

K.M. Shapen¹, A.T. Rakhymova^{1*}, A. Akhmedov², Zh.D. Baishemirov^{3,4}

¹L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

²Dallas College, Dallas, USA

³Kazakh-British Technical University, Almaty, Kazakhstan

⁴Narxoz University, Almaty, Kazakhstan

*e-mail: aigerim_rakhimova@mail.ru

BAYESIAN CAUSAL DEMAND MODELING WITH LATENT RANKING FACTORS IN E-COMMERCE

Abstract

E-commerce demand forecasting is complicated by the effects of latent ranking factors and deliberate price changes that distort observed sales relative to true demand. This paper presents a systematic review of modern approaches to demand modeling from the perspective of integrating probabilistic forecasting, Bayesian uncertainty assessment, and causal analysis. Three key areas are considered: probabilistic deep time series models (DeepAR, Temporal Fusion Transformer, N-BEATS, PatchTST, foundation models Chronos and TimesFM); Bayesian neural networks and ensemble methods for uncertainty assessment; and methods for causal modeling of price effects, including double machine learning and structural causal models. A comparative analysis of methods is conducted based on the criteria for accounting for uncertainty, ranking, and causality. A fundamental gap is identified: the absence of architectures that jointly account for Bayesian uncertainty, position bias, and causal price effects under conditions of unobservable product visibility. A direction for further research is proposed related to the development of integrated Bayesian-causal models with latent ranking variables for reconstructing true demand in e-commerce conditions.

Keywords: demand forecasting, e-commerce, Bayesian neural networks, causal modeling, position bias.

Қ.М. Шапен¹, А.Т. Рахымова¹А, А. Ахмедов², Ж.Д. Байшемиров^{3,4}

¹Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан

² Даллас колледжі, Даллас, АҚШ

³Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан

⁴Нархоз университеті, Алматы қ., Қазақстан

ЭЛЕКТРОНДЫҚ КОММЕРЦИЯДА ЛАТЕНТТІ ФАКТОРЛАРЫ БАР РАНЖИРЛЕУДІ ЕСКЕРЕТІН СҰРАНЫСТЫ БАЙЕСТІК СЕБЕП-САЛДАРЛЫҚ МОДЕЛДЕУ

Аңдатпа

Электрондық коммерцияға сұранысты болжау жасырын рейтинг факторларының және бақыланатын сатуды шынайы сұранысқа қатысты бұрмалайтын бағаның өзгеруінің әсерінен күрделене түседі. Бұл мақалада ықтималдық болжауды, Байес белгісіздік бағалауын және себеп-салдарлық талдауды біріктіру тұрғысынан сұранысты моделдеуге заманауи тәсілдердің жүйелі шолуы ұсынылған. Үш негізгі бағыт қарастырылады: ықтималдық терең уақыт қатарлары моделдері (DeepAR, Temporal Fusion Transformer, N-BEATS, PatchTST, Chronos және TimesFM негіздік моделдері); Байес нейрондық желілері және белгісіздікті бағалауға арналған ансамбльдік әдістер; баға әсерлерін себептік моделдеу әдістері, соның ішінде Double Machine Learning және құрылымдық себептік моделдер. Әдістердің салыстырмалы талдауы белгісіздік, рейтинг және себеп-салдарлыққа байланысты есепке алу критерийлеріне негізделіп жүргізілген. Байес белгісіздікті, позициялық ауытқуды және тауардың бақыланбайтын көрінгіштігі жағдайында бағаның себеп-салдарлық әсерлерін бірге ескеретін архитектуралардың болмауы - негізгі іргелі кемшілік болып анықталды. Электрондық коммерция жағдайында шынайы сұранысты анықтау үшін жасырын рейтингтік айнымалылары бар интеграцияланған Байес себеп-салдарлық моделдерді әзірлеуге қатысты әрі қарай зерттеу бағыттары ұсынылады.

Түйін сөздер: сұранысты болжау, электрондық коммерция, байес нейрондық желілері, себеп-салдарлық моделдеу, позициялық бейімділік.

К.М. Шапен¹, А.Т. Рахымова¹, А. Ахмедов², Ж.Д. Байшемиров³

¹ Евразийский национальный университет имени Л.Н. Гумилева, г.Астана, Казахстан

² Даллас колледж, Даллас, США

³ Казахстанско-Британский технический университет, г.Алматы, Казахстан

⁴ Университет Нархоз, г. Алматы, Казахстан

БАЙЕСОВСКО-КАУЗАЛЬНОЕ МОДЕЛИРОВАНИЕ СПРОСА С УЧЁТОМ ЛАТЕНТНЫХ ФАКТОРОВ РАНЖИРОВАНИЯ В ЭЛЕКТРОННОЙ КОММЕРЦИИ

Аннотация

Прогнозирование спроса в электронной коммерции осложняется действием латентных факторов ранжирования и целенаправленных изменений цен, искажающих наблюдаемые продажи относительно истинного спроса. Настоящая работа представляет систематический обзор современных подходов к моделированию спроса с позиций интеграции вероятностного прогнозирования, байесовской оценки неопределённости и каузального анализа. Рассмотрены три ключевых направления: вероятностные глубокие модели временных рядов (DeepAR, Temporal Fusion Transformer, N-BEATS, PatchTST, foundation models Chronos и TimesFM); байесовские нейронные сети и ансамблевые методы оценки неопределённости; методы причинного моделирования ценовых эффектов, включая Double Machine Learning и структурные причинные модели. Проведён сравнительный анализ методов по критериям учёта неопределённости, ранжирования и каузальности. Выявлен фундаментальный пробел - отсутствие архитектур, совместно учитывающих байесовскую неопределённость, position bias и причинные эффекты цены в условиях ненаблюдаемой видимости товара. Предложено направление дальнейших исследований, связанное с разработкой интегрированных байесовско-каузальных моделей с латентными переменными ранжирования для восстановления истинного спроса в условиях электронной торговли.

Ключевые слова: прогнозирование спроса, электронная коммерция, байесовские нейронные сети, каузальное моделирование, position bias.

Introduction

E-commerce is characterized by highly dynamic demand and limited observability of key sales drivers. Unlike traditional retail, where the seller independently determines product assortment, placement, and pricing, in e-commerce, the key driver of sales volume is product placement in search results, determined by the platform's algorithm and not directly observable. In these conditions, the central management task becomes inventory planning and pricing based on historical sales, which represent a distorted realization of true customer preferences. A fundamental feature of e-commerce data is the systematic discrepancy between observed demand and true demand. Observed sales y_t are formed as the product of latent demand d_t and latent visibility v_t , where v_t depends on the product position, competition, and the parameters of the ranking algorithm. In scientific literature, this phenomenon is described using the terms position bias and exposure bias. Position bias reflects the empirically established decrease in the likelihood of a click and purchase as a product moves away from the top search results, regardless of its relevance. Exposure bias means that products with low visibility do not receive enough impressions to realize their potential, resulting in their historical sales being systematically underestimated relative to actual demand [1-4].

For forecasting and decision-making purposes, latent visibility acts as an unobservable hidden factor. Traditional time series models interpret a decline in sales as a drop in demand, while in reality it may be due to a change in the ranking algorithm, increased competition, or a redistribution of traffic between categories. Ignoring this factor leads to systematic errors in assessing trends, seasonality, and reactions to promotions, as well as to a decrease in the generalizing ability of models during periods of structural changes in the platform. Existing approaches to demand forecasting only partially address this problem. Classical statistical methods, including ARIMA, ETS, and Prophet, provide interpretable results on stationary series and are widely used due to their computational simplicity. However, they are primarily focused on obtaining point forecasts and do not provide a reliable estimate of the uncertainty needed to calculate safety stocks in highly volatile environments [5].

The development of deep learning has led to the emergence of probabilistic architectures capable of modeling the full predictive distribution. The DeepAR model implements autoregressive probabilistic forecasting based on recurrent networks, the Temporal Fusion Transformer integrates an attention mechanism that accounts for static and dynamic covariates, the N-BEATS and N-HiTS architectures enable series decomposition without manual feature engineering, and PatchTST demonstrates effectiveness on long sequences thanks to its patch representation. The advent of foundation models, including Chronos and TimesFM, has enabled the use of pretraining on large time series corpora to improve the robustness of forecasts with limited observation history [6-11].

Despite the progress made in the accuracy of probabilistic forecasts, these models retain a common methodological premise: observed sales are considered a direct manifestation of demand. Price, promotional activity, and search ranking are introduced as standard correlation factors without distinguishing their causal nature. As a result, the models reproduce the observed patterns, but do not allow for a correct assessment of the consequences of targeted price changes, since they do not take into account the indirect influence of price on visibility through the ranking mechanism [12].

Of particular interest in this regard are Bayesian machine learning methods, which allow for quantitative assessment of prediction uncertainty and account for incomplete observed information. The use of Bayesian inference with variational approximations and ensemble methods provides a basis for joint modeling of latent demand, uncertainty parameters, and hidden visibility factors. Dropout-based approaches, as a Bayesian approximation, and deep ensembles enable scalable estimation of uncertainty associated with insufficient model knowledge in neural network architectures [13, 14].

Causal analysis methods based on the framework of structural causal models, directed acyclic graphs, and do-calculus provide a formal toolkit for estimating the effects of interventions. The Double Machine Learning approach provides consistent estimates for high-dimensional latent factors, provided they are fully observable. However, the application of these methods in e-commerce is limited because the key latent variable -product visibility- is latent and depends on unobservable parameters of the ranking algorithm [15-17].

To account for unobservable hidden factors, causal inference methods using proxy variables have been developed in recent years. These approaches utilize auxiliary proxy variables and bridge functions to identify causal effects in the presence of hidden factors. In e-commerce, proxy-based methods allow for reconstructing the impact of visibility based on external position measurements, impression data, and other indirect signals, creating the conditions for accurately estimating price elasticity in latent rankings [18, 19].

Thus, in the current literature, there is a significant gap between three approaches: probabilistic time series forecasting, Bayesian uncertainty assessment, and causal modeling of control actions. None of the existing approaches simultaneously accounts for probabilistic uncertainty, latent ranking factors, and the causal effects of pricing decisions. From a methodological perspective, the demand forecasting problem is transformed into the problem of reconstructing latent demand from indirect observations under conditions of uncertainty and structural causal dependencies [20].

This paper aims to systematize these areas and develop a unified theoretical model for demand analysis under conditions of hidden ranking and targeted price changes. The paper sequentially examines probabilistic deep models, Bayesian uncertainty assessment methods, approaches to causal modeling, and methods for accounting for positional bias. It then formulates a research agenda aimed at developing integrated Bayesian-causal architectures with latent visibility variables.

Research methodology

Classical demand forecasting methods

Historically, retail demand forecasting has relied on statistical time series models developed for conditions of relatively stable demand and full observability. ARIMA models and their seasonal extensions, SARIMA, formalize the dependence of the current value on past observations and forecast errors. Their strength lies in their interpretability and theoretical validity for stationary processes. In

e-commerce, these assumptions are often violated due to promotions, changes in ranking algorithms, the emergence of new competitors, and other factors that lead to structural breaks in the time series.

Exponential smoothing in ETS variants provides more flexible adaptation to trends and seasonality and is widely used due to its computational simplicity. Prophet is an additive time series model designed for automated accounting of trends, seasonality, and calendar effects. These models remain popular for practical applications due to their ease of implementation and low computational costs, but they are primarily focused on producing point forecasts and do not provide a reliable estimate of uncertainty.

A separate direction is associated with the use of gradient boosting, in particular XGBoost and LightGBM. With proper feature preparation, including previous sales values, moving averages, and calendar effects, these methods often outperform classical statistics in accuracy. Their limitation is the lack of a built-in probabilistic inference mechanism. Confidence intervals are constructed retroactively using quantile regression, which does not solve the problem of separating random and model uncertainty.

From a methodological perspective, the key limitation of the approaches considered is the assumption that observed sales represent an unbiased realization of demand $y_t = d_t + \varepsilon_t$. In e-commerce, this assumption is not met, since sales are distorted by visibility effects $y_t = d_t \cdot v_t$. The likelihood of purchase depends not only on consumer preferences but also on algorithmic selection and ranking mechanisms, which are described in the literature as position bias and exposure bias. As a result, models are trained on ranking noise, which reduces their robustness to changes in the distribution algorithms.

A comparative table illustrating the capabilities of classical methods is given below (Table 1.).

Table 1. Comparative analysis of classical methods

Method	Considers seasonal effects	Probabilistic forecast	Incorporates ranking	Causal interpretability	Interpretability
ARIMA	Yes	No	No	No	High
ETS	Yes	No	No	No	High
Prophet	Yes	Partially	No	No	Medium
XGBoost	Partially	No	No	No	Low

The table illustrates that none of the classical approaches simultaneously provides probabilistic estimation, accounting for ranking effects, and the possibility of causal interpretation. This circumstance necessitates a transition to probabilistic deep models and, further, to Bayesian and latent architectures capable of modeling v_t as a latent variable, as discussed in the following sections.

Deep Learning Models for Time Series

The development of deep learning has led to the emergence of a new generation of models for probabilistic forecasting and working with heterogeneous e-commerce data. Unlike classical approaches, these architectures are capable of training on thousands of product series simultaneously and accounting for complex nonlinear relationships.

The first widely adopted solution was DeepAR, which implements probabilistic forecasting based on recurrent networks. The model is trained to predict the parameters of the distribution of future sales, which allows for obtaining quantile estimates necessary for calculating safety stock. The most commonly used distributions are the negative binomial or Gaussian distributions, depending on the nature of the forecast series. Thanks to its probabilistic forecasting and ability to scale to a large number of product series, DeepAR has become one of the most widely used architectures in demand forecasting problems [6].

Further development involves the Temporal Fusion Transformer, which combines an attention mechanism with explicit consideration of static product characteristics and time-varying factors such

as price and promotional activity. A key advantage of TFT is its relative interpretability compared to other deep architectures, as the attention mechanism allows for the analysis of both feature importance and the time intervals that have the greatest impact on prediction [7].

The N-BEATS and N-HiTS architectures offer an alternative approach based on deep time series decomposition through stacks of fully connected blocks. N-BEATS provides automatic trend and seasonality separation without manual feature engineering, while N-HiTS complements this principle with hierarchical interpolation to improve accuracy over long horizons. Both models demonstrate high performance with limited external data [8].

A significant shift occurred with the advent of transformer solutions for long sequences. PatchTST implements a series partitioning into patches followed by the application of a self-attention mechanism, which allows for the effective capture of long-term dependencies and outperforms recurrent architectures on data with pronounced seasonality [9].

At the same time, the Time Series Foundation Models (TSFM) movement is being developed, represented by the Chronos, TimesFM, and Lag-Llama models. Unlike the traditional approach, where a model is trained on data from a single company, foundation models utilize knowledge extracted from a large number of heterogeneous time series, which allows for increased forecast robustness in the face of limited observation history [10, 11].

A particularly promising approach is combining time series with textual and visual representations of a product. Information from reviews, descriptions, and images allows for modeling hidden product quality characteristics that are not directly reflected in sales history. Thus, multimodal models partially address the problem of reconstructing latent demand.

It should be noted that the development of deep models has been accompanied by a significant increase in forecasting accuracy and the emergence of probabilistic estimates. However, the central object of forecasting has remained unchanged - most architectures are still focused on modeling observed sales y_t . The issue of separating true demand d_t from visibility v_t and ranking effects remains largely unresolved.

Despite progress, modern deep learning models treat price, stock, and search ranking as mere correlated features, ignoring their causal nature and the influence of hidden ranking factors. As a result, models are trained on data systematically skewed by position and exposure bias [1-3].

Thus, the development of deep learning has provided significant progress in forecasting observed sales, but has not resolved the fundamental problem of identifying true demand in the context of algorithmically driven product visibility and latent ranking factors. This limitation necessitates a transition from deterministic deep architectures to probabilistic models with explicit representations of uncertainty and latent variables, discussed in the next section.

Bayesian Approaches to Uncertainty Estimation

The transition from point forecasts to probabilistic estimates is a key development in e-commerce demand models. For inventory management tasks, knowing expected sales is insufficient; a quantitative assessment of the risk of stockouts and excess inventory is required. Bayesian machine learning methods allow for the formal consideration and quantification of various sources of forecast uncertainty.

It is common to distinguish between random and model components of uncertainty. Let y_t be sales at time t , and x_t be covariates (price, promotion). The predictive distribution is decomposed as:

$$p(y_t|x_t, \mathcal{D}) = \int p(y_t|x_t, \omega) p(\omega|\mathcal{D}) d\omega \quad (1)$$

where $p(y_t|x_t, \omega)$ reflects random noise, and $p(\omega|\mathcal{D})$ is the uncertainty of the parameters ω given $\mathcal{D}$.

The first reflects the unavoidable variability of consumer behavior, the second is associated with a lack of data, which is especially important for new products and periods of change in ranking algorithms.

The basic tool is Bayesian neural networks, in which weights are represented by distributions. Training is achieved through variational inference with optimization of the lower variational bound:

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\omega)}[\log p(\mathcal{D}|\omega)] - \text{KL}(q_{\theta}(\omega)||p(\omega)) \quad (2)$$

where $\mathcal{L}(\theta)$ is the variational lower bound of the marginal likelihood logarithm to be maximized with respect to the parameters θ ;

θ is the parameter of the variational approximation;

ω is the vector of model parameters;

$q_{\theta}(\omega)$ is the variational distribution approximating the true posterior distribution $p(\omega|\mathcal{D})$;

$\mathbb{E}_{q_{\theta}(\omega)}[\cdot]$ is the mathematical expectation for the distribution $q_{\theta}(\omega)$;

$p(\mathcal{D}|\omega)$ is the likelihood function of the training data $\mathcal{D}$ with fixed parameters ω ;

$\log p(\mathcal{D}|\omega)$ is the logarithm of the likelihood, characterizing the quality of the data approximation;

$p(\omega)$ is the prior distribution of the model parameters;

$\text{KL}(q_{\theta}(\omega)||p(\omega))$ is the Kullback-Leibler divergence between the variational distribution and the prior, acting as a regularizer.

The first term corresponds to the expected log-likelihood and ensures consistency with the data, the second term penalizes deviation from a priori assumptions and controls for model complexity.

In practice, the MC Dropout method, interpreted as approximate Bayesian inference, has become widely used. Instead of a normal approximation, empirical quantiles of the predictive distribution are used, which are more robust to asymmetric sales distributions.

With T stochastic passes, is constructed a quantile interval:

$$\begin{aligned} \hat{Q}_{\alpha/2} &= \text{Quantile}_{\alpha/2}(\{\hat{y}_t^{(i)}\}), \hat{Q}_{1-\alpha/2} = \text{Quantile}_{1-\alpha/2}(\{\hat{y}_t^{(i)}\}) \\ C_t^{1-\alpha} &= [\hat{Q}_{\alpha/2}, \hat{Q}_{1-\alpha/2}] \end{aligned} \quad (3)$$

where $\hat{Q}_{\alpha/2}$ is the estimate of the lower quantile of level $\alpha/2$ of the predictive distribution;

$\hat{Q}_{1-\alpha/2}$ is the estimate of the upper quantile of level $1 - \alpha/2$ of the predictive distribution;

$\text{Quantile}_p(\cdot)$ is the operator for calculating the empirical quantile of level p ;

$\{\hat{y}_t^{(i)}\}$ is the set of forecast samples for time t obtained by the Monte Carlo method, $i = \overline{1, N}$;

$\hat{y}_t^{(i)}$ is the i -th forecast realization corresponding to the sample of parameters $\omega^{\{i\}} \sim q_{\theta}(\omega)$;

α is the significance level, $\alpha \in (0, 1)$;

$1 - \alpha$ is the specified confidence level of the predictive interval;

$C_t^{1-\alpha}$ is the two-sided predictive interval of confidence level $1 - \alpha$ for the target variable y_t at time t ;

N is the number of Monte Carlo samples

A common alternative is to use deep ensembles of M independently trained networks:

$$p(y_t|x_t) \approx \frac{1}{M} \sum_{m=1}^M p(y_t|x_t, \omega_m) \quad (4)$$

where $p(y_t|x_t)$ is the predictive distribution of the target variable y_t given the features x_t ;

M is the number of Monte Carlo samples;

$\omega_m \sim q_{\theta}(\omega)$ is the m -th sample of the model parameters from the variational distribution;

$p(y_t|x_t, \omega_m)$ is the conditional distribution of the output given the fixed parameters ω_m ;

x_t is the vector of input features at time t ;

y_t is the predicted value.

For time series forecasting, Bayesian DeepAR and Bayesian TFT have been proposed. In DeepAR, the network models the parameters of the sales distribution:

$$y_t \sim \text{NegBin}(\mu_\theta(h_t), \sigma_\theta^2(h_t)), \quad h_t = f_\theta(y_{1:t-1}, x_{1:t}) \quad \theta \sim q_\theta \quad (5)$$

where y_t is the target variable at time t ;

$\mu_\theta(h_t), \sigma_\theta^2(h_t)$ are the mean and variance calculated by the network from the hidden state h_t ;

h_t is the hidden state;

f_θ is the recurrence function (LSTM/GRU);

$y_{1:t-1}$ are the past values of y

$x_{1:t}$ are the past and current features;

$\theta \sim q_\theta$ are the parameters sampled from the variance distribution.

Or in general terms

$$y_t \sim p_\theta(y_t | x_t) \quad (6)$$

where $y_t \sim p_\theta(y_t | x_t)$ is the target variable y_t that follows the parametric distribution p_θ given x_t ;

p_θ is a family of distributions;

θ are the distribution parameters, $\theta \sim q_\theta$;

x_t is generally the entire context: $x_t = \{y_{1:t-1}, x_{1:t}\}$;

$h_t = f_\theta(y_{1:t-1}, x_{1:t})$ is the deterministic latent representation of the context.

For inventory control, the quantile of the predictive distribution on the supply horizon L is used. The safety stock is defined as:

$$s_t = Q_\tau(y_{t:t+L} | x_t) - \mathbb{E}[y_{t:t+L} | x_t] \quad (7)$$

where s_t is the measure of forecast asymmetry/bias at time t ;

$Q_\tau(y_{t:t+L} | x_t)$ is the τ -quantile of the predictive distribution of the sum (or trajectory) $y_{t:t+L}$ given x_t ;

$\tau \in (0,1)$ is the quantile level ($\tau=0.9$ or 0.95 for the right tail);

$y_{t:t+L} = \sum_{k=0}^L y_{t+k}$ or the vector $(y_t, \dots, y_{t+L})$ is the forecast horizon of length L ;

$\mathbb{E}[y_{t:t+L} | x_t]$ is the mathematical expectation of the same predictive distribution;

x_t is the context at time t .

To calibrate confidence intervals, prediction methods with guaranteed coverage are used, providing finite-sample coverage guarantees at the $1 - \alpha$ level, provided that the order of observations does not affect their distribution:

$$P(y_{t+1} \in C_t(x_{t+1})) \geq 1 - \alpha \quad (8)$$

In the context of marketplaces, Bayesian inference allows one to treat true demand d_t as a latent variable. Observed sales are modeled taking into account latent visibility:

$$y_t \sim p(y_t | d_t, v_t), \quad d_t \sim p(d_t | x_t, w_d), \quad v_t \sim p(v_t | p_t, r_t, w_v) \quad (9)$$

where v_t characterizes the probability of displaying a product to a user and depends on platform ranking factors, including price – p_t and rating – r_t . The joint posterior is approximated through variational inference:

$$p(d_{1:T}, v_{1:T}, w_d, w_v | y_{1:T}) \propto \prod_{t=1}^T p(y_t | d_t, v_t) p(d_t | x_t, w_d) p(v_t | p_t, r_t, w_v) p(w_d) p(w_v) \quad (10)$$

where, $d_{1:T} = (d_1, \dots, d_T)$ is the latent demand trajectory for period $1:T$;

$v_{1:T} = (v_1, \dots, v_T)$ is the latent supply/visibility trajectory for period $1:T$;

w_d, w_v are the parameter vectors of the supply and demand models;

$y_{1:T}$ are the observed data (sales) for period 1: T ;

$p(y_t|d_t, v_t)$ is the likelihood of the observation at time t given demand d_t and supply v_t ;

$p(d_t|x_t, w_d)$ is the demand model dependent on features x_t and parameters w_d ;

$p(v_t|p_t, r_t, w_v)$ is the supply model, depending on the price p_t , rating/relevance r_t , and parameters w_v ;

$p(w_d)p(w_v)$ are the prior distributions of the parameters;

x_t are the exogenous demand features at time t ;

p_t is the price at time t ;

r_t is the rating or visibility measure at time t .

Joint estimation allows us to separate demand variation due to customer preferences from sales variation caused by ranking algorithms.

Thus, Bayesian approaches provide a quantitative assessment of forecast uncertainty and enable modeling of hidden factors influencing observed sales. This provides the foundation for moving from sales forecasting to true demand assessment and forms the methodological basis for the further application of latent and causal models.

Causal Modeling of Price Effects with Latent Ranking

Standard forecasting models estimate the conditional distribution $p(y_t|p_t, x_t)$. Price control requires an intervention distribution $p(y_t|do(p_t), x_t)$. The difference stems from the causal structure of the marketplace, where price influences sales across two channels.

The structural causal model is formalized with an explicit separation of the intervening variable and the common factor. Let u_t be the latent platform factor (e.g., a seasonal category trend, a marketplace promotional campaign), and v_t be the product visibility:

$$\begin{aligned} u_t &\rightarrow p_t, u_t \rightarrow v_t, u_t \rightarrow y_t, \\ p_t &\rightarrow d_t, p_t \rightarrow v_t, d_t \rightarrow y_t, v_t \rightarrow y_t, \end{aligned} \quad (11)$$

Here v_t is a mediator of the price effect through ranking, and u_t is an unobserved latent factor that creates a spurious correlation between price and sales.

To estimate the causal effect in a partially linear setting, the model is used:

$$y_t = \theta_0 p_t + g(x_t, u_t) + \varepsilon_t \quad (12)$$

where θ_0 represents the average price elasticity. When u_t is unobservable, the standard estimate is biased. Chernozhukov's Double Machine Learning method allows one to obtain an unbiased estimate under the condition that hidden factors are observable [16]. Two-step procedure:

$$\begin{aligned} \hat{l}(x_t) &= \mathbb{E}[p_t|x_t], \hat{m}(x_t) = \mathbb{E}[y_t|x_t] \\ \hat{\theta} &= \left(\frac{1}{n} \sum_t (p_t - \hat{l}_t)^2 \right)^{-1} \frac{1}{n} \sum_t (p_t - \hat{l}_t)(y_t - \hat{m}_t) \end{aligned} \quad (13)$$

In the presence of latent u_t this condition is violated.

To identify unobservable hidden factors, methods of proximal causal inference are used. Let z_t be a proxy variable for v_t (e.g., an external position estimate), and w_t be a proxy for u_t . Within the Miao-Tchetgen Tchetgen approach, a bridge function h is introduced that satisfies the integral equation [18]:

$$\mathbb{E}[y_t|p_t, z_t] = \int h(p_t, w) h(w|z_t) dw \quad (14)$$

This equation is a Fredholm equation of the first kind. If a solution exists, the causal effect is identified as:

$$\mathbb{E}[y_t|do(p^*)] = \int h(p^*, w) p(w) dw \quad (15)$$

Under the Bayesian causal framework with Pearl's do-operator, truncated factorization is applied. The joint model [15]:

$$p(y_t, d_t, v_t, u_t | p_t) = p(y_t | d_t, v_t, u_t) p(d_t | p_t, u_t) p(v_t | p_t, u_t) p(u_t) \quad (16)$$

Following the $do(p^*)$ intervention, the distribution is:

$$p(y_t | do(p^*)) = \iiint p(y_t | d_t, v_t, u_t) p(d_t | do(p^*), u_t) p(v_t | do(p^*), u_t) p(u_t) dd_t dv_t du_t \quad (17)$$

where $p(d_t | do(p^*), u_t) = p(d_t | p^*, u_t)$ and $p(v_t | do(p^*), u_t) = p(v_t | p^*, u_t)$ according to the truncated factorization rule. The expected effect:

$$\mathbb{E}(y_t | do(p^*)) = \iiint \mathbb{E}[y_t | d_t, v_t, u_t] p(d_t | p^*, u_t) p(v_t | p^*, u_t) p(u_t) dd_t dv_t du_t \quad (18)$$

This formula clearly separates the direct effect $p^* \rightarrow d_t$, the mediation effect $p^* \rightarrow v_t \rightarrow y_t$ and the influence of the latent hidden factor u_t .

Thus, correct causal modeling requires distinguishing between the ranking mediator and the latent underlying platform factor. Proximal inference methods enable identification in the presence of proxy variables, while the Bayesian formulation allows for the integration of parameter uncertainty assessment with causal identification.

Latent Models for Ranking and Visibility

Observed sales y_t are the result of the interaction of true demand d_t and algorithmic visibility v_t . The seller does not directly observe v_t , which leads to the problem of reconstructing hidden states from distorted observations.

The basic generative model is formulated as a hierarchical system:

$$d_t \sim p(d_t | x_t, w_d), v_t \sim p(v_t | p_t, r_t, u_t, w_v), y_t \sim p(y_t | d_t, v_t, w_y) \quad (19)$$

where u_t is the platform's latent factor. For counted sales, a visibility-controlled distribution is used:

$$y_t \sim \text{Poisson}(d_t \cdot v_t) \text{ or } y_t \sim \text{NegBin}(d_t \cdot v_t, \phi) \quad (20)$$

Visibility is modeled as a latent random variable on the interval $(0, 1)$. A beta distribution with parameters dependent on the ranking factors is used:

$$\begin{aligned} \text{logit}(\mu_{v,t}) &= \alpha_p \log p_t + \alpha_r r_t + \beta^T u_t \\ v_t &\sim \text{Beta}(a_t, b_t), a_t = \mu_{v,t} \tau, b_t = (1 - \mu_{v,t}) \tau \end{aligned} \quad (21)$$

where τ is the concentration parameter, the concentration parameter, $\alpha_p < 0$ reflects the penalty for a high price. Alternative parameterization via the logit-normal distribution:

$$\text{logit}(v_t) \sim \mathcal{N}(\alpha_p \log p_t + \alpha_r r_t + \beta^T u_t, \sigma_v^2) \quad (22)$$

Variational inference is used to estimate latent variables. Approximation of the posterior distribution:

$$q_\phi(d_{1:T}, v_{1:T}, u_{1:T}) = \prod_t q_{\phi_d}(d_t) q_{\phi_v}(v_t) q_{\phi_u}(u_t) \quad (23)$$

Variational lower bound on the log-likelihood for the latent ranking model:

$$\begin{aligned} \mathcal{L}(\phi, w) &= \\ &\mathbb{E}_q[\sum_t \log p(y_t | d_t, v_t)] - \text{KL}(q_{\phi_d} || p_{w_d}) - \text{KL}(q_{\phi_v} || p_{w_v}) - \text{KL}(q_{\phi_u} || p_u) \end{aligned} \quad (24)$$

Optimization is carried out using stochastic gradient descent with the reparameterization trick. To take into account the temporal dynamics, a state space model is introduced:

$$\begin{aligned} d_t &= p_d d_{t-1} + \eta_{d,t}, \text{logit}(v_t) = p_v \text{logit}(v_{t-1}) + \eta_{v,t}, \eta_{\cdot,t} \sim \mathcal{N}(0, \sigma^2) \\ y_t &\sim p(y_t | d_t, v_t) \end{aligned} \quad (25)$$

Hidden state filtering is performed using a particle filter or a variational Kalman filter. The posterior distribution:

$$p(d_t, v_t | y_{1:t}) \propto p(y_t | d_t, v_t) \int p(d_t, v_t | d_{t-1}, v_{t-1}) p(d_{t-1}, v_{t-1} | y_{1:t}) dd_{t-1} dv_{t-1} \quad (26)$$

To identify the model in the absence of direct observations v_t , proxy variables z_t and constraints on the structure are used:

$$z_t \sim p(z_t | v_t, \sigma_z^2), \mathbb{E}[v_t] = \bar{v} \quad (27)$$

where $\bar{v}$ is given by a priori knowledge of the average visibility of the category.

Integration with Bayesian forecasting is achieved through a joint predictive distribution:

$$p(y_{t+1} | y_{1:t}) = \int p(y_{t+1} | d_{t+1}, v_{t+1}) p(d_{t+1}, v_{t+1} | y_{1:t}) dd_{t+1} dv_{t+1} \quad (28)$$

This formula allows us to build forecasts that take into account both demand uncertainty and future visibility uncertainty.

Thus, latent ranking modeling enables the reconstruction of true demand from distorted sales observations. The use of hierarchical Bayesian models with time dependencies and proxy variables provides the basis for the correct estimation of price elasticity and inventory management under algorithmic visibility management.

Combining Bayesian and Causal Approaches to Price and Stock Management

Sections 4-6 comprise three independent components: forecast uncertainty assessment, identification of causal price effects, and reconstruction of latent visibility. Marketplace management tasks require their combined use within a unified probabilistic inference framework.

The complete generative model integrates a hierarchical structure:

$$p(y_{1:T}, d_{1:T}, v_{1:T}, w) = \prod_{t=1}^T p(y_t | d_t, v_t) p(d_t | p_t, x_t, u_t, w_d) p(v_t | p_t, r_t, u_t, w_v) p(u_t) p(w) \quad (29)$$

where $w = \{w_d, w_v\}$.

Posterior distribution for observed sales:

$$p(d_{1:T}, v_{1:T}, u_{1:T}, w | y_{1:T}) \propto p(y_{1:T}, d_{1:T}, v_{1:T}, u_{1:T}, w) \quad (30)$$

Variational approximation:

$$q_\phi(d, v, u, w) = q_{\phi_w}(w) \prod_t q_{\phi_d}(d_t) q_{\phi_v}(v_t) q_{\phi_u}(u_t) \quad (31)$$

An intervention predictive distribution is used for decision making. When setting the price p^* :

$$p(y_{t+1} | do(p^*), y_{1:t}) = \int p(y_{t+1} | d_{t+1}, v_{t+1}) p(d_{t+1} | p^*, u_{t+1}) p(v_{t+1} | p^*, u_{t+1}) p(u_{t+1} | y_{1:t}) dd_{t+1} dv_{t+1} du_{t+1} \quad (32)$$

This distribution takes into account both the epistemic uncertainty of the parameters w and the stochastic variability of demand and visibility.

The optimal price is determined by maximizing expected utility taking into account risk:

$$p_t^* = \arg \max_p \{ \mathbb{E}[\pi(p, y) | do(p), \mathcal{D}] - \lambda \cdot \text{Var}[\pi(p, y) | do(p), \mathcal{D}] \} \quad (33)$$

where $\pi(p, y) = (p - c) \cdot \min(y, s) - h \cdot \max(0, s - y)$ is a profit function with cost c , inventory s and storage costs h . The parameter λ reflects the degree of risk aversion.

The intervention distribution quantile is used for inventory control:

$$s_t(p^*) = Q_\tau(y_{t:t+L} | do(p^*), y_{1:t}) \quad (34)$$

The joint optimization of price and stock is formulated as:

$$(p^*, s^*) = \arg \max_{p, s} \mathbb{E}_{p(y | do(p))} [\pi(p, y, s)] \quad (35)$$

with a service level constraint $P(y_{t:t+L} \leq s | do(p)) \geq \tau$.

For computation, inference is carried out by sampling from the variational posterior. The algorithm:

1. Sample $w^{(i)} \sim q_{\phi_w}, u_{t+1}^{(i)} \sim q_{\phi_u}$
2. For a given p^* generate $d_{t+1}^{(i)} \sim p(d | p^*, u^{(i)}, w_d^{(i)}), v_{t+1}^{(i)} \sim p(v | p^*, u^{(i)}, w_v^{(i)})$
3. Generate $y_{t+1}^{(i)} \sim p(y | u^{(i)}, v^{(i)})$
4. Estimate $\mathbb{E}[\pi]$ and $\text{Var}[\pi]$ from the sample $\{y^{(i)}\}_{i=1}^M$.

Conformal calibration is applied to the intervention distribution to provide finite-sample guarantees of coverage of forecast intervals when the pricing policy changes.

Thus, the integration of Bayesian inference, causal identification, and latent ranking modeling forms a methodological basis for decision-making in the face of algorithmic uncertainty in marketplaces. The proposed framework enables the transition from correlational sales forecasting to risk-adjusted optimization of true demand.

Modern Architectures and Research Directions

The integration of Bayesian, causal, and latent components has been implemented in several classes of models developed since 2020.

Probabilistic Transformers. The Temporal Fusion Transformer architecture and its Bayesian variants use an attention mechanism to model long-term dependencies [7]

$$h_t = \text{Attention}(Q_t, K_{1:t}, V_{1:t}), y_t \sim p_\theta(y_t | h_t) \quad (36)$$

The Bayesian version introduces a distribution on the attention parameters $W_Q, W_K, W_V \sim q_\phi$, which allows for estimating forecast uncertainty. For count data, the Negative Binomial or Zero-Inflated Poisson output is used.

Deep state-space models. Deep State Space Models use neural networks to parameterize transition and emission distributions [21]:

$$p(y_* | x_*, C) = \int p(y_* | x_*, z) q(z | C) dz \quad (37)$$

where context C is constructed from similar products. The Bayesian version provides uncertainty estimates for items without history.

Causal Transformers. DeepIV's Causal Transformer and EconML architectures integrate instrumental variables into the attention mechanism [22]:

$$\hat{p}_t = f_1(z_t, x_t), \hat{y}_t = f_2(\hat{p}_t, x_t) \quad (38)$$

where z_t is an instrument orthogonal to u_t . Training with orthogonal loss ensures unbiased learning for high-dimensional latent factors.

Principal Constraints of Existing Approaches:

1. Scalability of Bayesian inference. Full variational inference for millions of SKUs (articles) requires stochastic approximations with inducing points or amortized inference.
2. Latent visibility identifiability. Without external proxies z_t for v_t , the model suffers from non-identifiability. Weakly supervised methods based on position data are needed.
3. Non-stationarity of ranking algorithms. The w_t parameters change with platform updates, violating the exchangeability assumption for conformal prediction. Adaptive methods with distribution shift detection are needed.
4. Computational complexity of do-inference. Estimation $\mathbb{E}[y|do(p^*)]$ requires multiple sampling from the joint posterior distribution. For real-time, surrogate models and amortized causal inference are used.

Promising areas include the development of fundamental demand models pre-trained on data from multiple marketplaces with subsequent adaptation through Bayesian retraining, as well as the integration of differential privacy methods for learning on platform data without disclosing ranking algorithms.

Thus, the current state of research is characterized by a transition from isolated forecasting methods to integrated probabilistic-causal systems capable of functioning under conditions of algorithmic uncertainty.

Results of the study

The analysis demonstrates a systematic shift in demand forecasting methods from frequentist ARIMA and ETS models to deep probabilistic architectures. DeepAR, Temporal Fusion Transformer, N-BEATS, and PatchTST models provide scalability to thousands of series and form a baseline for probabilistic forecasts. The introduction of Bayesian methods has enabled a transition from point estimates to quantitative uncertainty assessment with separation of aleatory and epistemic components.

Discussion

A key limitation of existing approaches is the assumption of observability of the data generation process. In marketplaces, this assumption is violated due to algorithmic visibility management. Observed sales y_t are the product of true demand d_t and latent visibility v_t , where v_t depends on price through a ranking mechanism. Standard models interpret the decrease in y_t with an increase in price as a manifestation of price elasticity, ignoring the indirect effect $p_t \rightarrow v_t \rightarrow y_t$.

A comparative analysis shows the absence of architectures that simultaneously satisfy three requirements: Bayesian estimation of parameter uncertainty, accounting for latent ranking, and identification of the causal effects of price interventions. Double Machine Learning methods provide consistency with observable latent factors, but lose their properties in the presence of latent u_t . Proximal causal methods need valid proxies for v_t - a condition only partially met in real-world marketplaces.

Computational limitations are related to the scale of the problem. Full Bayesian inference for the joint distribution $p(d_{1:T}, v_{1:T}, u_{1:T}, w|y_{1:T})$ has complexity $O(T \cdot M)$, where M is the number of samples. Catalogs with millions of SKUs require amortized inference schemes and distributed computation. Additional complexity arises from the non-stationarity of the ranking algorithms, which violates the exchangeability assumption required for conformal guarantees.

Promising areas include the development of fundamental demand models with pre-training on multi-platform data, the integration of differential privacy methods for training without disclosing ranking algorithms, and the creation of a digital twin of demand capable of simulating the consequences of pricing and assortment decisions under conditions of incomplete information.

Conclusion

An analysis of modern literature demonstrates the evolution of e-commerce demand forecasting methods from classical statistical models to deep probabilistic architectures and causal approaches. DeepAR, Temporal Fusion Transformer, N-BEATS, PatchTST, and foundation models have improved the accuracy of probabilistic forecasts and created a foundation for scaling. Bayesian methods have made it possible to formalize uncertainty assessment and separate the aleatory and epistemic components of risk.

It has been established that the key feature of e-commerce platforms - algorithmically controlled product visibility - remains outside the scope of most existing approaches. Observed sales represent a distorted representation of true demand, dependent on an unobservable ranking factor. Standard models interpret changes in visibility as fluctuations in preferences, leading to systematic errors in estimating price elasticity.

A comparative analysis revealed the lack of a universal method that simultaneously provides Bayesian uncertainty assessment, latent ranking considerations, and causal modeling of price interventions. This circumstance identifies a promising research direction related to the development of integrated Bayesian-causal architectures with latent visibility variables.

The proposed research program includes joint demand and ranking modeling, the development of proximal inference methods for unobservable latent factors, and the creation of computationally efficient schemes for intervention forecasting. Implementation of these areas will enable a transition from forecasting observed sales to reconstructing true demand and assessing the consequences of management decisions under conditions of incomplete information.

Thus, further development of the field is associated not with increasing the accuracy of point forecasts, but with the construction of models capable of taking into account the causal structure of marketplace data and ensuring stable decision-making in conditions of algorithmic uncertainty.

References

- [1] Joachims T., Granka L., Pan B., et al. *Accurately interpreting clickthrough data as implicit feedback*. SIGIR Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005. 154-161. doi.org/10.1145/1076034.1076063
- [2] Craswell N., Zoeter O., Taylor M., Ramsey B. *An experimental comparison of click position-bias models*. Web Search and Data Mining, 2008. 87-94. DOI:10.1145/1341531.1341545
- [3] Wang X., Bendersky M., Metzler D., Najork M. *Learning to rank with selection bias in personal search*. SIGIR, 2016. 115-124. doi.org/10.1145/2911451.2911537
- [4] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. *Correcting for Selection Bias in Learning-to-rank Systems*. In *Proceedings of The Web Conference 2020 (WWW '20)*, April 20–24, 2020, Taipei, Taiwan. ACM, New York, USA. 11 Pages. DOI: <https://doi.org/10.1145/3366423.3380255>
- [5] Hyndman R. J., Athanasopoulos G. *Forecasting: principles and practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. Accessed: Dec. 20, 2025. <https://otexts.com/fpp3/>
- [6] Salinas D., Flunkert V., Gasthaus J., Januschowski T. *DeepAR: Probabilistic forecasting with autoregressive recurrent networks*. International Journal of Forecasting, vol. 36, issue 3, 1181-1191. DOI: 10.1016/j.ijforecast.2019.07.001
- [7] Lim B., Arik S.O., Loeff N., Pfister T. *Temporal Fusion Transformers for interpretable multi-horizon time series forecasting*. International Journal of Forecasting, vol. 37(4), 2021. 1748-1764. DOI: 10.1016/j.ijforecast.2021.03.012
- [8] Oreshkin B.N., Carпов D., Chapados N., Bengio Y. *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting*. ICLR, 2020. 31 Pages. <https://doi.org/10.48550/arXiv.1905.10437>
- [9] Nie Y., Nguyen N.H., Sinthong P., Kalagnanam J. *A time series is worth 64 words: Long-term forecasting with transformers*. ICLR, 2023. 1-22. doi.org/10.48550/arXiv.2211.14730
- [10] Ansari A.F., Stella L., Turkmen C., et al. *Chronos: Learning the language of time series*. Transactions on Machine Learning Research, 2024. 1-43. DOI:10.48550/arXiv.2403.07815

- [11] Abhimanyu D., Weihao K., Rajat S., Yichen Zh. A decoder-only foundation model for time-series forecasting. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 2024. 10148-10167. doi.org/10.48550/arXiv.2310.10688
- [12] Januschowski T., Gasthaus J., Wang Y., et al. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1), 2020. 167-177. DOI:10.1016/j.ijforecast.2019.05.008
- [13] Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, 2016. 1050-1059. doi.org/10.48550/arXiv.1506.02142
- [14] Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, USA 2017. doi.org/10.48550/arXiv.1612.01474
- [15] Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. *Econometric Theory*, 19, 2003, 675–685. DOI:10.1017/S0266466603004109
- [16] Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen Ch., Newey Wh., Robins J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, Volume 21, Issue 1, 1 February 2018, 1–68. <https://doi.org/10.1111/ectj.12097>
- [17] Bottou L., Peters J., Quiñero-Candela J., et al. Counterfactual reasoning and learning systems. *Journal of Machine Learning Research* 14 (2013). 3207-3260. doi.org/10.48550/arXiv.1209.2355
- [18] Miao W., Geng Z., Tchetgen Tchetgen E.J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 2016. DOI:10.1093/biomet/asy038
- [19] Eric J Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, Wang Miao. An introduction to proximal causal learning. *Statist. Sci.* 39(3), 2024. 375-390. DOI: 10.1214/23-STS911
- [20] Schölkopf B., Locatello F., Bauer S., Ke N.R., Kalchbrenner N., Goyal A., Bengio Y. Toward Causal Representation Learning. *Proceedings of the IEEE*, Vol. 109, No. 5, 2021. 612–634. doi.org/10.48550/arXiv.2102.11107
- [21] Rangapuram S.S., Seeger M.W., Gasthaus J., Stella L., Wang Y., Januschowski T. Deep State Space Models for Time Series Forecasting. *Advances in Neural Information Processing Systems*. Vol. 31, 2018. 7785-7794. DOI: https://proceedings.neurips.cc/paper_files/paper/2018/file/5cf68969fb67aa6082363a6d4e6468e2-Paper.pdf
- [22] Hartford J., Lewis G., Leyton-Brown K., Taddy M. Deep IV: A Flexible Approach for Counterfactual Prediction. *Proceedings of the 34th International Conference on Machine Learning*. 2017. 1414-1423. DOI: <https://proceedings.mlr.press/v70/hartford17a.html>