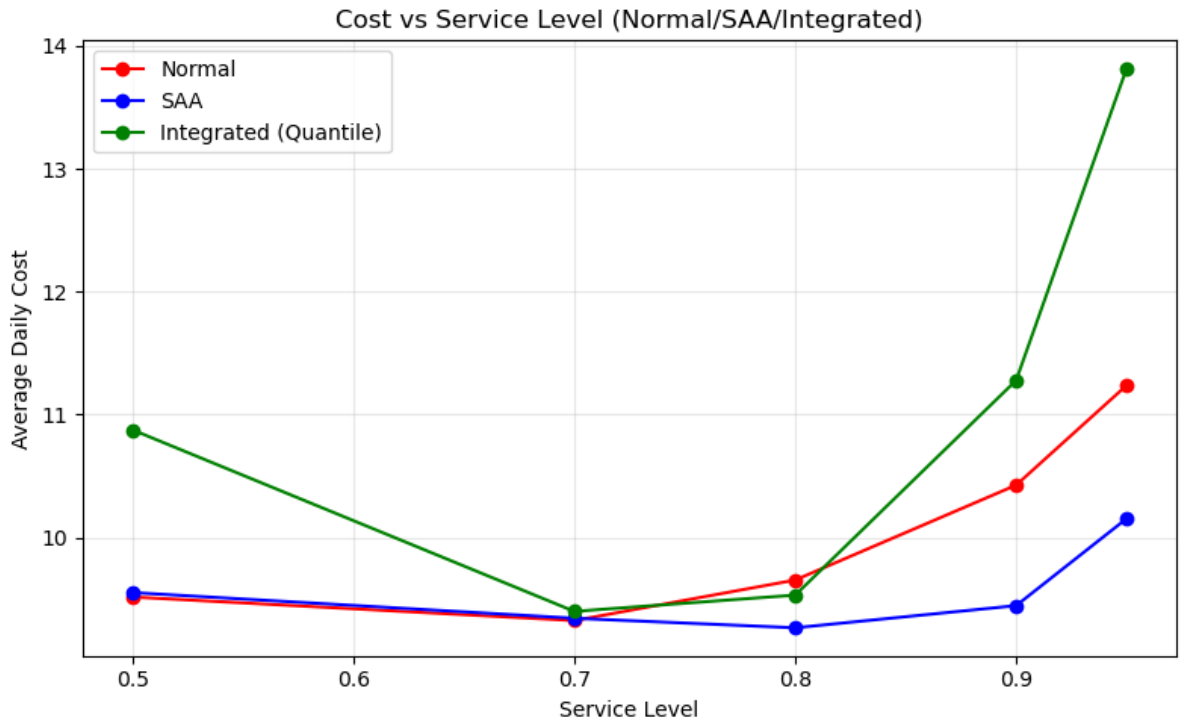

图 2: Level 2/3 综合：多服务水平平均成本对比（Normal/SAA/Integrated）

5 Understanding, Comments and Thinking

5.1 统计学视角的洞察

- **Prediction ≠ 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（Integrated Approach）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。

5.2 局限与改进

- **删失数据 (Censored Data)**：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。