Causal Product Networks: A Data-driven Methodology for Modeling Basket-Shopping Consumer Behavior

(Authors' names blinded for peer review)

Discovering and measuring the causal effects of the purchase of one product on the purchases of other products in a shopping basket is an important problem in retailing. It is also extremely challenging because the number of possible interactions among all the products in a retail store can be exponentially large. We present a graph-theoretic causal product network (CPN) representation of consumer basket-shopping behavior and a causal structure learning-based approach to learn the network and recover causal effects among product purchases from basket-shopping data. To address the unique challenge of data sparsity in such datasets, we propose a Euclidean distance-based independence test instead of the partial correlation-based test in the PC algorithm for causal learning. We validate our approach and demonstrate its value by utilizing a large-scale basket-shopping dataset from Numerator. Our results show that CPNs more accurately represents interactions among products than other machine-learning or theory-based models, require significantly fewer parameters, and reveal new insights into consumer behavior such as that the brick-and-mortar channel exhibits denser causal connections among product purchases in a basket than the online channel. Finally, we propose a two-stage graph learning algorithm to improve computational efficiency and demonstrate that the constructed CPN, when applied to a multi-category assortment optimization problem, increases total sales by 8.26%–20.62% compared to a benchmark within-category model, viz., the multinomial logit model. These results make a significant advancement towards utilizing large-scale basket-shopping data for generating new managerial insights and optimizing decisions.

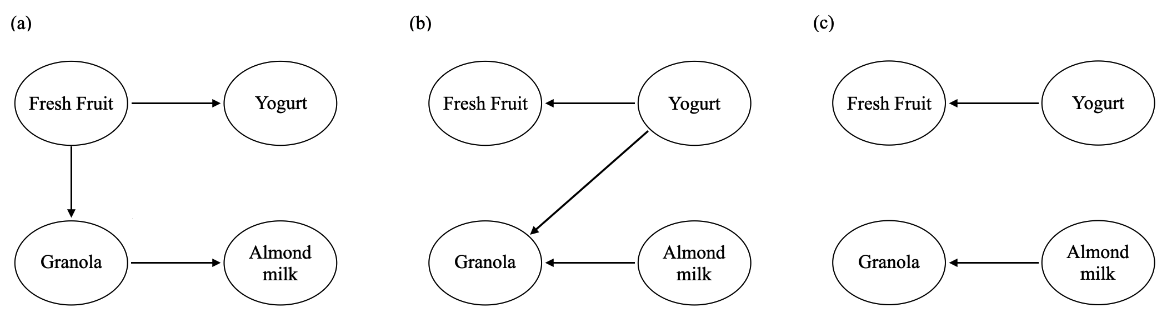
Key words: basket shopping, product network, assortment optimization, causal structure learning, empirical research

1. Introduction

Consider a customer purchasing multiple products across categories in a single shopping visit to a supermarket store. The study of such basket-shopping behavior is important for retailers for addressing different kinds of questions, such as: (i) utilizing the vast amount of historical basket purchase data, can we identify how buying one product *causally* affects the purchases of additional products, and (ii) using the causal interactions among product purchases, can we optimize retailing decisions, such as which products to include in the assortment strategy or what prices and promotion strategies to employ to maximize the total expected profit or sales.

We illustrate these questions in Figure 1 by showing different scenarios of interactions among four products: yogurt, fresh fruit, granola, and almond milk; the interactions are depicted as *causal* *product networks* (CPNs) where nodes represent products and directed edges denote causal relationships between their purchases. In Figure 1(a), fresh fruit purchase influences yogurt and granola purchases, with granola additionally inducing almond milk purchases. In contrast, Figure 1(b) suggests yogurt as the driver that influences fresh fruit and granola purchases, while almond milk directly impacts granola sales. Figure 1(c) features a sparser configuration, highlighting only the direct causal relationships between yogurt and fresh fruit, as well as between almond milk and granola. Given these differences, the assortment strategy must be tailored for each network: offering a wide variety of fresh fruit may be crucial in network (a) to initiate the purchasing sequence, while a focus on yogurt and almond milk variety can be essential in network (b) to exploit their role as purchase drivers, and network (c) suggests that jointly determining the assortments of yogurt with fresh fruit independently of the other products and likewise for almond milk with granola can be an effective strategy. Understanding the causal dynamics between these products is necessary to realize such benefits for optimizing the assortment to meet consumer buying habits.

Figure 1 **An illustration of alternative causal networks for four products from different categories**



While considerable advancement has been made in consumer choice modeling for purchases within a single category, it is only recently that new methods are emerging to characterize basket-shopping behavior of consumers across categories. In particular, the Multivariate MNL (MVMNL) and the Bundle Multivariate Logit Model (BundleMVL-K) models have been developed to accommodate a complementary effect across categories and a substitution effect within categories in modeling multiple-purchase decisions (Jasin et al. 2023, Tulabandhula et al. 2023). These models generalize the prior literature on Multinomial Logit Models (MNLs), Nested Multinomial Logit (NMNL), and Mixed MNL (MMNL) models and typically assume specific structures of basket-shopping behavior for tractability. However, there hasn't been sufficient empirical research thus far to assess the patterns of complementary and substitution behaviors that occur in multi-category baskets.

Discovering such causal relationships for basket-shopping consumers is a hard problem. For instance, for the above example with four products, there are 12 possible causal relationships and 2^{12} different causal graphs, which will likely yield different optimal product assortments. As the size of the product network increases, it gets astronomically difficult to conduct controlled experimental studies to identify which causal relationships are supported by data and evaluate their relative strength. In this paper, we propose a novel causal learning technique that is naturally suited to solve this problem. This approach constructs directed acyclic graphs (DAGs) that reveal the underlying dynamics among products using transaction data collected by retailers. For instance, in Figure 1, basket-shopping data would be used to relate the instances of purchasing fresh fruit, yogurt, granola, and almond milk with each other as well as with other products and iteratively eliminate possible conditional relationships until a DAG is obtained that is supported by data.

Before going further, it is important to recognize a caveat of our research. While we use the term *causal* based on the causal structure learning method that we employ, our study is not a causal study in the traditional economics sense of the term because we do not conduct controlled field experiments to measure each relationship. Instead, the causal structure learning approach is a data-driven approach that seeks to systematically eliminate relationships that do not have evidence in observational data. The network estimated from this approach will then have to be further tested by retailers and will also have to be examined for omitted variables, which can be done through selected field experiment studies. Our goal is to validate the applicability of this technique for analyzing retail transaction data and assess its potential for extracting important insights. Thus, we do not claim that the networks discovered in our estimation are truly causal.

Our research questions are as follows: (1) Is the partial correlation–based independence test used in PC algorithm, which has been utilized in computer science and economics for a variety of applications, well-suited for sparse, large-scale basket-shopping data, and how can we mitigate its shortcomings in estimating causal effects? (2) How do causal product networks perform in describing the relationships among product purchases compared to alternative methods for characterizing basket-shopping behavior, including both theory-driven and machine learning–based approaches? (3) What are the implications of the estimated network structure obtained from data on the optimal assortment strategy and to what extent does a multi-category assortment strategy based on the causal relationships among product purchases outperform a benchmark within-category strategy?

We conduct our analysis by collaborating with Numerator, a prominent market research company known for its first-party, consumer-sourced data. Numerator gathers purchase data for a large panel of consumers across various B&M and online channels; transaction sources include retailer loyalty data, mobile app purchases, email as well as paper receipts. Their extensive database comprises approximately 8.21 million consumer shopping baskets from more than 17,000 stores and around 3.90 million products in 36 categories, encompassing major retail chains such as Walmart, Costco, and Target, as well as small supermarkets and local grocery stores across the United States. We utilize the rich Numerator data by subdividing it into themes, including pasta, quick service restaurant

(QSR), bakery, prepared food, and unprepared food. We extract extensive data for these five themes, each comprising different product categories, and conduct our analysis on each theme to thoroughly study customers' basket-shopping behavior for different types of products. Thus, by presenting our empirical results separately for the five themes, we show evidence across different types of categories covering 250 products in 25 categories across B&M and online channels, with the number of baskets ranging from 5,719 to 13,056 for each theme. An exception is that the number of baskets for the QSR theme in the online channel is smaller than 2,500 due to its limited online availability.

With respect to the first research question, basket-shopping data is sparse because customers can only consider a small subset of the thousands of products available in a retail store during a single shopping trip. As a result, they typically purchase only a few items, leading to many zeros in the dataset that represent unpurchased products. The PC algorithm (Spirtes et al. 2000b), which is the most popular constraint-based causal structure learning algorithm, leverages probabilistic independence tests to infer these structures from observational data. We find that this method may yield erroneous results for our problem. For example, if fresh fruit and yogurt are almost always unpurchased in most baskets, then conditional on other products, the partial correlation test may spuriously detect a statistical association between them, incorrectly suggesting a causal relationship. To mitigate such biases, we propose a *Euclidean distance-based test* that captures variation across baskets without being distorted by the large number of zeros in sparse data. We characterize its computational complexity and utilize simulation analyses to demonstrate that our proposed approach consistently outperforms the partial correlation test used in the PC algorithm in learning causal effects from basket data under sparse conditions.

For the second question, our empirical analysis provides significant evidence confirming all three hypotheses that we formulate. First, we demonstrate that causal product networks provide a more accurate representation of product relationships in shopping baskets compared to: (1) Lasso- and Ridge-regularized complete networks, (2) singular value decomposition (SVD)– and eigenvalue decomposition (EVD)–based correlation networks, (3) collaborative filtering–based networks, and (4) association networks. Second, we compare different kinds of causal structures, such as category-level and product-level, and show that product-level causal effects most effectively describe the causal relationships in basket shopping, outperforming the alternative category-level restrictions. This result shows that consumers' basket-shopping behavior is driven by product-level causality, not category-level interactions as assumed in the prior literature. For example, we find that the categories of condiments and beverages are statistically independent, but purchasing fruit juice, a product in the beverages category, has a causal effect on the purchase of salad dressing, a product in the condiments category. Thus, product-level effects provide a more accurate description of basket-shopping consumer behavior. Moreover, we find that the number of edges in the product-level causal model is surprisingly only marginally larger than that in the alternative restriction-based specifications. These structural insights are useful for simplifying the empirical estimation problem because the potential number of parameters in this problem is extremely large. Third, we demonstrate that our technique can be utilized to evaluate different types of retail settings. Comparing online and brick & mortar (B&M) shopping baskets, we find that the online shopping channel exhibits fewer causal relationships among product purchases within shopping baskets compared to the B&M channel, which would imply different assortment strategies. Altogether, these results demonstrate the value of causal structure learning to characterize basket-shopping consumer behavior and provide insights into the relationships among product categories.

To answer the third research question, we conduct an optimization of product assortments using the causal product networks estimated from basket-shopping data across categories and examine its performance compared to the classic MNL model, which considers only within-category choice behavior. If causal modeling more accurately captures consumer behavior, it should naturally result in better assortment decisions than models that ignore causality across categories. Our goal in this analysis is to quantify the extent of the benefit and to obtain insights into the structure of the assortment. To address the scalability challenge of handling large product sets, we propose a two-stage graph learning model that explicitly incorporates category-level information, thereby improving computational efficiency. Considering 200 products from 25 different departments, we find that our causal model outperforms the MNL model by approximately 8.26%-20.62% in total sales across assortment sizes. Moreover, we observe distinct assortment strategies between causal product network models and MNL models, which reveals insights into the assortment implications of consumer behavior.

Thus, the contribution of our paper is that it presents a new representation of basket-shopping behavior as causal product networks, tests this representation rigorously against alternatives, and demonstrates its usefulness for generating new insights and improving retail decision-making. Our method shows that the large-scale basket-shopping data readily available to retailers can be extremely valuable as a cost-effective substitute for field experiments. We refine this method for our problem-setting to improve its performance with respect to data sparsity and scalability. We also note that our method can easily be generalized to incorporate not only purchase data but also supplementary product attributes, pricing and stockouts, or store characteristics to learn their causal implications from basket-shopping behavior and expand the scope of applications.

The remainder of this paper is organized as follows: §2 summarizes the related literature; §3 presents our causal product network model and describes our estimation procedure; §4 validates the performance of our proposed Euclidean distance-based approach in recovering causal effects compared to the partial correlation test used in the PC algorithm using synthetic data; §5 develops the hypotheses regarding basket-shopping purchase decisions; §6 presents the empirical results; §7 focuses on the assortment optimization problem; and §8 concludes the paper.

2. Literature Review

Our paper is related to the existing literature on (1) choice modeling and empirical analysis of basket-shopping behavior, (2) the use of graph models in retail analytics and (3) causal structure learning. In this section, we review the relevant literature on these topics and describe the contributions of our work. Choice modeling for basket-shopping customers seeks to discover and represent the complementary and substitution effects among product purchases with the goal of facilitating optimal decisions for product variety, pricing, and inventories. One common approach relies on the multinomial logit model (MNL), which is widely used in discrete choice modeling to capture substitution behavior within a category for either a single utility-maximizing choice (Börsch-Supan 1990, McFadden and Train 2000) or multiple purchase decisions (Aouad and Segev 2023, Bai et al. 2023, Aouad et al. 2024). The traditional MNL framework has been extended in the recent years to model multiple-purchase decisions across categories in basket shopping. For example, Cachon and Kök (2007) extend the nested-MNL model to describe the behavior of customers who choose one product from each of two categories and to demonstrate its performance in an assortment optimization problem. Jasin et al. (2023) develop the Multivariate Multinomial Logit (MVMNL) model and Tulabandhula et al. (2023) construct the Bundle Multinomial Logit (BundleMVL-K) model to account for both complementary and substitution effects across categories. In empirical research, Mani et al. (2022) use the MNL model to estimate substitution effects in a shopping basket. These papers have made considerable advances in developing the idea that purchasing one product can either increase or decrease the utility of purchasing another product leading to complementary or substitution effects between products. However, these models typically require further assumptions regarding the structure of the between-product effects for tractability, such as assuming that these effects are parameterized at the category-level rather than the product-level or that they are symmetric; these assumptions can significantly reduce the number of parameters but have not been empirically established.

Another stream of research uses data-driven methods to discover customer purchase patterns in shopping basket data. The foundational work by Agrawal et al. (1993) introduces the concept of mining association rules between product purchases, employing a breadth-first search strategy and candidate-generation to identify itemsets that frequently co-occur in shopping baskets. Manchanda et al. (1999) develop a multivariate probit model of multi-category purchase decisions to distinguish between factors such as coincidence and complementarity in purchases. In more recent papers, text mining and machine learning approaches have been applied to analyze co-purchase shopping behavior. For example, Jacobs et al. (2016) use probabilistic models, including Latent Dirichlet Allocation (LDA) and Mixtures of Dirichlet-Multinomials (MDM), to predict what a customer will purchase next by identifying latent customer topics. Gabel et al. (2019) customize a neural network language model to uncover latent product attributes by analyzing product co-occurrences in shopping baskets. Additionally, Xia et al. (2019) apply a Conditional Restricted Boltzmann Machine to explore cross-product patterns over time in large-scale consumer purchase data. In another recent paper, Ruiz et al. (2020) propose a random utility model of sequential discrete choice, where customers make purchase decisions one after another and the utility of subsequent purchases depends not only on the attributes of each product but also on all of the previous purchases. Collectively, these research articles demonstrate the value of processing large-scale purchase datasets to capture customer insights. They show the occurrence of co-purchase behavior, identify its various characteristics, and demonstrate the usefulness for predictive modeling. Our paper builds on this literature by introducing a model of causal purchase behavior. Our model is focused on applications such as assortment optimization, whereas the existing literature largely focuses on prediction problems such as sequential purchases, intertemporal variations, or the identification of co-purchase patterns.

Another set of recent research papers models choice and basket-purchases via product networks and directed acyclic graphs (DAGs). In e-commerce and information systems literature, this technique has been used for predictive modeling. Huang et al. (2007) represent the purchase behavior of customers in online marketplaces as bipartite consumer-product graphs and study the characteristics of these empirical graphs against behavioral predictions from random graphs. Oestreicher-Singer and Sundararajan (2012) construct co-purchase networks for books sold on Amazon.com and study the effect of network relationships on product demand. Dhar et al. (2014) also use the product network of books sold on Amazon.com and utilize it to predict future product demand as a function of the historical demand of a given product and its linkages with other products. In contrast to these papers, DAGs have been used differently in the operations management literature to represent preference ordering of products in non-parametric consumer choice models, such as by Honhon et al. (2010), Jagabathula and Vulcano (2018), and Jagabathula et al. (2022). These papers focus on within-category substitution models, whereas in our model, we allow DAGs to represent both complementary and substitution effects. We associate weights with the edges between products, where a positive weight represents complementarity and a negative weight represents substitution.

To discover the relationships in our model, we introduce a technique known as causal structure learning, which can infer causal relationships from observational data (Meek 2013, Glymour et al. 2019). Causal structure learning has been utilized in the fields of economics and policy-making in the recent years. Examples of this growing literature include uncovering latent macroeconomic models (Hall-Hoffarth 2022), the relationship between climate change and economic growth (Martinoli et al.

2023), inferring causal structures from observational data to validate instrumental variables for the well-known quarter-of-birth and proximity-to-college instruments to estimate the returns to education (Eberhardt et al. 2024), and combining short-term experimental data with long-term observational data to analyze the long-term effects of subsidies on healthy food products (Kaynar and Mitrofanov 2024). In addition to these studies in economics, causal structure learning has been widely applied in other areas, including bioinformatics, healthcare, and environmental science; we omit references to this literature for brevity.

Our study applies this approach to the field of retail operations, providing a comprehensive analysis of its effectiveness, testing it across different product categories and online versus brick & mortar channels, and addressing challenges related to the unique aspects of very large-scale retailing data, such as sparsity and heterogeneity. Our results demonstrate the value of this method compared to other state-of-the-art approaches and also provide a refinement to this method using Euclidean distance-based independence tests.

3. Causal Product Network Model and Estimation

This section presents our methodology. We begin by presenting a causal network model of basket-shopping behavior, then present our estimation method using causal structure learning and a robust Euclidean distance-based conditional independence test specifically designed to address the challenges posed by unobserved customer consideration sets. Finally, we describe how the resulting causal product networks can be parameterized using a linear structural equation model to estimate interpretable causal effects suitable for further optimization and decision-making.

3.1. Causal Product Network Representation of Basket-Shopping Behavior

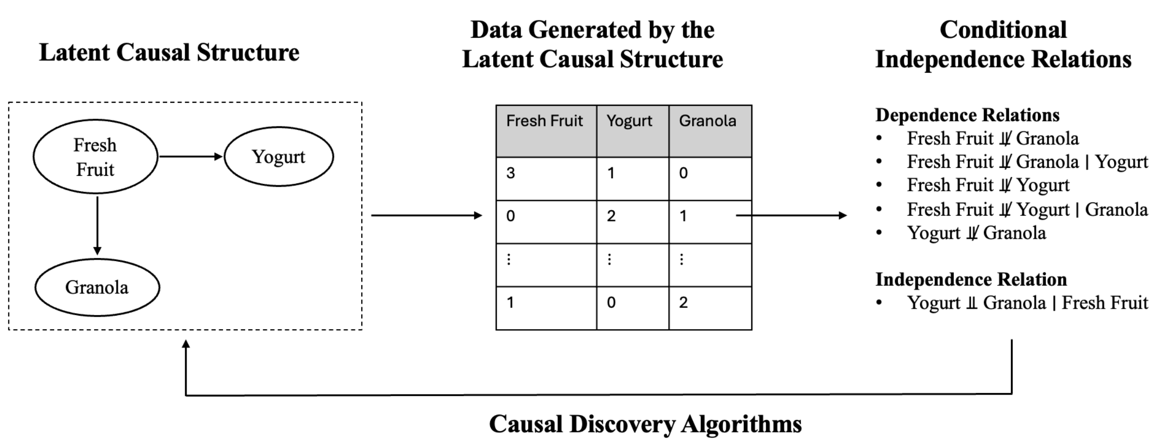
Suppose that we have basket-shopping data for a homogeneous set of customers. Let $J = \{1, 2, \ldots, m\}$ denote the set of product categories, where m is the total number of categories. For each category $j \in J$, let $I_j = \{1, 2, \ldots, n_j\}$ be the set of products, where n_j is the number of products in category j. The total number of products across all categories is $N = \sum_{j \in J} n_j$. We define $X_{ijt} \in \mathbb{Z}_{\geq 0}$ as a discrete random variable representing the quantity purchased of product $i \in I_j$ from category j in basket t, where $t = 1, \ldots, T$ and T is the number of observed baskets. For notational convenience, we also define $X_{ij} \in \mathbb{Z}_{\geq 0}^{1 \times T}$ as the vector representing the purchase quantities of product i in category j across all T baskets—that is, $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijT})$. The mapping of products into categories will be useful later on to formulate hypotheses on cross- and within-category causal relationships, but we do not use it as an input into the causal learning method. Indeed, an important aspect of our analysis is that we test the ability of the method to discover relationships without this knowledge.

Each basket is the outcome of purchase decisions made by a customer during the shopping trip. We assume that this purchase behavior of customers is represented by a directed, acyclic graph

$\mathcal{G} = (V, E)$, where each product purchase decision is a vertex, i.e., $V = \{X_{ij} \mid i \in I_j,\ j \in J\}$, and E denotes direct causal relationships from one product purchase to another. We say that an edge from product $X_{i'j'}$ to X_{ij} represents a causal relationship in the sense that the purchase decision of product $X_{i'j'}$ positively or negatively affects the purchase decision of product X_{ij}. Here, $X_{i'j'}$ is said to belong to the parent set of X_{ij}. If products $X_{i'j'}$ and X_{ij} are merely correlated due to other products and do not have a causal relationship, then there is no edge between products $X_{i'j'}$ and X_{ij}. Thus, this approach enables us to distinguish true causal relationships from correlations.

Using the causal graph, we model demand as follows. With each vertex X_{ij}, we associate $\alpha_{ij} \in \Re_{\geq 0}$ representing a baseline independent purchase rate for that product, and with each edge $(X_{i'j'}, X_{ij})$, we associate $\beta^{ij}_{i'j'} \in \Re$ representing the linear effect derived from each product $X_{i'j'}$ in the parent set $\text{pa}_{\mathcal{G}}(X_{ij})$ to product X_{ij}. The effect $\beta^{ij}_{i'j'}$ can take positive or negative values, representing situations where the demand for one product may cause an increase or a decrease in the demand for another product. The total mean demand rate for X_{ij} is thus given by $\alpha_{ij} + \sum_{X_{i'j'} \in \text{pa}_{\mathcal{G}}(X_{ij})} \beta^{ij}_{i'j'} X_{i'j'}$ and is constrained to be non-negative. Graph $\mathcal{G}$ can be partially connected. For example, if a product is not related to any other product, then it is represented as an isolated vertex with a null parent set. Similarly, if products within a category are substitutes and there is no relationship across categories, then the graph is separable into connected components corresponding to each product category. Thus, this model is flexible in representing all types of causal relationships between pairs of products and the resulting structure of demand. Our goal is to learn the graph $\mathcal{G}$ and the demand parameters from observed data.

Figure 2 **Example Causal Graph and the Implied Conditional Independence Relations**



We illustrate this conceptual framework with an example. Suppose a customer purchases fresh fruit, yogurt, and granola according to the relationships specified by the causal graph shown in Figure 2, i.e., the customer purchases fresh fruit, and with some independent likelihoods, buys yogurt and/or granola with the fresh fruit. Here, fresh fruit is the driver of the purchase and has direct causal effects on yogurt and granola purchases, while there is no direct causal edge between yogurt and granola. As a result, yogurt and granola may appear to be statistically dependent, but once we condition on fresh fruit purchases, the association between yogurt and granola disappears—they become conditionally independent given fresh fruit. If, instead, the purchase of yogurt directly influenced the purchase of granola, the graph would include a direct edge from yogurt to granola and the conditional independence would no longer hold: yogurt and granola would remain statistically dependent even after conditioning on fresh fruit. Thus, causal graphs are a versatile structure for representing purchase behavior and have important implications for retail decisions. For example, for the graph presented in Figure 2, a discount on fresh fruit would increase the purchase of yogurt and granola, whereas a discount on yogurt or granola would not increase the purchase of fresh fruit because fresh fruit is the driver product in the graph. As we show later, our hypotheses and empirical results reveal such insights into consumer purchasing behavior from the data.

Causal structure learning is a novel technique applied to observational data to identify the latent directed graph $\mathcal{G} = (V, E)$ of causal relationships, where V denotes the set of vertices representing the variables in the dataset and E denotes the set of directed edges to be discovered. The goal in this technique is to uncover these underlying causal relationships from purely observational data without the need for randomized experiments. This is possible because causal relationships impose specific constraints on patterns of statistical dependence and independence. Thus, we apply this technique in our setting by mapping product purchase decisions to the vertices and causal relationships between them to the edges of the graph.

Causal learning rests on two fundamental assumptions, Causal Markov Condition and Faithfulness. The causal Markov condition guarantees that every conditional-independence statistical inference in the data can be traced back to the graph, and thus, permits us to transition from the causal graph to the observed probabilistic independencies (Pearl 2009). Faithfulness rules out "coincidental" independencies (e.g., exact cancellation of opposing pathways), ensuring a one-to-one correspondence between the graph and the observed independence patterns (Spirtes et al. 2000a). Under these two assumptions, statistical independencies in the data faithfully mirror the underlying causal structure. Causal discovery algorithms, therefore, leverage these conditional-independence patterns in the data to iteratively evaluate candidate graph structures with respect to different conditioning sets and recover the causal graph $\mathcal{G}$ whose implied independencies best align with the data. In Appendix A, we provide brief theoretical background on causal structure learning, illustrative examples, and complete pseudocode for the PC algorithm (Spirtes et al. 2000b).

It is worth noting that some of the previous literature has studied the problem of sequential purchases and has modeled how to predict the next item a customer will buy based on the items already added to a shopping cart. These research papers are based on data in which the actual purchase sequence is observed, either directly from online retailing or through video tracking of customers. The identification of causal relationships, studied in our paper, is a different problem. First, causal relationships are *hidden or latent*, while the study of purchase sequences relies on the ability to observe the sequence. Moreover, causal relationships need not be the same as the purchase sequence, which can be guided by the store layout or the design of the e-commerce website. In particular, the relationships revealed by our analysis can be based on how customers construct their shopping lists or build a recipe rather than how they actually shop. The knowledge of these relationships would be valuable for retail decisions, such as assortment optimization, pricing, or recommendations. Thus, we focus on discovering these hidden customer purchase patterns from basket-shopping data.

3.2. Estimation Method

We estimate the causal model in two steps: first, we learn the graph G using causal discovery as discussed in this section and in Section 3.2.1, then we estimate the parameters α_{ij} and $\beta^{ij}_{i'j'}$ for all pairs of products as discussed in Section 3.2.2. One of the most well-known algorithms for causal discovery is the PC algorithm (Spirtes et al. 2000b). We adopt this method to systematically test conditional-independence relations in the data to recover the underlying causal structure.

Typically, the PC algorithm relies on a partial-correlation statistic to accurately detect conditional independence relations. However, we find that standard statistical tests for conditional independence are ill-suited for basket-shopping data due to their unique structural sparsity. In this setting, sparsity does not arise from missing data or random variation but from systematic behavioral constraints: customers consider only a limited subset of available products before making purchases in a shopping trip, and these consideration sets are unobserved. As a result, zeros in the data may reflect non-consideration rather than a true decision not to purchase. These structured zeros can induce misleading correlations or mask true dependencies, distorting the conditional independence relations that underlie causal graph recovery. Traditional methods, which treat all observed zeros as equally informative, fail to account for this latent selection mechanism.

To illustrate the challenge, suppose that we have 1,000 shopping baskets, but only 100 of them include both product X_{ij} and product $X_{i'j'}$ in the customer's consideration set. In these 100 baskets, assume that the purchases of X_{ij} and $X_{i'j'}$ are truly independent—customers may choose to purchase either, both, or neither. In contrast, the remaining 900 baskets do not include either product in the consideration set. As a result, both X_{ij} and $X_{i'j'}$ take the value zero in these baskets. If we naively apply a conditional independence test using the full dataset of 1,000 baskets, the large number of shared zeros from the 900 non-consideration baskets will create the appearance of strong positive correlation, even though no dependence exists in the subset of baskets where the products were actually considered. One might attempt to correct for this by removing baskets where both X_{ij} and $X_{i'j'}$ are zero. However, this introduces a new problem: in the reduced sample, whenever one product is not purchased, the other must be. This creates artificial dependence, distorting the true relationship between the variables. The problem becomes even more acute when testing for conditional independence given additional variables: conditioning splits the data into smaller subsets, and the number of baskets where all relevant products are jointly considered can become very small, increasing the likelihood of spurious conclusions.

To address the limitations of existing conditional independence tests in the context of sparse basket-shopping data, we propose a new method based on Euclidean distances.

3.2.1. Euclidean Distance-Based Solution for Data Sparsity. Our proposed test procedure is analogous to the partial-correlation test and proceeds in two steps. First, we define the degree of dissimilarity between two products as the squared Euclidean distance between their purchase vectors. When a conditioning set of products is present, this step becomes more challenging, as we must isolate variation not explained by the conditioning variables. To address this, we use QR decomposition to project each purchase vector onto the orthogonal complement of the subspace spanned by the conditioning set, and then compute the squared Euclidean distance between the resulting residual vectors. In both settings, the Euclidean distance measures the difference in product purchase behavior and remains stable in sparse data because matching zeros across vectors do not contribute to the distance. This makes it well-suited for basket-shopping data, where many products are frequently unpurchased in the same baskets. Algorithm 1 summarizes this procedure for computing the residuals and squared Euclidean distance, both with and without a conditioning set. In the second step, we characterize how this distance should behave if the residual vectors were independent—deriving the mean and variance of the squared distance. This allows us to formally test for distance-based independence by comparing the observed distance to its expected null distribution. This approach mitigates bias introduced by systematic sparsity while maintaining a consistent framework for assessing conditional and unconditional independence.

Next, we characterize how the Euclidean distance d between residuals should behave if R_{ij} and $R_{i'j'}$ are independent. To derive these properties, we assume that the original purchase vectors $X_{ij} \in V$ are normally distributed for all products in the dataset. This assumption is fundamental, as it ensures that the residuals R_{ij} and $R_{i'j'}$, obtained through the projection step in Algorithm 1, remain normally distributed. Consequently, the squared Euclidean distance d can be expressed as a sum of squared Gaussian differences, enabling the derivation of closed-form expressions for its mean and variance under the null hypothesis of independence. In practice, since purchase data have

Algorithm 1: Euclidean distance calculation for two products given a conditioning set

Input: $X_{ij}, X_{i'j'}$; $C \subseteq V \setminus \{X_{ij}, X_{i'j'}\}$.
Output: Residuals $R_{ij}, R_{i'j'}$ and distance d.
1. Let W be $W = [\, X : X \in C \,] \in \mathbb{R}^{T\times|C|}$.
2. **If** W is empty:
   $R_{ij} = X_{ij},\ R_{i'j'} = X_{i'j'}$.
3. **Else**:
   3.1. Compute QR decomposition: $W = Q\,R$.
   3.2. Projection: $P = Q\,Q^{\top}$.
   3.3. Residuals: $R_{ij} = (I-P)\,X_{ij},\ R_{i'j'} = (I-P)\,X_{i'j'}$.
4. Compute distance: $d = \sum_{t=1}^{T}(R_{ijt} - R_{i'j't})^2$.
5. **Return**: $R_{ij}, R_{i'j'}, d$.

count distributions, we apply the Anscombe transformation (Anscombe 1948) to the dataset to stabilize variances and approximate normality; these assumptions are required for both the Euclidean distance-based and partial correlation-based independence tests.[1]

Proposition 1 characterizes the mean and variance of the squared Euclidean distance when the residuals are independent. The proof is detailed in Appendix B.

Proposition 1 (Expectation and variance of squared Euclidean distance between residual purchase vectors)**.** *If X_{ij} and $X_{i'j'}$ follow a normal distribution, then the residuals R_{ij} and $R_{i'j'}$, obtained by applying Algorithm 1 with respect to the conditioning set C, also follow a normal distribution.*

Let $\mu_{R,ij}$ and $\mu_{R_{i'j'}}$ denote the means and $\sigma^2_{R,ij}$ and $\sigma^2_{R_{i'j'}}$ denote the variances of the residual purchase vectors R_{ij} and $R_{i'j'}$, respectively. Assuming that R_{ij} and $R_{i'j'}$ are independent, the expectation and variance of the squared Euclidean distance between these residuals, d, are given by $\mathbb{E}[d] = T\left(\mu_R^2 + \sigma_R^2\right)$ and $\mathrm{Var}(d) = T\left(4\mu_R^2\sigma_R^2 + 2\sigma_R^4\right)$, where T is the number of baskets and $\mu_R = \mu_{R_{ij}} - \mu_{R_{i'j'}}$ and $\sigma_R^2 = \sigma^2_{R_{ij}} + \sigma^2_{R_{i'j'}}$.

Having established the mean and variance of the squared Euclidean distance under the null hypothesis of independence, we now formalize the proposed independence test.

Definition 1 (Squared Euclidean Distance-Based Conditional Independence Test)**.** *Let X_{ij} and $X_{i'j'}$ be purchase vectors, and let $C \subseteq V \setminus \{X_{ij}, X_{i'j'}\}$ denote a conditioning set. Define the test as follows:*

(i) ***Residualization:*** *Compute residual vectors R_{ij} and $R_{i'j'}$ by applying Algorithm 1 to remove linear dependence on the conditioning set C.*

(ii) ***Test statistic:*** *Calculate the squared Euclidean distance $d = \sum_{t=1}^{T}(R_{ijt} - R_{i'j't})^2$ and standardize it as $Z = \frac{d-\mathbb{E}[d]}{\sqrt{\mathrm{Var}(d)}}$, where $\mathbb{E}[d]$ and $\mathrm{Var}(d)$ are as given in Proposition 1.*

(iii) ***Decision rule:*** *At significance level α, we reject the null hypothesis of distance-based conditional independence if $|Z| > z_{\alpha/2}$.*

[1] Testing for conditional independence using partial correlation assumes joint normality because a zero partial correlation implies conditional independence under multivariate normality.

The validity of the test rests on the fact that the mean and variance of the squared Euclidean distance are derived under the null hypothesis of independence. By standardizing the observed distance using these quantities, the resulting Z-statistic follows a known reference distribution under the null. This allows us to formally assess whether the observed distance is consistent with independence. A significant deviation indicates that the residual vectors, after accounting for the conditioning set, exhibit statistical dependence.

We apply the Euclidean distance test within the PC algorithm replacing the partial correlation test. Thus, its implementation is analogous to partial correlation. Since we deal with very large-scale data, we next show that the computational complexity of the Euclidean distance-based approach is at least the same as that of the partial-correlation-based approach. The main computational step involves projecting the purchase vectors onto the orthogonal complement of the conditioning set using QR decomposition followed by the calculation of pairwise distances. The following proposition provides the total computational complexity of evaluating the independence tests between all pairs of products for a given conditioning set $\boldsymbol{C}$ with the proof given in Appendix B:

Proposition 2 (Computational Complexity of Euclidean Distance-based Independence Test)**.** *Let* $N = |V|$ *and* $\boldsymbol{C} \subseteq V$ *be a conditioning set of size* $k = |C|$*. The computational complexity of conducting conditional independence tests for all pairs of products* $X_{ij}, X_{i'j'} \in V \setminus \boldsymbol{C}$ *with respect to* $\boldsymbol{C}$*, via computing the corresponding standardized test statistics* Z *as in Algorithm 1, is* $O(T(k^2 + k(N-k) + (N-k)^2)) = O(TN^2)$*.*

This complexity aligns closely with that of conditional independence tests based on partial correlations; that method requires inversion of covariance submatrices corresponding to conditioning sets (e.g., (Kalisch and Bühlman 2007)) to efficiently compute partial correlations, whereas we use QR decompositions to compute projections. The proof of Proposition 2 takes advantage of the fact that as the size of the conditioning set increases, the number of pairs of products for which a conditional independence test is to be performed decreases. Moreover, the QR decomposition of each conditioning set has to be computed only once to calculate the orthogonal projections of all other products. The number of conditioning sets is still exponential in the total number of products, and thus, the complete PC algorithm can be computationally intensive for both the Euclidean distance and the partial correlation-based metrics.

In practice, the total number of tests conducted by the PC algorithm depends on the density of the underlying causal graph and the number of neighbors of each node. Sparse graphs can be learned faster than dense graphs. Additionally, there are a few avenues to improve computational efficiency. First, when a given conditioning set is expanded by adding one product, orthogonal components of all other products can be incrementally updated rather than recalculated from scratch. Second, since the Euclidean distance is unaffected by baskets containing zero purchases of the products being considered, the dataset can be preprocessed before computing conditional independence statistics, thus taking advantage of sparsity. Finally, in our real data-based study in Section 7, we present a hierarchical implementation of the PC algorithm, which improves scalability.

Our proposed QR-residual Euclidean independence test adds to the recent literature on the development of distance-based methods to test the independence of random vectors. Székely et al. (2007) introduce distance correlation as a new test of independence that does not require normality and can be implemented in arbitrary dimensions. Székely and Rizzo (2014) extend this approach to conditional settings by proposing a partial distance correlation, Wang et al. (2015) present a different non-parametric extension that involves computing a conditional distance correlation using characteristic functions, and Zhang et al. (2011) proposes a kernel-based conditional independence test that allows for nonlinearity and non-Gaussian noise. Compared to these approaches, our approach is based on computing distances using the orthogonal projections of variables. This is similar to Székely and Rizzo (2014), but is computationally easier because of our simplifying assumption of multivariate normality.

In the next section, we compare the accuracy of the Euclidean distance-based test with the partial correlation-based conditional independence test.

3.2.2. Parameter Estimation using a Structural Equations Model. Causal structure learning can identify *which* product purchases are causally connected, but does not quantify *how strongly* these products influence one another. Thus, having learned the graph $\mathcal{G}$, our next step is to estimate the parameters α_{ij} and $\beta^{ij}_{i'j'}$, defined in Section 3.1, for all vertices and edges in the graph. For this purpose, we convert the estimated DAG into a linear structural equations model (SEM). Specifically, using the edges to each product X_{ij} from its parent set, we have:

$$X_{ij} = \alpha_{ij} + \sum_{X_{i'j'} \in \mathrm{pa}_{\mathcal{G}}(X_{ij})} \beta^{ij}_{i'j'} X_{i'j'} + \varepsilon_{ij}, \tag{1}$$

where α_{ij} and $\beta^{ij}_{i'j'}$ are as defined earlier and ε_{ij} represents the error term. We suppress the index for baskets for ease of notation. This model is estimated using the basket-shopping transaction data, where each basket represents a row in the estimation dataset. We separate our dataset into two parts, using one part to learn the graph $\mathcal{G}$ and the other to estimate the above regression model. We estimate these equations by ordinary least squares using basket-level transaction data.

When $\mathcal{G}$ is a fully oriented DAG, the system can be triangularized and estimated equation-by-equation without endogeneity bias. However, causal structure learning methods cannot always determine the direction of every edge, leaving some relationships between product purchases unoriented (see Appendix A for a detailed discussion). To incorporate undirected edges, we define pa_{ij} to include both directed and undirected neighbors, so that any undirected edge enters the equations for both neighboring variables. We note that a challenge in parameter estimation is that such unoriented or bidirectional edges can introduce simultaneity bias when estimating the SEM. The benchmarking models are more prone to this challenge because they often contain numerous bidirectional edges, making identification difficult. In practice, retailers can include additional variables such as product prices or promotional deals in the SEM to address this issue. In our analysis, we mitigate this issue through two approaches: (i) by using different samples to construct the causal product network and estimate the SEM, and (ii) applying dimensionality-reduction techniques such as regularization to simplify the benchmark models, thus minimizing the occurrence of bidirectional relationships.

4. Numerical Validation of the Euclidean Distance-Based Test

The partial correlation test used in the PC algorithm is known to be effective for causal inference and has been used in many applications (Spirtes et al. 2000b, Maathuis et al. 2009). In this section, we evaluate the performance of the proposed Euclidean distance–based independence test for recovering causal effects, particularly under varying levels of sparsity in basket-shopping data, by comparing it with the partial correlation test in the PC algorithm. Since the true causal graph is latent in real-world settings, we conduct a simulation analysis to thoroughly evaluate both methods with respect to a ground truth.

Our analysis proceeds in the following stages. First, we construct ground-truth CPNs $\mathcal{G}$ by generating random DAGs over 4-12 products grouped into 2-4 categories. The resulting graph defines the causal structure across products with directed edges indicating potential influence on purchasing behavior. We parameterize this graph by assigning each product a base purchase probability α_{ij} and each edge a causal effect coefficient $\beta^{ij}_{i'j'}$. Second, we simulate synthetic basket data based on the specified CPN. Third, we implement a consideration set mechanism to reflect the possibility that not all products are evaluated in every basket. Specifically, each product is independently excluded from the consideration set with probability p_1 in a randomly selected proportion p_2 of the sample baskets. If a product is excluded, its purchase count is set to zero, and it does not exert any causal influence on downstream products in the graph. Product purchase decisions are updated in topological order to preserve causal ordering. Fourth, we transform the basket data using the Anscombe transformation, then recover causal graphs using the PC algorithm with each test, resulting in two estimated graphs: $\bar{\mathcal{G}}_E$ from the Euclidean distance–based test and $\bar{\mathcal{G}}_{PC}$ from the partial correlation test. Finally, we estimate the parameters of each graph via SEM and evaluate the estimation accuracy using the root mean square error (RMSE) between the estimated and true parameter values α_{ij} and $\beta^{ij}_{i'j'}$. We present the complete details of this analysis, parameter values, and computational algorithm in Appendix C.

We create different sparsity conditions in two ways, first by modeling consideration sets as described above, and second, by varying the parameters of the CPN. Specifically, we consider three simulation scenarios that vary base purchase rates and causal effect magnitudes: (1) **Low-Causality Sparse Case**: a scenario with low base purchase rates and small magnitudes of causal effects; (2) **Non-Sparse Case**: a scenario with high base purchase rates and small magnitudes of causal effects; and (3) **High-Causality Sparse Case**: a scenario with low base purchase rates and large magnitudes of causal effects.

Table 1 **RMSE Performance of Euclidean Distance–Based and Partial Correlation Independence Tests**

	Scenario 1 (Low-Causality Sparse) α ~U(0, 0.01) β ~U(0, 1)			Scenario 2 (Non-Sparse) α ~U(0, 0.1) β ~U(0, 1)			Scenario 3 (High-Causality Sparse) α ~U(0, 0.01) β ~U(0, 2)		
p_1	Euclidean	Partial Correlation	Diff.	Euclidean	Partial Correlation	Diff.	Euclidean	Partial Correlation	Diff.
20%	0.30	0.33	-0.03*** (0.07)	0.23	0.22	0.01 (0.06)	0.61	0.92	-0.31*** (0.60)
40%	0.31	0.33	-0.02*** (0.07)	0.22	0.23	-0.01 (0.06)	0.54	0.78	-0.24*** (0.42)
60%	0.30	0.36	-0.06*** (0.10)	0.23	0.25	-0.02 (0.14)	0.58	0.93	-0.34*** (0.60)
80%	0.30	0.34	-0.04** (0.16)	0.22	0.25	-0.03* (0.13)	0.64	0.84	-0.20*** (0.58)

Note: (1) *p<0.1; **p<0.05; ***p<0.01; (2) α, β, and p_1 represent the base purchase rate, the magnitude of the causal effect, and the probability that a product is excluded from the consideration set, respectively; (3) The 'Euclidean' and 'Partial Correlation' columns report the RMSEs for the Euclidean distance–based and partial correlation–based independence tests; (4) Reported significance levels are based on a one-sided paired t-test under the null hypothesis $\text{RMSE}_\text{E} - \text{RMSE}_\text{C} \geq 0$; (5) Each result is based on 50 independent simulation replicates.

Table 1 presents the RMSE performance of the Euclidean distance–based and partial correlation–based methods for each scenario and each level of p_1. In Scenario 1, where both the base purchase rates and causal effect magnitudes are low, the Euclidean method consistently achieves lower RMSEs than the partial correlation method. This pattern holds across all values of p_1, which governs how often products are excluded from the customer's consideration set. The differences are statistically significant at the 0.05 level, indicating that the Euclidean method more accurately recovers causal effects even when many products are frequently not considered. In Scenario 2, where purchase rates are high but causal effects are small, both methods yield similar performance. The Euclidean method shows a statistically significant advantage only when product exclusion is most frequent ($p_1 = 80\%$), while at lower levels of exclusion ($p_1 = 20\%$, 40%, and 60%), the differences are not significant. This suggests that when most products are typically considered and causal effects are weak, the performance of the two methods converges. In Scenario 3, where purchase rates are low and causal effect magnitudes are high, the Euclidean method clearly outperforms partial correlation. RMSE differences are statistically significant at the 0.01 level across all levels of p_1, showing that the Euclidean method remains effective even when many products are regularly excluded from consideration. We conclude this analysis with two main findings. First, the proposed Euclidean distance–based approach consistently outperforms the partial correlation test in estimating causal effects from basket data in sparse settings, and performs comparably even when the data is less sparse. Second, this advantage holds across varying levels of product exclusion from the consideration set (p_1), indicating that the approach remains robust when products are not always included in a customer's consideration set. Based on these findings, we use the Euclidean distance–based test for conditional independence within the PC algorithm for all subsequent analyses. That is, whenever the algorithm tests whether two products are conditionally independent given a set of other products, it applies the Euclidean distance–based criterion.

5. Hypothesis Development

In this section, we develop our hypotheses to evaluate the ability of CPN to characterize basket-shopping consumer behavior, compare it with alternative theory- and machine learning–based network structures, and utilize it to provide insights.

5.1. Comparison of Alternative Network Structures

We compare our approach to several alternative methods for characterizing basket-shopping consumer behavior, drawing from both theoretical foundations and machine learning techniques. Specifically, we consider four different ways to represent interactions among products in a shopping basket:

(a) **Complete network:** Inspired by Tulabandhula et al. (2023), this model assumes that each product is interconnected with every other product in the basket.

(b) **Correlation-based network:** Product connections are determined by the Pearson correlation test, capturing the linear association between product purchase patterns.

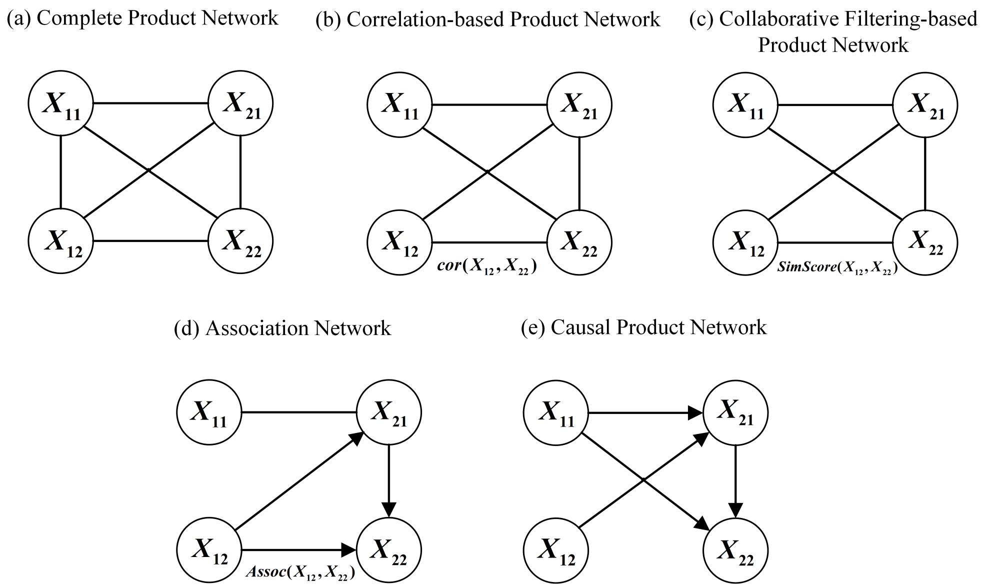
(c) **Collaborative filtering–based network:** Product connections are defined using cosine similarity, capturing behavioral similarity based on their co-occurrence in shopping baskets.

(d) **Association network:** This network is inferred using the procedure by Agrawal et al. (1993), which identifies directed edges based on association rules mined using the Apriori algorithm.

(e) **Causal network:** This network is inferred using the PC algorithm, which identifies edges based on conditional independence relations among products.

Figure 3 illustrates examples of these networks showing interactions among four products across two categories. We note that all the networks in Figure 3 can be alternative representations obtained from the same real-world shopping basket dataset. For instance, if we assume correlation fully captures relationships between product purchases and use only a correlation-based network, we might incorrectly infer a direct connection between purchases of x_{12} and x_{22} when, in reality, this relationship could be mediated by purchases of x_{21}. Similarly, in the collaborative filtering–based network, cosine similarity captures only pairwise relationships between product purchases. For the association

Figure 3 Different types of product interaction networks that can arise from the same purchase data by making different modeling assumptions



Note: Figure 3 illustrates how the same basket-level purchase data can give rise to different product interaction networks depending on the modeling assumption. The complete network assumes full connectivity; the correlation-based network encodes statistically significant correlations; the collaborative filtering–based network reflects similarity in co-occurrence; the association network reflects association rules that exceed both the minimum support and confidence thresholds, as defined by Agrawal et al. (1993); and the causal network, inferred using conditional independence tests, represents directional relationships learned through causal structure learning.

network, the association rules are based on co-occurrence patterns, and thus, may reflect spurious patterns. Unlike these network structures, the proposed causal structure can remove spurious connections between product purchases and provide directional relations by conducting rigorous conditional independence tests for each relationship. This results in an accurate and parsimonious representation of purchase relationships. Thus, as the first application of our methodology, we compare these network representations on our dataset by testing the following hypothesis.

Hypothesis 1 (Network Structures)**.** *A causal product network more accurately represents product purchase relationships in basket shopping compared to complete, correlation-based, collaborative filtering-based, and association networks.*

The structural equations models for the above network structures are described in Appendix D.1. For the estimation of the complete network model, we apply Lasso and Ridge regularization by adding ℓ_1- and ℓ_2-penalties, respectively, to the OLS loss function in order to induce sparsity, reduce estimation variance, and improve interpretability. We refer to the resulting models as the *Lasso-regularized* and *Ridge-regularized complete networks*. For the correlation-based model, we apply matrix decomposition techniques, specifically singular value decomposition (SVD) and eigenvalue decomposition (EVD), to obtain low-rank approximations of the correlation matrix. These denoised representations, known as the *SVD-* and *EVD-decomposed correlation networks*, help reveal underlying correlation structure and reduce noise in sparse basket data (Eckart and Young 1936, Alter et al. 2000, Wall et al. 2003). Thus, altogether, we compare the CPN approach to six state-of-the-art alternatives from the literature.

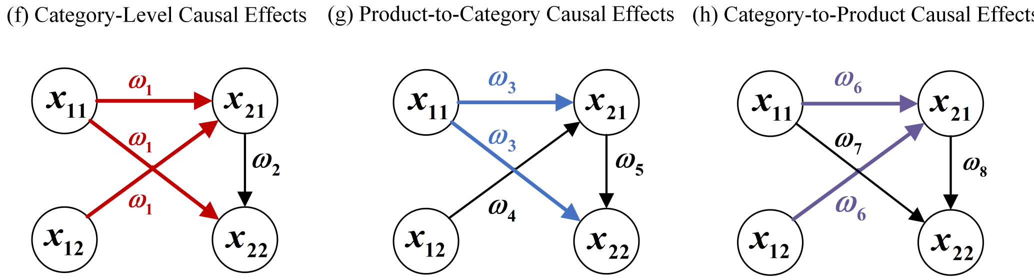
Our test of this hypothesis is important for modeling basket-shopping behavior because each of these network representations reflects a different set of assumptions about how product purchases relate to each other. The complete network assumes a fully connected structure but relies on regularization techniques to induce sparsity during estimation. The correlation-based, similarity-based, association-based, and causal networks impose structural constraints that limit potential product interactions based on statistical significance, similarity, association rules, or inferred causal relationships. Our statistical comparisons of these models presented in Section 6 show that CPN provides a more accurate representation of basket-shopping behavior than the alternatives.

5.2. Specifications for Causal Representations in Basket Shopping

Building on the hypothesis that a causal network more accurately represents basket-shopping behavior, we now investigate various specifications within causal structures. The prior literature shows that there can be different kinds of assumptions regarding consumers' basket-shopping behavior. At the one extreme, Song and Chintagunta (2006) and Jasin et al. (2023) assume that interactions between two products from distinct categories depend solely on the categories to which they belong. At the other extreme, our CPN model allows arbitrary interactions between any pairs of products. There can be many other possibilities in between. Figure 4 illustrates these specifications with reference to the causal network shown in Figure 3(e). The products x_{11},x_{12} and x_{21},x_{22} belong to categories 1 and 2, respectively. The parameters ω_p for $p = 1, 2, \ldots, 8$ correspond to different magnitudes of causal effects. Figures 4(f)-(h) show different conceptual representations of consumer behavior. Specifically, Figure 4(f) considers category-level specification based on Song and Chintagunta (2006) and Jasin et al. (2023), ensuring that the magnitude of causal effects from x_{11} to x_{21}, x_{11} to x_{22}, and x_{12} to x_{21} are identical. Figure 4(g) considers product-to-category specification, ensuring the magnitude of causal effects from x_{11} to x_{21} and x_{11} to x_{22} are identical. Finally, Figure 4(h) shows a category-to-product specification, where we restrict the magnitude of causal effects from x_{11} to x_{21} and x_{12} to x_{21} to be identical.

The study of these specifications is useful for modeling and estimating consumer shopping behavior because the identification of the correct model has implications for decision-making tasks such as assortment planning, pricing, and promotions. The four models, including Figure 3(e) and Figures 4(f)-(h), exhibit varying levels of generalization: considering causal effects between products represents the most general structure while considering causal effects between categories is the most restricted. These restrictions, if proven accurate, can have critical managerial implications. Additionally, they are also computationally relevant because the number of parameters to be estimated

Figure 4 **Alternative causal product network specifications to Figure 3(d) with different types of imposed restrictions**



decreases with the restrictions. On the one hand, the more general models allow parameters to be tailored to individual products, but on the other hand, having more parameters might be computationally burdensome and might not provide the best model fit. Thus, balancing this tradeoff, we hypothesize that models capturing product-level interactions provide a more accurate representation of shopping data even when penalizing for model complexity.

Hypothesis 2 (Specification Comparison)**.** *Across the four causal specifications, the product-level causal effects model illustrated in Figure 3(e) most effectively describes the causal relationships in basket shopping (in terms of model fit) compared to alternatives illustrated in Figures 4(f)-(h). In other words, the restrictions imposed by the alternative models do not hold.*

We provide the formal equations defining the SEMs for these specifications in Appendix D.2. The three specifications shown in Figure 4 are applied as restrictions to the product-level model presented in Equation (6). This enables us to test Hypothesis 2 using the AIC criterion and likelihood ratio test for nested models.

5.3. Sparsity of Causal Product Networks Across Channels

The power of our causal network learning approach for archival data is that it can be not only used to construct evidence regarding different model specifications but also applied separately to different retailing contexts to analyze differences in basket-shopping behavior of customers. We now turn our attention to investigating differences in these networks across brick-and-mortar and online channels. Indeed, the prior literature has examined assortment optimization problems across these two channels. For example, Lo and Topaloglu (2022) examines an assortment problem for a physical store aimed at maximizing the total expected revenue of an omnichannel retailer. Gopalakrishnan et al. (2023) compare the impacts of assortment width, defined as the number of unique categories in which a retailer offers products, and assortment depth, defined as the average number of products offered within each category, on order delivery timeliness. Sapra and Kumar (2023) examine the joint assortment optimization strategy for omnichannel retailers, taking into account three classes of customers who use only the brick-and-mortar channel, only the online channel, or both channels.

The models in the above literature are based on differences in both the cost economics and consumer behavior across the two channels. With regard to empirical evidence for the latter, Chintala et al. (2023) show that there are systematic differences between online and offline grocery shopping. Specifically, online purchases have significantly lower shopping basket variety than brick-and-mortar. However, they do not study causal relationships in consumer purchase behavior and highlight the need for considering causal relationships. Building on this, we hypothesize that the online channel exhibits fewer causal relationships among product purchases within shopping baskets compared to the brick-and-mortar channel (Hypothesis 3). We believe that this difference in sparsity can provide insights for omnichannel assortment optimization by revealing how consumer behavior varies across online and offline environments.

Hypothesis 3 (Sparsity)**.** *Compared to the brick-and-mortar channel, the online channel exhibits fewer causal relationships among product purchases within shopping baskets.*

Our method for testing Hypothesis 3 differs slightly from those used for the previous hypotheses. We first construct the causal graphs separately for online and brick-and-mortar baskets using 30 samples. Then, instead of estimating an SEM model, we compare the total number of edges in the resulting graphs using the paired t-test.

6. Empirical Evidence

In this section, we introduce our dataset, describe the sample construction process, and provide statistical summaries of the samples in Section 6.1. We then empirically test the proposed hypotheses and present the results of our analysis in Sections 6.2 and 6.3.

6.1. Data

We collaborate with Numerator, a prominent market research company known for its first-party, consumer-sourced data. Numerator gathers purchase data from both brick-and-mortar and online channels for a large panel of consumers through multiple sources: retailer loyalty data, mobile app purchases, email receipts, as well as uploaded paper receipts. Their extensive database comprises customer purchase records from more than 17,000 stores, encompassing major retail chains such as Walmart, Costco, and Target, as well as small supermarkets and local grocery stores across the United States. We use their dataset for the year 2021.

Since the number of products in a retail chain is vast, it is not managerially meaningful to construct a single CPN across all the products. Thus, for our analysis, we select five different *themes* for our analysis: each theme is a collection of multiple representative product categories and each category includes several products. For each theme, we consider the most frequently purchased products in each selected department to represent shopping baskets as these products are representative of shopping baskets. Further, to include information about other products within and outside these departments in the product networks, we create aggregated control variables for the total amount of purchases of other products within each selected department and of products outside the selected departments. These control variables help account for information about other product purchases in shopping baskets, thereby supporting the causal sufficiency assumption in our empirical study. This design of our study provides us with sufficient information to construct and analyze product networks for the five themes and demonstrate the usefulness of this approach. It enables us to conduct five tests of our hypotheses on different themes for a comprehensive evaluation.

Specifically, we focus on the following themes: pasta, quick service restaurant (QSR), bakery, prepared food, and unprepared food, labeled sequentially as Themes 1 through 5. We chose these themes based on two criteria: (1) items within each theme are available in both brick and mortar and online shopping channels, and (2) the themes are distinct enough to test the hypotheses across a range of scenarios. Theme 1 comprises five departments: pasta & noodle, meat, produce, dairy, and condiments. Theme 2 includes ten departments: QSR beverages, QSR breakfast, QSR sandwiches & wraps, QSR sauces & condiments, QSR Mexican, QSR snack & Sides, QSR desserts, QSR entrees, QSR Italian, and QSR salads. Theme 3 comprises five departments: bakery sweet goods, in-store bakery, packaged bakery, baking & cooking and dairy. Theme 4 includes five departments: frozen foods, canned, deli & prepared foods, beverages, and condiments. Theme 5 comprises five departments: produce, shelf stable meals, meat, herbs & spices, and condiments. Note that product category and department are used synonymously in our data set. The lists of the most frequently purchased product sets for the five themes across both brick-and-mortar and online channels, in the brick-and-mortar channel only, and in the online channel only are shown in Appendix E.

To test Hypotheses 1 and 2, we extract the basket data for each theme including purchases from both brick-and-mortar and online channels. For each theme, we identify the most frequently purchased products within each department in 2021 combined across both channels and choose a total of 50 products. In Themes 1, 3, 4, and 5, which consist of five departments each, we select the top 10 products from each department. In Theme 2, which includes ten departments, we select the top 5 products from each department. After establishing the product set for each theme using the entire dataset, we retrieve all the shopping basket data, including both online and brick-and-mortar channels, for 500 randomly chosen customers who had purchased products from at least three departments within the theme in a single basket. We use these samples as the graph learning datasets for causal structure learning, and collectively refer to them as Data I. Similarly, we extract another set of samples for 500 customers each as the parameter estimation samples for the SEM step and refer to them as Data II. This approach of separate samples for graph learning and parameter estimation avoids overfitting.

To test Hypothesis 3, since the most frequently purchased products differ across the two channels, we consider two product sets for each channel: the most frequently purchased products in brick-and-mortar stores and the most frequently purchased products online. We refer to these as the top brick-and-mortar products and top online products, respectively. For each product set, we retrieve baskets for each theme in the same manner as above, separating brick-and-mortar and online purchases, termed brick-and-mortar baskets and online baskets. Thus, we create four datasets for each theme as combinations of {brick-and-mortar, online} channels $\times$ {top brick-and-mortar, top online} product sets. We refer to this collection of datasets as Data III, which comprises twenty datasets ($= 4$ scenarios $\times$ 5 themes). By considering the most popular products from both brick-and-mortar stores and online platforms and analyzing purchases made through both these channels, we aim to show that the differences between brick-and-mortar purchases and online purchases are consistent across the datasets regardless of the type of products being considered.

Tables 2(a), 2(b), and 2(c) present the statistical summaries for the total number of products purchased per basket for Data I, Data II, and Data III, respectively. As shown in Table 2(a), the median number of purchased items is the same across the five themes, and the mean number is similar, ranging from 2.52 to 3.02. In contrast, the maximum quantity varies more substantially, ranging from 29 to 194 items. Similarly, in Table 2(b), the median number of purchased items remains the same across the five themes, and the mean number is similar, ranging from 2.55 to 2.94. The maximum quantity also varies considerably, ranging from 35 to 236 items. In Table 2(c), the median quantity of purchased items in B&M baskets remains at 2, whereas the median quantity in online baskets varies between 1 and 4. On average, online baskets also contain more items compared to B&M baskets. Additionally, the number of baskets for each theme varies between 5,000 and 15,000 across Data I, Data II, and Data III, with the exception of Theme 2 (Quick service restaurants theme) sample in the online channel. Note that the sample size for Theme 2 in the online channel is smaller, with only 1,776 baskets for the top brick-and-mortar channel products and 2,437 baskets for the top online channel products. We believe this is due to the nature of QSRs, making it attractive for customers to make purchases in person rather than waiting for online orders and delivery.

6.2. Tests of Hypotheses 1 and 2: Evidence from Causal Product Networks

To test Hypothesis 1, we generate the causal product network, the regularized complete networks, the matrix decomposition-based correlation networks, and the collaborative filtering-based network for each theme. Using Data I (graph learning sample), we first construct causal product networks for each theme using the PC algorithm as detailed in §3, and using a significance level of 0.05 for our proposed Euclidean distance-based approach to test for conditional independence. To ensure order independence and maximal informativeness, we use the complete PC algorithm, which applies

Table 2 Summary statistics of shopping baskets in Data I, Data II and Data III

(a) Data I (Graph Learning Sample), used for discovering causal DAGs in Hypotheses 1 and 2

	Total Quantity of Items in Each Basket							
	Min	25%	50%	75%	Max	Mean	SD	N
Theme 1 (Pasta)	1	1	2	4	29	3.02	2.96	14,689
Theme 2 (QSR)	1	1	2	3	105	2.74	2.84	8,858
Theme 3 (Bakery)	1	1	2	3	32	2.52	2.23	10,481
Theme 4 (Prepared Food)	1	1	2	4	61	2.78	2.70	8,345
Theme 5 (Unprepared Food)	1	1	2	3	194	2.62	4.05	14,017

(b) Data II (Parameter Estimation Sample), used for testing Hypotheses 1 and 2

	Total Quantity of Items in Each Basket							
	Min	25%	50%	75%	Max	Mean	SD	N
Theme 1 (Pasta)	1	1	2	4	97	2.94	2.82	14,595
Theme 2 (QSR)	1	1	2	3	35	2.69	2.62	9,408
Theme 3 (Bakery)	1	1	2	3	35	2.55	2.27	10,204
Theme 4 (Prepared Food)	1	1	2	3	99	2.63	2.74	12,530
Theme 5 (Unprepared Food)	1	1	2	3	236	2.67	3.80	12,892

(c) Data III, used for testing Hypothesis 3

	Top Brick-and-Mortar Products															
	Brick-and-Mortar Channel								Online Channel							
	Min	25%	50%	75%	Max	Mean	SD	N	Min	25%	50%	75%	Max	Mean	SD	N
Theme 1	1	1	2	4	29	2.92	2.61	14,472	1	2	4	6	27	4.80	3.67	6,664
Theme 2	1	1	2	3	47	2.75	2.57	9,247	1	1	1	2	48	1.91	2.90	1,776
Theme 3	1	1	2	3	41	2.53	2.20	10,114	1	2	3	4	25	3.39	2.39	6,125
Theme 4	1	1	2	3	112	2.64	2.86	11,744	1	1	3	4	60	3.50	3.28	5,719
Theme 5	1	1	2	4	170	3.16	3.77	12,901	1	2	4	6	46	4.81	4.07	5,914
	Top Online Products															
	Brick-and-Mortar Channel								Online Channel							
	Min	25%	50%	75%	Max	Mean	SD	N	Min	25%	50%	75%	Max	Mean	SD	N
Theme 1	1	1	2	4	258	2.89	4.10	13,023	1	2	5	8	62	5.78	4.54	6,613
Theme 2	1	1	2	3	44	2.52	2.32	8,159	1	1	1	2	25	1.56	1.37	2,437
Theme 3	1	1	2	3	72	2.72	2.51	9,219	1	2	3	5	24	3.69	2.66	6,502
Theme 4	1	1	2	4	39	2.82	2.89	10,361	1	2	3	5	53	4.02	3.64	6,240
Theme 5	1	1	2	4	165	2.93	3.09	13,056	1	2	4	7	35	5.39	4.42	6,340

Note: (1) Themes 1-5 correspond to Pasta, Quick Service Restaurant (QSR), Bakery, Prepared Food Items, and Unprepared Food Items, respectively, where each theme consists of several product categories; (2) N refers to the total number of baskets; (3) SD refers to the standard deviation of product quantity in shopping baskets; (4) Each theme includes 50 focal products and six control products and the above statistics are for the focal products; (5) Quantity of items refers to the total number of products in a basket for the respective theme.

all applicable rules to orient the edges of the causal skeleton. In cases where different rules suggest conflicting edge orientations, we retain both directions, resulting in undirected edges. This approach allows for flexibility in representing both directed and undirected relationships among product purchases, making the model more realistic and practically useful. The implementation of the PC algorithm is conducted using the *pcalg* package (Markus Kalisch et al. 2012).

Table 3 Model fit results for the Five network structures in Hypotheses 1

AIC Scores							
Network Structure	(a1)	(a2)	(b1)	(b2)	(c)	(d)	(e)
Theme 1 (Pasta)	567,990	571,658	559,445	559,393	578,273	563,451	**557,721**
Theme 2 (QSR)	448,529	447,874	437,961	438,198	514,212	441,388	**436,085**
Theme 3 (Bakery)	324,774	324,075	323,959	308,765	380,065	314,650	**304,645**
Theme 4 (Prepared Food)	652,517	654,361	647,528	646,568	718,641	647,204	**638,588**
Theme 5 (Unprepared Food)	928,613	932,847	988,911	928,083	1,063,438	857,323	**822,830**
Number of Parameters							
Network Structure	(a1)	(a2)	(b1)	(b2)	(c)	(d)	(e)
Theme 1 (Pasta)	762	2,450	498	474	500	157	213
Theme 2 (QSR)	1,424	2,450	714	600	500	109	254
Theme 3 (Bakery)	1,739	2,450	396	420	500	90	203
Theme 4 (Prepared Food)	1,159	2,450	424	364	500	155	176
Theme 5 (Unprepared Food)	246	2,450	262	266	500	92	210

Note: (1) Models a1 and a2 represent the Lasso-regularized complete network and the Ridge-regularized complete network, respectively; models b1 and b2 represent the SVD-decomposed correlation network and the EVD-decomposed correlation network; model c represents the collaborative filtering-based network; model d represents the association network; and model e represents the causal product network; (2) The size of product sets for all the themes is 50; (3) AIC scores are computed by first estimating the networks on Data I, then computing fit statistics for the SEM models on Data II; (4) The likelihood ratio test (LRT) statistics are reported in Appendix F, and all results are statistically significant at the 1% significance level.

We then generate Lasso-regularized and Ridge-regularized complete product networks for each theme in Data I by regressing each variable on all other variables using Lasso and Ridge regression, which automatically selects important variables through regularization. The regularization parameters (i.e. the strength of the penalty) are selected by cross-validation. In the SVD-decomposed and EVD-decomposed correlation networks, we first construct correlation matrices where edges between nodes are determined by pairwise Pearson correlation tests at the 0.05 significance level. We then apply Singular Value Decomposition (SVD) or Eigenvalue Decomposition (EVD) to the correlation matrix and retain the top components that explain 80% of the variance, yielding a low-rank approximation of the original structure. The simplified matrix is then thresholded to construct a sparse correlation network that reduces noise. For the collaborative filtering–based network, we construct edges based on item-to-item similarity, computed using cosine similarity between variables representing product purchases. For each product, we follow the top-K rule by connecting it to the 10 most similar products based on cosine similarity scores, and use these connections to define the network structure (Sarwar et al. 2001). For the association network, we follow the procedure of Agrawal et al. (1993) to construct directed edges based on association rules mined using the Apriori algorithm that exceed the minimum support and confidence thresholds. Since the basket shopping data are sparse, we set the minimum support to 0.1% and the minimum confidence to 15%.[2] Having

[2] In our study, we find that varying the minimum support between 0.01% and 0.1%, and the minimum confidence between 5% and 15%, yields similar model fit and does not affect the comparative results.

obtained all five types of networks, we formulate and estimate the corresponding SEMs using the procedure described in Appendix D.

Table 3 presents the model fit results using Data II (Parameter Estimation Sample) for all the methods. The top half of the table shows comparative analysis using Akaike Information Criterion (AIC) scores, which balances the number of parameters and model fit, and the bottom half provides insights into the number of statistically significant parameters in each model. As shown by Burnham and Anderson (2004), an AIC difference of more than 10 provides strong evidence in favor of the model with the lower AIC. We further supplement the AIC criterion with likelihood ratio tests presented in Appendix F. Note that each estimate is based on an SEM consisting of 50 equations for 50 products. The number of independent variables varies across the specifications and is different in Theme 2 since it consists of 10 categories.

Comparing the results across all the themes in Table 3, we find that the causal product network consistently exhibits the lowest AIC scores. The AIC differences are all much greater than 10. Furthermore, likelihood ratio test results reported in Appendix F confirm that the causal product network outperforms the benchmark models at the 1% significance level. Therefore, we conclude that the causal network provides a more accurate representation of product relationships in shopping baskets than all the other methods, thereby confirming Hypothesis 1. The matrix decomposition approaches provide the second best performance, and collaborative filtering consistently has the worst performance. Additionally, turning to the bottom half of the table, we observe that the CPN requires significantly fewer parameters compared to all other network structures, except for the association network. For example, in the pasta theme, the Lasso- and Ridge-regularized complete networks require 762 and 2,450 parameters, respectively; the SVD- and EVD-decomposed correlation networks require 498 and 474 parameters; the collaborative filtering-based network requires 1,970 parameters; whereas the causal product network requires only 213 parameters.[3] On the other hand, association network has only 157 edges. This sparsity arises precisely because association networks cannot detect substitution effects—it can account *only* for complementarity effects. In contrast, the causal product network effectively captures both substitution and complementarity while maintaining a parsimonious structure, thus reducing estimation complexity.

The causal product network empirically outperforms alternative network structures because it captures sparse, directional, and interpretable relationships among product purchases. Complete and correlation-based networks often result in dense graphs that obscure direct relationships and lead

[3] The Ridge-regularized complete network retains all edges between variables because Ridge regression applies an L2 penalty, which shrinks coefficients toward zero but does not set them exactly to zero. As a result, all variables remain in the model with small but non-zero weights, leading to a fully connected network. In contrast, the Lasso-regularized network uses an L1 penalty, which promotes sparsity by shrinking some coefficients exactly to zero.

to overfitting. In contrast, the causal product network uses Euclidean distance-based conditional independence tests to isolate direct causal influences, enhancing both interpretability and generalization. Collaborative filtering focuses on co-occurrence patterns and is prone to popularity bias, offering limited insight into the underlying structure of product relationships. Association rule-based networks primarily capture complementary relationships and tend to favor high-demand products, often neglecting substitution effects and underrepresenting interactions involving low-demand items. As a result, they provide a biased and incomplete view of product interactions. The causal product network addresses these limitations by modeling both complementary and substitutive effects within a causal framework. For these reasons, our empirical evidence supports the hypothesis that the causal product network provides a more suitable representation of product relationships.

Table 4 **Model fit results for the Four network structures in Hypotheses 2**

Network Structure	(e)	(f)	(g)	(h)
Theme 1 (Pasta)	**557,721**	571,693	569,227	561,288
Theme 2 (QSR)	**436,085**	452,019	447,819	440,309
Theme 3 (Bakery)	**304,645**	320,222	316,940	309,806
Theme 4 (Prepared Food)	**638,588**	647,984	646,745	641,099
Theme 5 (Unprepared Food)	**822,830**	922,520	915,581	881,728
Number of Parameters				
Network Structure	(e)	(f)	(g)	(h)
Theme 1 (Pasta)	213	33	94	121
Theme 2 (QSR)	254	72	140	156
Theme 3 (Bakery)	203	34	107	122
Theme 4 (Prepared Food)	176	35	93	114
Theme 5 (Unprepared Food)	210	36	90	118

Note: (1) Model e represents the causal product network; models f, g, and h represent the category-level causal effects model, the product-to-category causal effects model, and the category-to-product causal effects model, respectively; (2) The size of product sets for all the themes is 50; (3) Results are computed based on test data (Data II); (4) The likelihood ratio test (LRT) statistics are reported in Appendix F, and all results are statistically significant at the 1% significance level.

To test Hypothesis 2, we begin our analysis with the final causal graph obtained above. To this graph, we impose the category-to-category, product-to-category, and category-to-product specifications in the corresponding simultaneous equations models as discussed in §5.2. Table 4 presents the model fit results. Comparing the AIC score of (e) with the three alternative causal specifications (f), (g), and (h), we find that the product-level causal model consistently shows the lowest scores. Again, the differences in AIC scores exceed 10, and likelihood ratio tests (detailed in Appendix F) confirm that the causal product network outperforms other model specifications at the 1% significance level. This shows that the product-level specification most effectively captures the causal relationships in basket shopping and the restrictions imposed by the other models do not hold, substantiating

Hypothesis 2. This result provides us with an important finding that consumers' basket-shopping behavior is driven by product-level causality, not category-level interactions, which is relevant for the design of choice models for basket-shopping behavior. Moreover, note that there is a very highly significant difference in the AIC scores of model (e) versus all the other models, demonstrating the quality of the fit.

As shown in the bottom half of the table, the number of edges in the product-level causal model is slightly larger than in the alternative restriction-based specifications. For example, in the pasta theme, the causal product network includes 213 parameters, compared to 33 for the category-level specification, 94 for the product-to-category specification, and 121 for the category-to-product specification. We make two inferences from these results: (i) the number of causal relationships in basket-shopping behavior is not too large, making it computationally tractable and managerially feasible to analyze and manage these relationships, and (ii) the correct specification of these relationships is critical for model fit, which should materially affect decision-making tasks based on choice modeling.

6.3. Test of Hypothesis 3: Brick-and-Mortar Vs Online Basket-Shopping Behavior

To test Hypothesis 3, we create causal product networks separately for each of the 20 datasets in Dataset III representing both brick-and-mortar and online channels. For each theme and product set, we extract 30 subsamples from the corresponding dataset, each consisting of orders from 100 randomly selected customers. We then apply the PC algorithm with our Euclidean distance-based metric to discover the causal graph for each sample, and compare the sparsity of the brick-and-mortar and online causal product networks using a paired t-test. Each pair in the test represents the total number of edges in the brick-and-mortar causal network and the corresponding number in the online causal network in the same subsample (undirected edges and directed edges are counted only once to ensure consistency in our analysis).

The results of the paired t-test are presented in Table 5. We report the average number of edges across the 30 discovered causal graphs for both the brick-and-mortar and online channels, along with the differences between these two statistics. Our results show that the online channel has a smaller number of edges than the brick-and-mortar channel in 9/10 cases; eight of these are statistically significant at a 0.01 significance level and one at 0.1. For Theme 4, there is a significant difference when top online products are considered, but not when the top brick-and-mortar products are considered. Overall, these results support our hypothesis that online causal product networks are sparser than their brick-and-mortar counterparts. This implies that purchases in a basket in the online channel are more likely to be independent of each other, with limited complementary and substitution effects, than those in the brick-and-mortar channel. Thus, assortments and inventories

Table 5 Average number of directed edges in DAGs across 30 simulations: Test for Hypothesis 3

	Top B&M Products				Top Online Products			
	B&M	Online	Difference	Num. of Pairs	B&M	Online	Difference	Num. of Pairs
Theme 1 (Pasta)	63.10	55.83	7.27*** (11.10)	30	97.17	81.77	15.40*** (17.05)	30
Theme 2 (QSR)	45.70	36.07	9.63*** (9.55)	30	56.60	35.97	20.63*** (16.94)	30
Theme 3 (Bakery)	68.13	61.83	6.30*** (5.42)	30	64.60	48.57	16.03*** (9.64)	30
Theme 4 (Prepared Food)	12.37	13.93	-1.57 (5.67)	30	13.03	10.50	2.53*** (4.37)	30
Theme 5 (Unprepared Food)	17.23	13.97	3.27*** (4.83)	30	15.87	14.50	1.37* (5.11)	30

Note: (1) *p<0.1; **p<0.05; ***p<0.01; (2) The values for B&M, Online, and Difference correspond to the average number of directed edges across 30 causal graphs discovered from randomly selected subsets for B&M, Online, and the B&M–Online comparison, respectively.

for online channels can be planned independently for products more easily than for the brick-and-mortar channel. Moreover, this also suggests that there is a greater need for recommendation systems and cross-promotions to increase basket size in online channel than in the brick-and-mortar channel.

We next compare the structures of directed graphs obtained for the brick-and-mortar and online channels to obtain further insights into consumer behavior. To present our findings, we focus specifically on the pasta theme (theme 1) within the top brick-and-mortar (B&M) product set.

Our analysis of causal edges reveals clear behavioral differences between these two channels. Considering a total of 2,750 edges for 50 focal products and 6 control variables, we find that 57.4% of edges do not exist in either channel, 12.3% of the edges exist in both channels, and the remaining 30.3% of edges appear exclusively in either the B&M or the online channel. The large difference between 12.3% and 30.3% shows that the two channels have different consumer behavior patterns. Exploring further, we find that edges involving produce items appear substantially more frequently in B&M channel, showing an 86% higher frequency compared to online channel. Specifically, produce-to-condiment edges are notably stronger in B&M channel (132% higher), indicating increased causal connections from fresh produce toward condiments, consistent with meal-planning or recipe-driven shopping. Additionally, edges involving impulse-oriented products, such as lunch packs, are significantly more frequent in B&M channel (108% higher), further highlighting spontaneous purchasing behaviors unique to physical stores. Conversely, online causal edges predominantly involve refrigerated staples such as dairy and processed meats, accounting for approximately 43% and 27% of the online edges, respectively. This pattern reflects routine, convenience-focused replenishment purchasing in the online channel rather than pantry-oriented purchasing. In fact, edges involving shelf-stable pantry categories remain relatively limited online, despite their suitability for convenient online shopping. Collectively, our findings underscore a clear distinction: physical grocery stores are characterized by spontaneous fresh-meal assembly, pantry categories, and impulse purchasing, whereas online grocery shopping predominantly involves regular, convenience-driven replenishment, particularly evident in dairy and processed-meat categories.

Further, the magnitude of causal effects also differs between the B&M and online channels. For each channel, we aggregate all the graphs obtained from the 30 subsamples, considering a causal edge to exist if it appears at least once in a subsample, and run SEM to estimate the sizes of causal effects based on this aggregated CPN. We find that the average causal effect magnitude in the B&M channel is 0.019, which is lower than 0.027 observed in the online channel. Additionally, the standard deviation of causal effect magnitudes in B&M is 0.028, also lower than 0.048 in the online channel. This indicates that causal effects in the B&M channel tend to be smaller and less variable compared to those in the online channel even though the number of directed edges tends to be larger as shown in Table 5. Considering all 2,750 edges, 338 (12.3%) exist in both channels with non-zero estimates. Among these, 175 edges (6.4%) have positive causal effects in both B&M and online channels, while 53 edges (1.9%) have negative effects in both channels. The remaining 110 edges (4%) exhibit causal effects with opposite signs across channels.

In addition, in the pasta theme, we observe more positive within-category causal effects in the online channel compared to the B&M channel. Specifically, there are 9 positive and 16 negative causal effects in B&M, versus 23 positive and 10 negative effects in the online channel. This suggests that within-category substitution effects for pasta and noodles are stronger when customers shop in physical stores, whereas complementary effects are more prominent when customers shop online. Outside the Pasta & Noodle department, a similar pattern emerges. In the Meat, Produce, Dairy, and Condiments departments, the B&M channel exhibits 22, 46, 26, and 27 positive effects respectively, alongside 13, 24, 24, and 21 negative effects. In contrast, the online channel shows 25, 52, 27, and 22 positive effects and 8, 13, 22, and 13 negative effects in the same departments. These findings further support the notion that substitution effects tend to be stronger in physical stores, while complementary effects dominate in online shopping.

Thus, we conclude that CPNs demonstrate considerable differences in consumer behavior across the brick-and-mortar and online channels. Using these insights, retailers can tailor their recommendation, marketing, and assortment strategies to better align with each channel.

7. Scalable Learning and Assortment Optimization

Thus far, we have demonstrated that the CPN model captures purchase relationships in shopping basket more effectively than alternative econometric or machine learning–based methods. In this section, we quantify the value of this modeling approach by comparing the assortment computed using the CPN consumer purchase behavior model with a within-category benchmark assortment.

Our goals are to assess the impact of accurately modeling complementary and substitution downstream causal effects across categories versus a within-category model and to obtain insights into the structure of the assortment.

We begin by proposing a two-stage graph learning estimation approach for our CPN model to improve its computational efficiency. This approach explicitly incorporates category-level information to reduce the number of causal relationships that have to be tested. Then, we determine the optimal product assortment by accounting for causal interactions among products through the learned product-level causal graph, which includes both across- and within-category complementary and substitution effects. Finally, we measure the impact of this approach by comparing it to a purely within-category choice model for the basket purchase data of all five themes considered.

7.1. Two-Stage Scalable Graph Learning for Basket Purchase Data

Since there is a large number of product categories in a retail store, we simplify the computational burden of causal learning by exploiting the category classification of products. Specifically, we first perform a category-level analysis, aggregating product-level data into category totals and estimating dependencies between categories. Subsequently, we conduct product-level graph learning, leveraging insights from the category-level step to focus the causal discovery process on product pairs within dependent categories. This two-stage procedure substantially reduces the complexity of causal inference, enabling efficient identification of relevant product-level relationships while maintaining the detailed granularity needed for assortment optimization. We begin with the assumption of aggregate-level independence between categories:

Assumption 1 (No Exact Cancellation of Product-Level Covariances)**.** *For two distinct categories j and j', define the aggregates $W_j = \sum_{i \in \mathcal{I}_j} X_{ij}$ and $W_{j'} = \sum_{i' \in \mathcal{I}_{j'}} X_{i'j'}$. If the aggregates are uncorrelated (i.e.,* $\operatorname{corr}(W_j, W_{j'}) = 0$*), then every cross-category, product-level correlation is also zero,* $\operatorname{corr}(X_{ij}, X_{i'j'}) = 0$ *for all $i \in \mathcal{I}_j$ and $i' \in \mathcal{I}_{j'}$.*

Note that Assumption 1 involves only the correlation structure among observed variables: it explicitly rules out scenarios in which positive and negative product-level correlations exactly offset each other, creating a misleading appearance of independence at the aggregate level. We analyze the estimation results in Section 6.2 for the causal product networks and find that 91.50% of causal effects between departments share the same sign on average across all five themes. This consistency indicates that offset effects are rare, supporting our assumption. Proposition 3 extends this assumption into a causal implication. Specifically, under the standard Faithfulness assumption for linear–Gaussian SEMs, nonzero correlations among variables indicate the presence of at least one causal or confounding path between them. Given Assumption 1, zero correlation between category-level aggregates implies zero correlation at the product level, which, by Faithfulness, implies no causal connection between any products in these two categories. Thus, Proposition 3 formally establishes that aggregate-level independence implies causal independence, allowing us to simplify causal discovery by focusing only on products in categories with non-zero aggregate correlations.

Proposition 3. *Suppose the product-level purchase quantities X_{ij} follow a linear Gaussian SEM satisfying Assumption 3 (Faithfulness), and define category-level aggregates $W_j = \sum_{i \in \mathcal{I}_j} X_{ij}$. Under Assumption 1 (No Exact Cancellation of Product-Level Covariances), if two aggregates are uncorrelated, $\mathrm{corr}(W_j, W_{j'}) = 0$, then there is no causal edge between any product in category j and any product in category j'.*

Intuitively, Proposition 3 shows that causal effects between products across categories would generate detectable nonzero correlations when aggregated, unless exact cancellations occurred, as Assumption 1 explicitly prohibits this possibility. Therefore, aggregate independence can guide efficient causal discovery by reducing the number of candidate product-level causal relationships. Building on these theoretical insights, we propose the following two-stage algorithm:

Algorithm 2: Two-Stage Graph Learning for Basket Purchase Data

Input: Basket data X_{ij}^t for $t = 1, \dots, T$, $i \in \mathcal{I}_j$, $j \in J$.
Output: Product-level graph $\mathcal{G}$.
Step 1: Category-Level Analysis
1.1. For each category j, compute category totals $W_j = \sum_{i \in \mathcal{I}_j} X_{ij}$.
1.2. Define the set of uncorrelated category pairs $\mathcal{U} = \{(j, j') \mid \mathrm{Cov}(W_j, W_{j'}) = 0\}$.
Step 2: Product-Level Graph Learning
2.1. Initialize a fully connected graph over all product pairs.
2.2. For each category pair $(j, j') \in \mathcal{U}$, remove all edges between products (i, j) and (i', j') for all $i \in \mathcal{I}_j$, $i' \in \mathcal{I}_{j'}$.
2.3. Apply a causal-discovery algorithm to this constrained graph to obtain the refined product-level graph $\mathcal{G}$.

This two-stage algorithm leverages category structure to streamline causal discovery in basket data. First, it aggregates product-level purchases into category-level totals, identifying pairs of categories whose aggregates are uncorrelated. Proposition 3 ensures that no causal links exist between products in these category pairs. Second, it eliminates all product-level causal edges between such independent categories, significantly reducing the complexity of the causal graph. Finally, a standard causal-discovery algorithm (e.g., PC algorithm) is applied to this reduced, constrained graph, efficiently identifying relevant causal relationships without sacrificing granularity.

7.2. Assortment Optimization Performance of CPN and MNL Models

Using the Numerator basket-shopping dataset, we examine the value of incorporating cross-category causal effects into assortment optimization. Our main results are that: (1) the CPN model results in an increase of 8.3% to 20.6% in total number of units sold across various problem scenarios compared to the MNL model, and (2) the complementarity and substitution effects in the CPN model lead to significantly different optimal assortments compared to the MNL model.

For this analysis, we analyze the 200 unique products spanning 25 distinct departments across the five themes that constitute the dataset used for testing our hypotheses.[4] Similar to the approach used previously, we randomly select 1,000 customers, extracting a total of 30,583 baskets containing at least one product from our selected product set. We then divide these baskets into two distinct samples: a graph learning sample (15,296 baskets) and a parameter estimation sample (15,297 baskets). First, we employ Algorithm 2 to estimate the causal graph structure $\mathcal{G}$ via the PC algorithm, applying a Euclidean distance-based conditional independence test. Since each shopping basket typically includes only a small subset of the 200 products, we restrict the maximum conditioning set size to three. This choice follows a standard practice in causal discovery as conditioning on up to three variables has been shown to capture most of the underlying causal structure while preserving statistical power (Spirtes et al. 2000a, Uhler et al. 2013, Kalisch and Bühlman 2007). Second, we estimate the base purchase probabilities $\boldsymbol{\alpha}$ and causal effect parameters $\boldsymbol{\beta}$ over the graph $\mathcal{G}$ using Equation 1 on the parameter estimation sample. Finally, we use this estimated demand model and causal graph as input to solve the assortment optimization problem under CPN subject to a parameter K for the number of products to be stocked.

We compare the above solution with the optimal assortment under the traditional MNL model. For this, we first use the same purchase data as above to estimate customer preference parameters separately for each product category. Then, we implement a multi-category greedy optimization algorithm to find the optimal assortment given the required number of products K. Our algorithm for solving the CPN-based and the MNL-based assortment optimization problems are presented in Appendix G.1 and Appendix G.2, respectively.

Table 7 Comparison of the CPN- and MNL-based assortment strategies in expected sales

Assortment Size	$K = 10$	$K = 15$	$K = 20$	$K = 30$	$K = 50$
CPN	1.17	1.49	1.74	2.12	2.62
MNL	0.97	1.25	1.52	1.90	2.42
Percentage Difference	20.62%	19.20%	14.47%	11.58%	8.26%

Note: The percentage difference between the CPN and MNL models is calculated as (Expected Sales from CPN/Expected Sales from $MNL - 1) \times 100\%$.

To compare the two assortment strategies, we compute optimal assortments and corresponding total expected demand under both the CPN and MNL models across varying assortment sizes

[4] In Data I and II, certain departments appear in multiple themes. After removing these duplicates and their associated products, we are left with 200 unique products across the 25 departments.

($K = 10, 15, 20, 30, 50$). Table 7 presents the expected sales, calculated under the CPN demand model, for assortments generated by both the CPN and MNL approaches. As expected, the CPN-based assortments consistently outperform those generated by the MNL model. We find that the amount of improvement ranges from 8.26% to 20.62%, and as the assortment size K becomes smaller or more constrained, the advantage of explicitly modeling causal relationships increases. We believe this occurs because product selection becomes more critical in smaller assortments, whereas larger assortments inevitably include more products that are strongly connected through causal relationships even when those relationships were not explicitly included in the model.

We further find that the choice of products in each category varies across the two models due to the effect of cross-category complementarities. For example, the produce category illustrates this difference most prominently. The MNL assortments consistently include fewer produce items compared to the CPN assortments across all assortment sizes. This difference is directly attributable to the explicit consideration of causal relationships between products by the CPN model. Specifically, items such as *Steak* and *Ground Beef*, included in both assortments, exhibit strong positive causal effects on several produce items. For instance, the causal effect coefficients from these proteins to produce are as follows: $\beta_{\text{Steak}\to\text{Potatoes}} = +0.069$, $\beta_{\text{Steak}\to\text{Lettuce}} = +0.011$, $\beta_{\text{Steak}\to\text{Strawberries}} = +0.021$, $\beta_{\text{Ground Beef}\to\text{Potatoes}} = +0.021$, $\beta_{\text{Ground Beef}\to\text{Lettuce}} = +0.041$, and $\beta_{\text{Ground Beef}\to\text{Strawberries}} = +0.017$. Because the CPN optimization incorporates these positive spillover effects, it systematically includes *Potatoes*, *Lettuce*, and *Strawberries*, whereas the MNL model, evaluating products solely on individual demand without considering these network effects, consistently omits these produce items from its selected assortments.

8. Conclusion

In conclusion, our paper takes the first step towards using causal structure learning to model basket-shopping behavior. Our analysis suggests several important directions for future research. First, our model can be enriched by incorporating additional retail data, such as prices, promotions, store attributes and workforce decisions. Expanding the model to include these factors would allow for different kinds of specific applications to investigate factors that influence consumer behavior, potentially leading to even more effective assortment strategies and more accurate identification of causal relationships. Second, our study assumes a homogeneous customer population. Future research could explore differences in consumer preferences by estimating our model separately for different customer segments or types of retailers. Such empirical research would allow the model to capture individual-level variations, enabling more personalized and accurate predictions of basket-shopping. Finally, while our approach leverages observational data to infer causal relationships, it lacks the experimental validation typically required to ascertain causality. Future research can focus on conducting controlled field experiments to investigate the causal networks identified through our model or to potentially integrate field experiments with our historical data-based approach.

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216.

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* 97(18):10101–10106.

Anscombe FJ (1948) The transformation of poisson, binomial and negative-binomial data. *Biometrika* 35(3/4):246–254.

Aouad A, Feldman J, Segev D, Zhang DJ (2024) The click-based mnl model: A framework for modeling click data in assortment optimization. *Management Science* .

Aouad A, Segev D (2023) The stability of mnl-based demand under dynamic customer substitution and its algorithmic implications. *Operations Research* 71(4):1216–1249.

Bai Y, Feldman J, Segev D, Topaloglu H, Wagner L (2023) Assortment optimization under the multi-purchase multinomial logit choice model. *Operations Research* ISSN 0030-364X.

Burnham KP, Anderson DR (2004) Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research* 33(2):261–304.

Börsch-Supan A (1990) On the compatibility of nested logit models with utility maximization. *Journal of Econometrics* 43(3):373–388, ISSN 0304-4076.

Cachon GP, Kök AG (2007) Category management and coordination in retail assortment planning in the presence of basket shopping consumers. *Management Science* 53(6):934–951, ISSN 0025-1909.

Chintala SC, Liaukonytė J, Yang N (2023) Browsing the aisles or browsing the app? how online grocery shopping is changing what we buy. *Marketing Science* ISSN 0732-2399.

Dhar V, Geva T, Oestreicher-Singer G, Sundararajan A (2014) Prediction in economic networks. *Information Systems Research* 25(2):264–284, ISSN 1047-7047.

Eberhardt F (2017) Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics* 3:81–91, ISSN 2364-415X.

Eberhardt F, Kaynar N, Siddiq A (2024) Discovering causal models with optimization: Confounders, cycles, and instrument validity. *Management Science* .

Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. *Psychometrika* 1(3):211–218.

Gabel S, Guhl D, Klapper D (2019) P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research* 56(4):557–580.

Glymour C, Zhang K, Spirtes P (2019) Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10:524, ISSN 1664-8021.

Gopalakrishnan S, Matta M, Yourdshahy MI, Choudhary V (2023) Go wide or go deep? assortment strategy and order fulfillment in online retail. *Manufacturing & Service Operations Management* 25(3):846–861.

Hall-Hoffarth EW (2022) Causal discovery of macroeconomic state-space models .

Honhon D, Gaur V, Seshadri S (2010) Assortment planning and inventory decisions under stockout-based substitution. *Operations research* 58(5):1364–1379.

Huang Z, Zeng DD, Chen H (2007) Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management science* 53(7):1146–1164, ISSN 0025-1909.

Jacobs BJ, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. *Marketing Science* 35(3):389–404.

Jagabathula S, Mitrofanov D, Vulcano G (2022) Personalized retail promotions through a directed acyclic graph–based representation of customer preferences. *Operations Research* 70(2):641–665.

Jagabathula S, Vulcano G (2018) A partial-order-based model to estimate individual preferences using panel data. *Management Science* 64(4):1609–1628, ISSN 0025-1909.

Jasin S, Lyu C, Najafi S, Zhang H (2023) Assortment optimization with multi-item basket purchase under multivariate mnl model. *Manufacturing & Service Operations Management* ISSN 1523-4614.

Kalisch M, Bühlman P (2007) Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* 8(3).

Kaynar N, Mitrofanov D (2024) Long-term policy impact estimation with causal structure learning .

Lo V, Topaloglu H (2022) Omnichannel assortment optimization under the multinomial logit model with a features tree. *Manufacturing & Service Operations Management* 24(2):1220–1240, ISSN 1523-4614.

Maathuis MH, Kalisch M, Bühlmann P (2009) Estimating high-dimensional intervention effects from observational data .

Manchanda P, Ansari A, Gupta S (1999) The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing science* 18(2):95–114.

Mani V, Thomas DJ, Bansal S (2022) Estimating substitution and basket effects in retail stores: Implications for assortment planning. *Management Science* 68(7):5002–5024.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, Peter Bühlmann (2012) Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47(11):1–26.

Martinoli M, Moneta A, Pallante G (2023) Calibration and validation of macroeconomic simulation models: A general protocol by causal search. *Available at SSRN 4325509* .

McFadden D, Train K (2000) Mixed mnl models for discrete response. *Journal of applied Econometrics* 15(5):447–470, ISSN 0883-7252.

Meek C (2013) Causal inference and causal explanation with background knowledge .

Oestreicher-Singer G, Sundararajan A (2012) The visible hand? demand effects of recommendation networks in electronic markets. *Management science* 58(11):1963–1981, ISSN 0025-1909.

Pearl J (2009) *Causality* (Cambridge university press).

Ruiz FJ, Athey S, Blei DM (2020) Shopper. *The Annals of Applied Statistics* 14(1):1–27.

Sapra A, Kumar S (2023) Optimal joint assortment for an omni-channel retailer. *Information Systems Research* ISSN 1047-7047.

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web*, 285–295.

Song I, Chintagunta PK (2006) Measuring cross-category price effects with aggregate store data. *Management Science* 52(10):1594–1609, ISSN 0025-1909.

Spirtes P, Glymour C, Scheines R, Kauffman S, Aimale V, Wimberly F (2000a) Constructing bayesian network models of gene expression networks from microarray data .

Spirtes P, Glymour CN, Scheines R (2000b) *Causation, prediction, and search* (MIT press), ISBN 0262194406.

Székely GJ, Rizzo ML (2014) Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* 42(6):2382–2412.

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6):2769–2794.

Tulabandhula T, Sinha D, Karra SR, Patidar P (2023) Multi-purchase behavior: Modeling, estimation, and optimization. *Manufacturing & Service Operations Management* 25(6):2298–2313, ISSN 1523-4614.

Uhler C, Raskutti G, Bühlmann P, Yu B (2013) Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* 436–463.

van Ryzin G, Mahajan S (1999) On the relationship between inventory costs and variety benefits in retail assortments. *Management Science* 45(11):1496–1509.

Verma T, Pearl J (1990) Equivalence and synthesis of causal models. *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, 255–270.

Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, 91–109 (Springer).

Wang X, Pan W, Hu W, Tian Y, Zhang H (2015) Conditional distance correlation. *Journal of the American Statistical Association* 110(512):1726–1734.

Xia F, Chatterjee R, May JH (2019) Using conditional restricted boltzmann machines to model complex consumer shopping patterns. *Marketing Science* 38(4):711–727.

Zhang K, Peters J, Janzing D, Schölkopf B (2011) Kernel-based conditional independence test and application in causal discovery. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, 804–813 (AUAI Press).

Appendix A: Theoretical Foundations of Causal Structure Learning and the PC Algorithm

Causal structure learning seeks to infer causal relationships from observational data by exploiting patterns of statistical associations and conditional independencies. Inferring causal relationships solely from observational data—without experimental intervention—is inherently challenging. In this section, we review the theoretical foundations of the data-generating mechanism and the key assumptions that enable such inference. Our aim here is to provide a brief overview of the theory behind causal structure learning; for a comprehensive discussion, see Spirtes et al. (2000b) and Eberhardt (2017).

In causal modeling, the latent graph $\mathcal{G} = (V, E)$ represents the underlying structural-equation model whose joint distribution gives rise to the data we observe. Here, V denotes the set of vertices (each corresponding to a random variable), and E denotes the set of directed edges (each encoding a direct causal relationship). In our setting, V represents the set of product purchase variables that we aim to model. Specifically, we define $V = \{X_{ij} \mid i \in I_j,\ j \in J\}$, where each X_{ij} is a random variable indicating the quantity of product i purchased from category j. Given the causal graph $\mathcal{G} = (V, E)$, the parent set $pa_{\mathcal{G}}(X_{ij})$ of a node $X_{ij} \in V$ includes all products whose purchases directly influence the purchase decision of product X_{ij}. Formally, $pa_{\mathcal{G}}(X_{ij}) = \{X_{i'j'} \mid (X_{i'j'} \to X_{ij}) \in E\}$.

In practice, the underlying causal graph $\mathcal{G}$ is not directly observed; we only observe basket-level purchase data, which we assume are generated according to $\mathcal{G}$. A key assumption underlying causal structure learning is that each variable is generated as a function of its parents plus an independent noise term. Specifically, we assume a structural equation model (SEM) of the form

$$X_{ij} = f_{ij}\big(pa_{\mathcal{G}}(X_{ij})\big) + \epsilon_{ij}, \quad \forall i \in I_j, j \in J,$$

Here, $f_{ij}(\cdot)$ is a deterministic function and ϵ_{ij} is a noise term statistically independent of all other variables and noise terms.[5]

Given the close correspondence between the latent graph $\mathcal{G}$ and the data-generation process, the observed basket-level purchases can be viewed as independent realizations drawn from the structural model over $\{X_{ij}\}$. Causal structure learning methods aim to infer the latent graph $\mathcal{G}$ based on patterns of conditional independence detected in the data. Specifically, by systematically testing conditional independence relations among the observed product purchases, causal structure learning algorithms seek to recover the latent causal structure $\mathcal{G}$ that best explains these observational dependencies.

We define $P_{\mathcal{G}}(V)$ as the joint probability distribution over all product purchases. It specifies how likely different realizations of purchases are under the data-generating process implied by the graph $\mathcal{G}$. Next, we introduce the causal Markov condition, a key assumption that links the structure of $\mathcal{G}$ to the joint distribution $P_{\mathcal{G}}(V)$. Under this condition, the joint distribution factorizes so that each variable is conditionally independent of all its non-descendants given its parent set—thereby specifying the data-generating mechanism and yielding the conditional-independence relations we test to recover the true causal graph.

[5] Most causal structure learning methods do not require imposing any parametric form on f_{ij}; for estimation, however, we impose a linear form for estimation in the next section.

Assumption 2 (Causal Markov Condition)**.** *Each variable $X_{ij} \in V$ is conditionally independent of its non-descendants given its parent set $pa_{\mathcal{G}}(X_{ij})$ in the causal graph $\mathcal{G} = (V, E)$.*

The causal Markov condition, following from the probability distribution definition in terms of the causal structure (Pearl 2009), enables derivation of observed probabilistic independencies from the causal graph.

We next show how graph's structure specifically affects the joint probability distribution. Under the Causal Markov assumption, the joint distribution $P_{\mathcal{G}}(X_{i_1 j_1}, \dots, X_{i_{|V|} j_{|V|}})$ decomposes into the product of each variable's conditional distribution given its parent set. We now derive this graph-driven factorization.

Applying the chain rule to the joint distribution for any permutation of V,

$$P_{\mathcal{G}}(X_{i_1 j_1}, \dots, X_{i_{|V|} j_{|V|}}) = \prod_{k=1}^{|V|} P_{\mathcal{G}}(X_{i_k j_k} \mid X_{i_1 j_1}, \dots, X_{i_{k-1} j_{k-1}}).$$

Let π be a topological ordering of the nodes in $\mathcal{G}$, so that if $(X_{i'j'} \to X_{ij}) \in E$ then $\pi(i', j') < \pi(i, j)$. Re-expressing the chain rule in this order gives

$$P_{\mathcal{G}}(X_{i_1 j_1}, \dots, X_{i_{|V|} j_{|V|}}) = \prod_{k=1}^{|V|} P_{\mathcal{G}}(X_{\pi(k)} \mid X_{\pi(1)}, \dots, X_{\pi(k-1)}).$$

By the Causal Markov Condition, each $X_{\pi(k)}$ is independent of all its non-parent predecessors once conditioned on its parent set $\mathrm{pa}_{\mathcal{G}}(X_{\pi(k)})$. Hence every factor simplifies to

$$P_{\mathcal{G}}(X_{\pi(k)} \mid X_{\pi(1)}, \dots, X_{\pi(k-1)}) = P_{\mathcal{G}}(X_{\pi(k)} \mid \mathrm{pa}_{\mathcal{G}}(X_{\pi(k)})),$$

and we obtain the desired graph-driven factorization

$$P_{\mathcal{G}}(X_{i_1 j_1}, \dots, X_{i_{|V|} j_{|V|}}) = \prod_{k=1}^{|V|} P_{\mathcal{G}}(X_{\pi(k)} \mid \mathrm{pa}_{\mathcal{G}}(X_{\pi(k)})).$$

By eliminating the reference to the topological ordering, the factorization can be written as

$$\prod_{(i,j) \in V} P_{\mathcal{G}}(X_{ij} \mid \mathrm{pa}_{\mathcal{G}}(X_{ij})).$$

This factorization establishes a direct link between the graph structure and the observable joint distribution. We then invoke the *faithfulness* assumption, which, together with the Causal Markov assumption, allows us to infer the underlying causal graph from the conditional-independence relations testable in data.

The conditional independence relations observed in the data must exactly match those entailed by the causal graph structure: no spurious independencies may occur, and none implied by the graph may be absent. We formalize this requirement as the faithfulness assumption, which guarantees that the data exhibit precisely the conditional independence patterns specified by the graph. This assumption ensures that observed independencies reflect the underlying causal structure rather than, for example, exact cancellations between causal pathways (Spirtes et al. 2000a).

Assumption 3 (Faithfulness)**.** *The only independencies present in the probability distribution are those implied by the graph structure through the causal Markov condition.*

Lastly, the causal sufficiency assumption ensures that all common causes of any pair of variables in the set V are also contained within V, thereby excluding the existence of any unobserved confounders.

Assumption 4 (Causal Sufficiency). *No variable external to the set V directly affects more than one variable contained within V.*

These assumptions establish a correspondence between the observed data and the underlying graph, enabling causal structure learning algorithms to leverage probabilistic independence relations implied by the data to learn the underlying causal relations. The following theorem formalizes this correspondence:

Theorem 1. ***(Spirtes et al. 2000b)***. *Let graph $\mathcal{G}$ be the true data generating graph. Under the causal Markov condition, faithfulness and causal sufficiency assumptions, we have:*

(i) for all $X_{ij}, X_{i'j'} \in V$, X_{ij} and $X_{i'j'}$ are adjacent in true graph $\mathcal{G}$ if and only if X_{ij} and $X_{i'j'}$ are dependent conditional on any $C \in \mathcal{C}_{(ij)(i'j')}$ and

(ii) for all $X_{ij}, X_{i'j'}, X_{i''j''} \in V$ such that X_{ij} is adjacent to $X_{i'j'}$, $X_{i'j'}$ is adjacent to $X_{i''j''}$, and X_{ij} and $X_{i''j''}$ are not adjacent in the true graph $\mathcal{G}$, $X_{ij} \rightarrow X_{i'j'} \leftarrow X_{i''j''}$ is in the true graph $\mathcal{G}$ if and only if X_{ij} and $X_{i''j''}$ are dependent conditional on every set $C \in \mathcal{C}_{(ij)(i''j'')}$ such that $X_{i'j'} \in C$,

where $\mathcal{C}_{(ij)(i'j')} = \{C \mid C \subseteq V \setminus \{X_{ij}, X_{i'j'}\}\}$ stores all possible conditioning sets for $X_{ij}, X_{i'j'} \in V$.

Theorem 1 establishes that both the adjacency relations and the orientation of edges in a graph can be inferred from conditional independence relations. The PC algorithm, which is a popular and widely used method in causal structure learning, utilizes this correspondence between (conditional) independence relations and the underlying graph to iteratively learn the underlying causal relations from observational data (Spirtes et al. 2000b). Specifically, the PC algorithm uses the principles outlined in Theorem 1 and operates in two main phases: the first phase discovers the skeleton of the causal structure without specifying the directions of discovered edges and the second phase handles orienting the edges of the skeleton. The pseudo-codes for these two phases of the PC algorithm are provided in Algorithms A1 and A2, respectively.

In Algorithm A1, we start with a complete undirected graph in which each pair X_{ij} and $X_{i'j'}$ in V is connected by an undirected edge. We then evaluate each pair of variables, X_{ij} and $X_{i'j'}$, for probabilistic independence by considering subsets C in $\mathcal{C}_{(ij)(i'j')} = \{C \mid C \subseteq V \setminus \{X_{ij}, X_{i'j'}\}\}$, i.e., all subsets of V excluding the focal pair of variables. The size of each subset, denoted by $|C|$, incrementally increases throughout the testing process. If X_{ij} and $X_{i'j'}$ are found to be independent given a subset C, we remove the edge between them, following part (i) of Theorem 1. The output of Algorithm A1 is the skeleton of the causal structure, which represents the potential causal relations without directionality among the variables and it serves as the input for the next phase of the algorithm. Algorithm A2 focuses on orienting the edges within the graph following part (ii) of Theorem 1. We proceed by examining conditional independence relations between variable triplets $X_{ij}, X_{i'j'}, X_{i''j''} \in V$ such that X_{ij} is adjacent to $X_{i'j'}$ and $X_{i'j'}$ is adjacent to $X_{i''j''}$ and X_{ij} and $X_{i''j''}$ are not adjacent in the skeleton returned by Algorithm A1. If X_{ij} and $X_{i''j''}$ are probabilistically dependent with respect to conditioning set $C = \{X_{i'j'}\}$ then we orient $X_{ij} — X_{i'j'} — X_{i''j''}$ as $X_{ij} \rightarrow X_{i'j'} \leftarrow X_{i''j''}$. The algorithm then employs Meek's rules for further edge orientation, ensuring that all the orientations are consistent with Theorem 1 and the resulting graph does not include cycles (Meek 2013). The PC algorithm offers flexibility in selecting independence tests appropriate for the specific domain and enables the choice of preferred methods for correcting multiple hypothesis testing. These choices depend on factors such as sample

Algorithm A1: PC algorithm - Skeleton Discovery (Spirtes et al. 2000b)

Input: I_j, J and $V = \{X_{ij}$ for $i \in I_j, j \in J\}$.
Output: $\mathcal{G} = (V, E)$.
Initialization: $\mathcal{G} = (V, E)$ where $E = \left\{(X_{ij} — X_{i'j'}),\ \ \forall X_{ij}, X_{i'j'} \in V \text{ such that } X_{ij} \neq X_{i'j'}\right\}$,
$\mathcal{N}_{\mathcal{G}}(X_{ij}) = \{X_{i'j'} \in V \setminus X_{ij}\}$.
1. Initialize $\ell = 0$.
2. **for** an ordered pair $X_{ij}, X_{i'j'} \in V$ where $X_{i'j'} \in \mathcal{N}_{\mathcal{G}}(X_{ij})$ and $|\mathcal{N}_{\mathcal{G}}(x_{ij}) \setminus X_{i'j'}| \geq l$:
3. **for** $C \in C_{(ij)(i'j')}$ with $|C| = \ell$:
4. **if** $X_{ij} \perp\!\!\!\perp X_{i'j'} \mid C$:
5. Remove the edge $(X_{ij} — X_{i'j'})$ from E.
6. Update $\mathcal{N}_{\mathcal{G}}(X_{ij}) = \mathcal{N}_{\mathcal{G}}(X_{ij}) \setminus X_{i'j'}$ and $\mathcal{N}_{\mathcal{G}}(X_{i'j'}) = \mathcal{N}_{\mathcal{G}}(X_{i'j'}) \setminus X_{ij}$.
7. **Break**.
8. **if** $\ell < |V| - 2$:
9. Update $\ell = \ell + 1$. Go to Step 2.
10. **else** $\ell = |V| - 2$:
11. **Break**.
12. Return $\mathcal{G} = (V, E)$.

Algorithm A2: PC algorithm - Orienting Edges (Spirtes et al. 2000b)

Input: $\mathcal{G} = (V, E)$.
Output: $\mathcal{G} = (V, E)$.
1. **for** $X_{ij}, X_{i'j'}, X_{i''j''} \in V$ where $\{(X_{ij} — X_{i'j'}), (X_{i'j'} — X_{i''j''})\} \subseteq E$ & $\{(X_{ij} — X_{i''j''})\} \nsubseteq E$:
2. **if** $X_{ij} \not\perp\!\!\!\perp X_{i''j''} \mid X_{i'j'}$:
3. Orient $X_{ij} — X_{i'j'} — X_{i''j''}$ as $X_{ij} \rightarrow X_{i'j'} \leftarrow X_{i''j''}$.
4. Update $E = \left(E \setminus \{(X_{ij} — X_{i'j'}), (X_{i'j'} — X_{i''j''})\}\right) \cup \{(X_{ij} \rightarrow X_{i'j'}), (X_{i'j'} \leftarrow X_{i''j''})\}$.

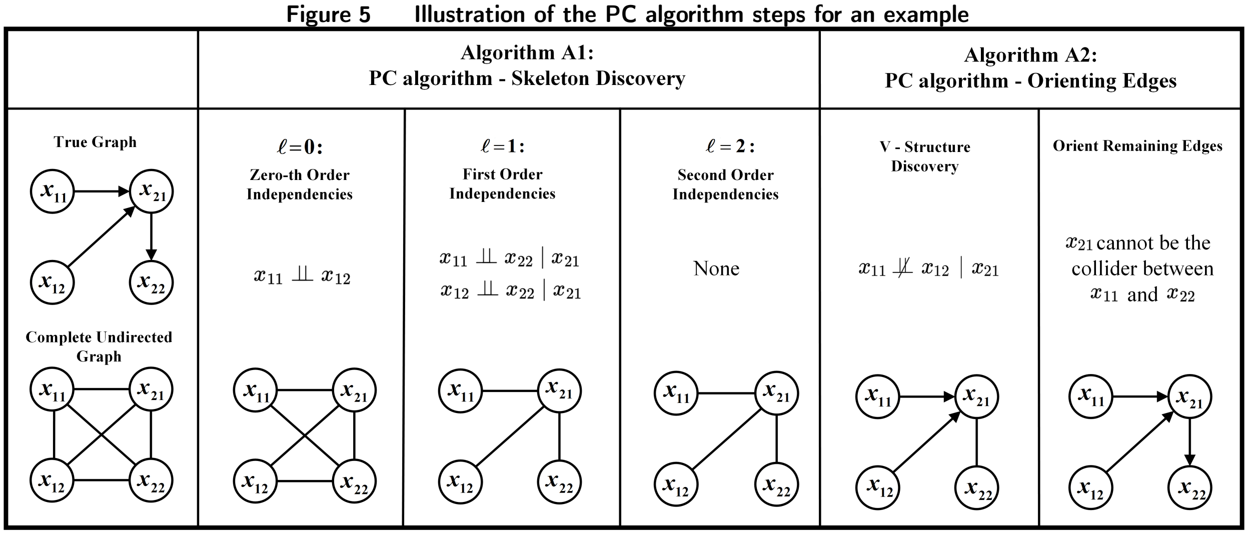
Meek rules (See Figure 6 for illustration):
4. **(R1) for** $X_{ij}, X_{i'j'}, X_{i''j''} \in V$ where $X_{ij} \notin \mathcal{N}_{\mathcal{G}}(X_{i''j''})$ and $\{(X_{ij} \rightarrow X_{i'j'}), (X_{i'j'} — X_{i''j''})\} \subseteq E$:
5. Update $E \leftarrow \left(E \cup \{(X_{i'j'} \rightarrow X_{i''j''})\}\right) \setminus \{(X_{i'j'} — X_{i''j''})\}$.
6. **(R2) for** $X_{ij}, X_{i'j'}, X_{i''j''} \in V$ where $\{(X_{ij} \rightarrow X_{i'j'}), (X_{i'j'} \rightarrow X_{i''j''}), (X_{ij} — X_{i''j''})\} \subseteq E$:
7. Update $E \leftarrow \left(E \cup \{(X_{ij} \rightarrow X_{i''j''})\}\right) \setminus \{(X_{ij} — X_{i''j''})\}$.
8. **(R3) for** $X_{ij}, X_{i'j'}, X_{i''j''}, x_{i'''j'''} \in V$ where $X_{i'j'} \notin \mathcal{N}_{\mathcal{G}}(x_{i'''j'''})$ &
9. $\{(X_{ij} — X_{i'j'}), (X_{i'j'} \rightarrow X_{i''j''}), (X_{ij} — x_{i'''j'''}), (x_{i'''j'''} \rightarrow X_{i''j''}), (X_{ij} — X_{i''j''})\} \subseteq E$:
10. Update $E \leftarrow \left(E \cup \{(X_{ij} \rightarrow X_{i''j''})\}\right) \setminus \{(X_{ij} — X_{i''j''})\}$.
11. **(R4) for** $X_{ij}, X_{i'j'}, X_{i''j''}, x_{i'''j'''} \in V$ where $X_{ij} \notin \mathcal{N}_{\mathcal{G}}(X_{i''j''})$ &
12. $\{(X_{ij} \rightarrow X_{i'j'}), (X_{i'j'} \rightarrow X_{i''j''}), (X_{ij} — x_{i'''j'''}), (x_{i'''j'''} — X_{i''j''}), (X_{i'j'} — x_{i'''j'''})\} \subseteq E$:
13. Update $E \leftarrow \left(E \cup \{(x_{i'''j'''} \rightarrow X_{i''j''})\}\right) \setminus \{(x_{i'''j'''} — X_{i''j''})\}$.
14. **Return** $\mathcal{G} = (V, E)$ where $E = E$.

size, the number of variables, whether the variables are categorical or continuous, and the assumptions made about the parametric nature of the causal relationships.

In general, the independence structure seen in observational data is not guaranteed to uniquely identify the underlying causal graph. Two graphs with different structures are said to be *Markov equivalent* if they have the same independence structure (Verma and Pearl 1990). In the context of causal structure learning, the Markov equivalence class of the true, data-generating graph is the limit of what can be learned about the causal structure from the independence structure in the data. The PC algorithm has been shown to be asymptotically correct, meaning in the large-sample limit it discovers the true data-generating graph up to an equivalence class, given Assumptions 2-4 (Spirtes et al. 2000b). This implies that the edge set E returned by Algorithm A1 represents a Markov equivalence class of graphs and it might include undirected edges. In the context of our study, having $(X_{i'j'} \to X_{ij}) \in E$ implies purchasing product i' from category j' has a *direct* causal impact on purchasing product i from category j. Whereas having $(X_{i'j'} — X_{ij}) \in E$ implies an unspecified causal relationship between the products. In other words, $(X_{i'j'} — X_{ij})$ suggests that there is an edge between $X_{i'j'}$ and X_{ij}, but it could point in either direction.

Figure 5 illustrates the steps of the PC algorithm for an example. For a given true data-generating graph over product purchases x_{11} and x_{21} in category 1, and x_{12} and x_{22} in category 2, the PC algorithm conducts a series of conditional independence tests using the data generated accordingly with this graph. Algorithm A1 starts with testing for marginal independencies ($\ell = 0$). Since x_{11} and x_{12} are marginally independent, denoted as $x_{11} \perp\!\!\!\perp x_{12}$, the edge between x_{11} and x_{12} is removed. Then, we start testing for conditional independencies. At the first order ($\ell = 1$), since $x_{11} \perp\!\!\!\perp x_{22} \mid x_{21}$ and $x_{12} \perp\!\!\!\perp x_{22} \mid x_{21}$, the edges between x_{11} and x_{22} and x_{12} and x_{22} are removed. Since no further independencies are found at the second order ($\ell = 2$), we obtain the skeleton of the graph. Using this skeleton, Algorithm A2 orients the edges using the dependency relations. Since $x_{11} \not\perp\!\!\!\perp x_{12} \mid x_{21}$, we must have $x_{11} \to x_{21} \leftarrow x_{12}$ following part (ii) of Theorem 1. Finally, we orient the remaining edges using Meek's rules, resulting in $x_{21} \to x_{22}$, forming the directed acyclic graph that captures the causal relationships. In this example, we are able to identify a unique graph, indicating that there is only one graph in the Markov equivalence class of the underlying graph. [6]

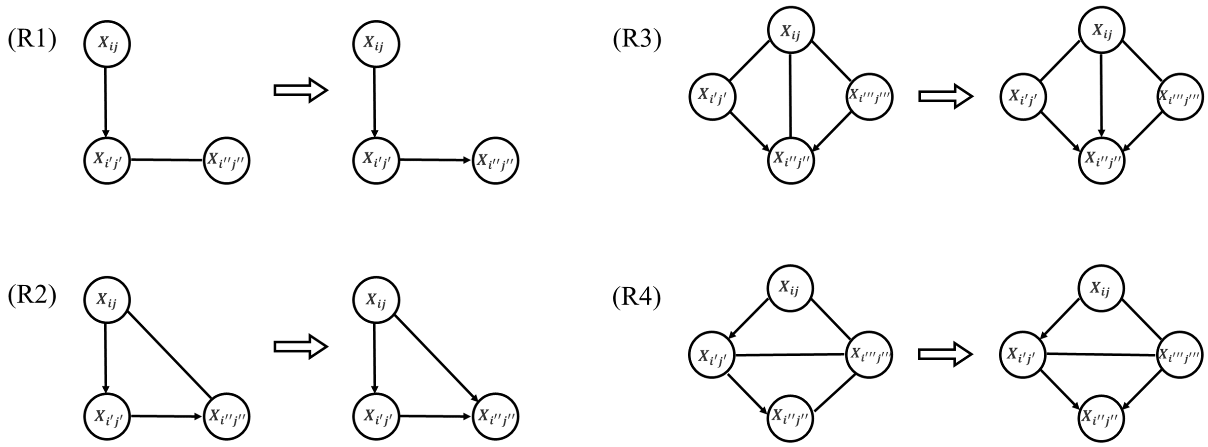
Figure 5 **Illustration of the PC algorithm steps for an example**



Furthermore, Figure 6 graphically illustrates the four Meek rules that correspond to (R1), (R2), (R3), and (R4) in Algorithm A2.

[6] As previously mentioned, unique identification of the causal graphs is not always guaranteed.

Figure 6 **Meek rules**



Appendix B: Proofs

Proof of Proposition 1. Define for each $t \in \{1,\dots,T\}$, $Z_t = R_{ijt} - R_{i'j't}$. Since the difference of two independent normal random variables is also normal, we have: $Z_t \sim \mathcal{N}(\mu_R, \sigma_R^2)$, where $\mu_R = \mu_{R_{ij}} - \mu_{R_{i'j'}}$ and $\sigma_R^2 = \sigma_{R_{ij}}^2 + \sigma_{R_{i'j'}}^2$.

Part (i): Expected Squared Distance. We aim to compute the expectation of the squared Euclidean distance, $\mathbb{E}[d] = \sum_{t=1}^{T} \mathbb{E}[Z_t^2]$. Since each $Z_t \sim \mathcal{N}(\mu_R, \sigma_R^2)$, we use the identity $\mathbb{E}[Z_t^2] = \mathrm{Var}(Z_t) + (\mathbb{E}[Z_t])^2 = \sigma_R^2 + \mu_R^2$. Summing over T independent dimensions, we obtain $\mathbb{E}[d] = T(\mu_R^2 + \sigma_R^2)$.

Part (ii): Variance of Squared Distance. To compute the variance of the squared distance, we use the fact that the Z_t are i.i.d., so $\mathrm{Var}(d) = \sum_{t=1}^{T} \mathrm{Var}(Z_t^2)$. The variance of Z_t^2 requires the fourth moment, which for a normal distribution is $\mathbb{E}[Z_t^4] = \mu_R^4 + 6\mu_R^2\sigma_R^2 + 3\sigma_R^4$. Subtracting the square of the second moment, $\mathbb{E}[Z_t^2]^2 = (\mu_R^2 + \sigma_R^2)^2 = \mu_R^4 + 2\mu_R^2\sigma_R^2 + \sigma_R^4$, gives $\mathrm{Var}(Z_t^2) = 4\mu_R^2\sigma_R^2 + 2\sigma_R^4$. Therefore, $\mathrm{Var}(d) = T(4\mu_R^2\sigma_R^2 + 2\sigma_R^4)$.
In conclusion, $\mathbb{E}[d] = T(\mu_R^2 + \sigma_R^2)$ and $\mathrm{Var}(d) = T(4\mu_R^2\sigma_R^2 + 2\sigma_R^4)$. □

Proof of Proposition 2. Algorithm 1 proceeds in three main stages:

Case 1: $|C| > 0$.

1. QR Decomposition of the Conditioning Matrix. Form the matrix $W = [\, X : X \in C \,] \in \mathbb{R}^{T \times |C|}$. A full QR decomposition $W = Q\,R$ on a $T \times |C|$ matrix requires $O\big(T\,k^2\big)$ arithmetic operations where $k = |C|$.

2. Residualization of Each Test Vectors For each $x \in V \setminus C$, we compute the residual $R_x = x - Q(Q^\top x)$. Forming $Q^\top x \in \mathbb{R}^{|C|}$ requires $O(T\,k)$ operations, and multiplying $Q(Q^\top x) \in \mathbb{R}^T$ costs another $O(T\,k)$. Therefore, computing each residual costs $O(T\,k)$. Since there are $N - k$ test vectors, obtaining the residuals of all the test vectors costs $O(T\,k(N-k))$.

3. Squared Euclidean Distance Computation. Once the residuals have been obtained, we compute $d = \sum_{t=1}^{T} \big(R_{ijt} - R_{i'j't}\big)^2$, for all pairs of residual vectors associated with the test vectors in $V \setminus C$. Each such computation costs a single pass through T entries, i.e. $O(T)$. Since there are $O((N-k)^2)$ such pairwise computations, the total burden for computing all the squared Euclidean distances is $O(T(N-k)^2)$.

Summing these three stages, we obtain $O\big(Tk^2\big) + O\big(Tk(N-k)\big) + O(T(N-k)^2) = O\big(TN^2\big)$. Therefore, the overall complexity is $O\big(TN^2\big)$.

Case 2: $|C| = 0$.

When the conditioning set is empty, no QR decomposition or residualization is needed; the algorithm reduces to computing the squared Euclidean distance between all pairs X_{ij} and $X_{i'j'}$ in V. Each such computation requires a single pass through T entries and costs $O(T)$. Since there are $O(N^2)$ pairs, the total complexity is again $O\big(TN^2\big)$. □

Proof of Proposition 3. Assume, for contradiction, that the category aggregates are uncorrelated, $\mathrm{Cov}(W_j, W_{j'}) = 0$, while there exists at least one causal edge between a product in category j and a product in category j'; without loss of generality let that edge be $X_{ij} \to X_{i'j'}$ for some $i \in \mathcal{I}_j$ and $i' \in \mathcal{I}_{j'}$.
Because the SEM is faithful, the edge $X_{ij} \to X_{i'j'}$ results in statistical dependence between the two variables, hence we must have (i) $\mathrm{Cov}(X_{ij}, X_{i'j'}) \neq 0$. Yet Assumption 1 (No Exact Cancellation of Product-Level Covariances) states that $\mathrm{Cov}(W_j, W_{j'}) = 0$ can occur only when (ii) $\mathrm{Cov}(X_{kj}, X_{k'j'}) = 0$ for all $k \in \mathcal{I}_j$, $k' \in \mathcal{I}_{j'}$. In particular, (i) asserts $\mathrm{Cov}(X_{ij}, X_{i'j'}) = 0$.
Statements (i) and (ii) cannot both be true, so our initial assumption must be false. Therefore, if $\mathrm{Cov}(W_j, W_{j'}) = 0$, no causal edge can connect any product in category j to any product in category j'. □

Appendix C: Supplement to the Numerical Validation of the Euclidean Distance-Based Test

In this Appendix, we present the details of the numerical validation of the Euclidean Distance-based Test presented in Section 4 of the paper.

Our analysis proceeds in the following stages as summarized in Algorithm C1. First, we construct a ground-truth causal product network $\mathcal{G}$ by generating a random directed acyclic graph (DAG) over 4-12 products. Products are grouped into 2-4 categories (each containing 2-3 products), and edges are assigned between products based on an edge inclusion probability sampled uniformly from 0.3 to 0.8. The resulting graph defines the causal structure across products, with directed edges indicating potential influence on purchasing behavior. Then, we parameterize this graph by assigning each product a base purchase probability α_{ij}, drawn uniformly from $[0, \alpha_H]$. For each parent of a product in the graph, we draw a causal effect coefficient $\beta^{ij}_{i'j'}$ from a uniform distribution on $[0, \beta_H]$. Second, we simulate synthetic basket data based on the specified structural equation model. Because we aim to model discrete purchase counts, we draw each product's count from a Poisson distribution, where the rate equals its base purchase probability plus the cumulative influence from its purchased parent products (as defined in Equation (1)). We simulate a total of $2T$ baskets where $T \sim U(10^3, 10^4)$, and divide them equally into graph learning and parameter estimation samples. Third, we implement a consideration set mechanism to reflect the possibility that not all products are evaluated in every basket. Specifically, in a randomly selected proportion p_2 of graph learning sample, each product is independently excluded from the consideration set with probability p_1. If a product is excluded, its purchase count is set to zero, and it does not exert any causal influence on downstream products in the graph.

Algorithm C1: Causal Effect Estimation Procedure on Simulated Basket Shopping Data

Inputs: α_H (upper bound on base purchase probabilities), β_H (upper bound on causal effect sizes), p_1 (product-level consideration set dropout probability).

Outputs: True graph $\mathcal{G}$, learned graphs $\bar{\mathcal{G}}_E$, $\bar{\mathcal{G}}_{PC}$, and estimation errors err_β^E, err_β^{PC}.

Step 1: Graph Generation and Parameterizing the Graph

1. Sample number of categories $m \sim U\{2,4\}$ and number of products $n_j \sim U\{2,3\}$ for $j \in \{1,\dots,m\}$.
2. Sample edge inclusion probability $p \sim U(0.3, 0.8)$.
3. Construct a random DAG $\mathcal{G}$ with N nodes using edge probability p where $N = \sum_{j=1}^m n_j$.
4. For each product X_{ij}, define its parent set $\mathrm{pa}(X_{ij})$ from $\mathcal{G}$.
5. Let $\mathcal{O} = (X_{[1]}, X_{[2]}, \dots, X_{[N]})$ denote a topological ordering of all products in the graph $\mathcal{G}$, such that for any $X_{ij} \in \mathrm{pa}(X_{i'j'})$, X_{ij} precedes $X_{i'j'}$ in $\mathcal{O}$.
6. For each product $X_{ij} \in \mathcal{O}$:
7. Sample $\alpha_{ij} \sim U(0, \alpha_H)$ and $\beta_{i'j'}^{ij} \sim U(0, \beta_H)$ for $X_{i'j'} \in \mathrm{pa}(X_{ij})$ and $\beta_{i'j'}^{ij} = 0$ otherwise.

Step 2: Simulating Shopping Data

8. Sample number of baskets $T \sim U(10^3, 10^4)$.
9. For each basket $t \in \{1,\dots,2T\}$ and for each product $X_{ij} \in \mathcal{O}$:
10. Generate purchase count $X_{ijt} \sim \mathrm{Poisson}\left(\alpha_{ij} + \sum_{X_{i'j'} \in \mathrm{pa}(X_{ij})} \beta_{i'j'}^{ij} X_{i'j't}\right)$.
11. Split data equally into graph learning and parameter estimation sets.

Step 3: Consideration Set Mechanism

12. Sample $p_2 \sim U(0.3, 0.8)$, the probability that a basket is subject to product consideration dropout.
13. Sample a subset of graph learning baskets $\mathcal{T}_1 \subseteq \{1,\dots,T\}$, where each $t \in \mathcal{T}_1$ with probability p_2.
14. For each basket $t \in \mathcal{T}_1$:
15. Sample subset $\mathcal{N}_t^{\mathrm{off}} \subseteq \mathcal{O}$, where each $X_{ij} \in \mathcal{N}_t^{\mathrm{off}}$ with probability p_1.
16. For each product $X_{ij} \in \mathcal{O}$:
17. $X_{ijt} \leftarrow 0$ if $X_{ij} \in \mathcal{N}_t^{\mathrm{off}}$, else sample
$$X_{ijt} \sim \mathrm{Poisson}\left(\alpha_{ij} + \sum_{X_{i'j'} \in \mathrm{pa}(X_{ij})} \beta_{i'j'}^{ij} X_{i'j't}\right).$$
18. Apply Anscombe transformation to graph learning set for normalization.

Step 4: Graph Recovery, Estimation, and Error Computation

Euclidean Distance-Based Estimation:

19. Run Euclidean distance-based PC algorithm on the graph learning set to get $\bar{\mathcal{G}}_E$.
20. Estimate causal effects $\bar{\beta}_{i'j',E}^{ij}$ on parameter estimation set using graph $\bar{\mathcal{G}}_E$.
21. Compute $err_\beta^E = \left\{\beta_{i'j'}^{ij} - \bar{\beta}_{i'j',E}^{ij}\right\}$.

Partial Correlation-Based Estimation:

22. Run partial correlation-based PC algorithm to get $\bar{\mathcal{G}}_C$.
23. Estimate causal effects $\bar{\beta}_{i'j',C}^{ij}$ on parameter estimation set using graph $\bar{\mathcal{G}}_C$.
24. Compute $err_\beta^C = \left\{\beta_{i'j'}^{ij} - \bar{\beta}_{i'j',C}^{ij}\right\}$.
25. Return $\mathcal{G}$, $\bar{\mathcal{G}}_E$, $\bar{\mathcal{G}}_C$, err_β^E, err_β^C.

Product decisions are updated in topological order to preserve causal ordering. Fourth, we transform the graph learning data using the Anscombe transformation technique (Anscombe 1948) to stabilize variance and approximate normality—assumptions required for both the Euclidean distance–based and partial correlation–

based independence tests.[7] We then recover causal graphs using the PC algorithm with each test, resulting in two estimated graphs: $\bar{\mathcal{G}}_E$ from the Euclidean distance–based test and $\bar{\mathcal{G}}_{PC}$ from the partial correlation test. Using each graph, we estimate causal effects on the test data via structural equation modeling (SEM) and evaluate the estimation accuracy by computing the root mean square error (RMSE) between the estimated and true effect values. For each scenario and each level of p_1, we perform this exercise on 50 samples.

To comprehensively evaluate the performance of the two methods, we consider the following three simulation scenarios that vary base purchase rates and causal effect magnitudes:
(1) **Low-Causality Sparse Case**: a scenario with low base purchase rates and small causal effect sizes;
(2) **Non-Sparse Case**: a scenario with high base purchase rates and small causal effect magnitudes;
(3) **High-Causality Sparse Case**: a scenario with low base purchase rates and large causal effect magnitudes.

We define $\alpha \sim U(0, 0.01)$ as a low base purchase rate and $\alpha \sim U(0, 0.1)$ as high, consistent with empirical findings from real data, where the average quantity purchased per product is below 0.01 per basket across 200 products (see Section 6.1). Similarly, we define small causal effects as $\beta \sim U(0, 1)$ and large effects as $\beta \sim U(0, 2)$. In our setting, observed purchase data are shaped not only by preferences but also by latent consideration sets—subsets of products that consumers evaluate before making a decision. Because these consideration sets are unobserved, as discussed earlier the resulting sparsity in the data does not necessarily indicate true non-purchase, but rather non-consideration. This structured sparsity can distort the conditional independence patterns needed for accurate causal discovery. To model the strength of this latent selection mechanism, we vary the probability $p_1 \in \{0.2, 0.4, 0.6, 0.8\}$, which governs the exclusion of products from the consumer's consideration set. To maintain computational tractability across the resulting 12 configurations, we restrict the number of variables; however, the observed patterns are expected to generalize to larger graphs. By comparing Scenarios 1 and 2, we assess whether the Euclidean distance–based approach remains effective across different base purchase rates, while comparison between Scenarios 1 and 3 isolates its sensitivity to changes in causal effect magnitudes.

The results of this analysis are presented in Section 4 in the main body of the paper.

Appendix D: Structural Equations Models for Network Structures and Causal Representation Specifications

D.1. SEMs for Network Structures

The complete network representation of product interactions is specified by

$$X_{ij} = \alpha_{ij}^{comp} + \sum_{i' \in I_{j'},\, j' \in J} \beta_{i'j'}^{ijcomp} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j,\, j \in J. \tag{2}$$

To induce sparsity, reduce variance in the estimated network, and obtain a more interpretable subset of relevant connections, we estimate the model using Lasso and Ridge regularization techniques by adding an ℓ_1-penalty and an ℓ_2-penalty, respectively, to the OLS loss function.

[7] Testing for conditional independence using partial correlation assumes joint normality. Under multivariate normality, a zero partial correlation implies conditional independence.

Let cor_{ij} store product purchases significantly correlated with the purchase of product i from category j (i.e., p-value ≤ 0.05). The corresponding SEM for the correlation network representation is defined as:

$$X_{ij} = \alpha_{ij}^{cor} + \sum_{X_{i'j'} \in cor_{ij}} \beta_{i'j'}^{ijcor} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J. \tag{3}$$

However, in the presence of sparse basket data, many pairwise correlations may be based on limited co-occurrence and thus subject to high variance and overfitting. To mitigate this, we apply matrix decomposition techniques—specifically, singular value decomposition (SVD) and eigenvalue decomposition (EVD)—to obtain denoised, low-rank approximations of the correlation matrix. These methods are commonly used to reduce noise and reveal the underlying network structure, including correlation structures (Eckart and Young 1936, Alter et al. 2000, Wall et al. 2003). The resulting network structures, derived from the denoised matrices, are the SVD-decomposed correlation network and the EVD-decomposed correlation network, respectively.

Let sim_{ij} denote the set of product purchases that have high cosine similarity with product X_{ij}, where i indexes the product and j the category. Let $SimScore(X_{ij}, X_{i'j'})$ represent the cosine similarity score between product purchases X_{ij} and $X_{i'j'}$. Following the standard item-item regression-based collaborative filtering approach (Sarwar et al. 2001), we rank all $X_{i'j'}$ by $SimScore(X_{ij}, X_{i'j'})$, and select the top-$K$ most similar products to form the similarity set sim_{ij}. The representation model is defined as follows:

$$X_{ij} = \alpha_{ij}^{sim} + \sum_{X_{i'j'} \in sim_{ij}} \beta_{i'j'}^{ijsim} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J. \tag{4}$$

Let $\mathcal{A}_{ij}$ denote the set of products that have strong association rules directed toward product X_{ij}, where i indexes the product and j the category. Let $Assoc(X_{i'j'}, X_{ij})$ indicate the association rule from product $X_{i'j'}$ to X_{ij}, which is retained if it satisfies the minimum support and confidence thresholds (Agrawal et al. 1993). The association set $\mathcal{A}_{ij}$ is defined as the collection of all such $X_{i'j'}$ that have valid association rules pointing to X_{ij}. The association network–based representation model is then defined as follows:

$$X_{ij} = \alpha_{ij}^{assoc} + \sum_{X_{i'j'} \in \mathcal{A}_{ij}} \beta_{i'j'}^{ijassoc} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J. \tag{5}$$

Let $\mathcal{G} = (V, E)$ be the causal graph derived from the PC algorithm and $\mathrm{pa}_{\mathcal{G}}(X_{ij})$ store the parents of the purchase of product i from category j in graph $\mathcal{G}$. The SEM for the causal network representation is defined as:

$$X_{ij} = \alpha_{ij} + \sum_{X_{i'j'} \in \mathrm{pa}_{\mathcal{G}}(X_{ij})} \beta_{i'j'}^{ij} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J. \tag{6}$$

D.2. SEMs for Specification of Causal Representations

Let $\mathcal{G} = (V, E)$ be the causal graph derived by the PC algorithm. First, we introduce the category-level specification. Among the existing causal edges in E, let $E_{j'}^{j} \subseteq E$ store the edges from products in category j' to products in category j, i.e., $E_{j'}^{j} = \{(X_{i'j'} \to x_{ij}) \in E,\ \forall i' \in I_{j'}, i \in I_j\}$. We restrict the coefficients of all the edges in $E_{j'}^{j}$ to be identical for any $j, j' \in J$. The corresponding SEM for this structure is defined as:

$$X_{ij} = \alpha_{ij} + \sum_{X_{i'j'} \in pa_{ij}} \beta_{i'j'}^{ij} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J \tag{7a}$$

$$\text{subject to} \quad \beta_{i_1'j'}^{i_1 j} = \beta_{i_2'j'}^{i_2 j}, \quad i_i \in I_j, i_1' \in I_{j'}, j, j' \in J : (X_{i_1'j'} \to X_{i_1 j}), (X_{i_2'j'} \to X_{i_2 j}) \in E_{j'}^{j}. \tag{7b}$$

Next, we introduce the product-to-category specification. Let $E^{j}_{i'j'} \subseteq E$ store the edges from product i' in category j' to any product in category j, i.e.,$E^{j}_{i'j'} = \{(X_{i'j'} \to X_{ij}) \in E,\ \forall i \in I_j\}$. We restrict the coefficients of all the edges in $E^{j}_{i'j'}$ to be identical for any $i' \in I_{j'}, j, j' \in J$. The corresponding SEM for this structure is defined as:

$$X_{ij} = \alpha_{ij} + \sum_{X_{i'j'} \in pa_{ij}} \beta^{ij}_{i'j'} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J \tag{8a}$$

$$\text{subject to} \quad \beta^{i_1 j}_{i'j'} = \beta^{i_2 j}_{i'j'} \quad i_1, i_2 \in I_j, i' \in I_{j'}, j, j' \in J : (X_{i'j'} \to x_{i_1 j}), (X_{i'j'} \to x_{i_2 j}) \in E^{j}_{i'j'}. \tag{8b}$$

Lastly, we introduce the category-to-product specification. Let $E^{ij}_{j'} \subseteq E$ store the edges from any product in category j' to product i from category j, i.e., $E^{ij}_{j'} = \{(X_{i'j'} \to X_{ij}) \in E,\ \forall i' \in I_{j'}\}$. We restrict the coefficients of all the edges in $E^{ij}_{j'}$ to be identical for any $i \in I_j, j, j' \in J$. The corresponding SEM for this structure is defined as:

$$X_{ij} = \alpha_{ij} + \sum_{X_{i'j'} \in pa_{ij}} \beta^{ij}_{i'j'} X_{i'j'} + \epsilon_{ij}, \quad \forall i \in I_j, j \in J \tag{9a}$$

$$\text{subject to} \quad \beta^{ij}_{i'_1 j'} = \beta^{ij}_{i'_2 j'} \quad i \in I_j, i'_1, i'_2 \in I_{j'}, j, j' \in J : (X_{i'_1 j'} \to X_{ij}), (X_{i'_2 j'} \to X_{ij}) \in E^{ij}_{j'}. \tag{9b}$$

These SEMs correspond to different restrictions on the causal effects in Figure 3(d) as follows: Figure 4(e) visualizes the specification in Equation (7) where the causal effects between categories are identical, i.e., $\beta^{22}_{11} = \beta^{21}_{12} = \omega_1$; Figure 4(f) corresponds to Equation (8) where the causal effects from any product to a category are identical, i.e., $\beta^{21}_{11} = \beta^{22}_{11} = \omega_3$; Figure 4(g) corresponds to Equation (9) where the causal effects from any category to a product are identical, i.e., $\beta^{21}_{11} = \beta^{21}_{12} = \omega_6$. Thus, the three specifications are applied as restrictions to the product-level model presented in Equation (6). This enables us to test Hypothesis 2 using a likelihood ratio test or the AIC criterion.

Appendix E: Product Lists of Five Themes

Tables 8, 9, and 10 display the most frequently purchased products across both the brick-and-mortar and online channels, the brick-and-mortar channel only, and the online channel only, respectively. Due to Numerator's data disclosure policy, we have concealed the brand information and are only displaying the product lists without brand details.

(See next page for the tables.)

Table 8 Top B&M and online products

Theme 1 (Pasta)	Pasta & Noodle	Spaghetti, Dry, Macaroni, Penne, Rotini
	Meat	Steaks, Ground Beef, Chicken, Lunch Packs, Whole Cuts & Roasts, Pork Chop, Sausage-Pork
	Produce	Bananas, Onions, Tomatoes, Apples, Grapes, Avocado, Cucumber, Potatoes, Lettuce, Strawberries
	Dairy	Milk, Coffee Creamers, Shredded Cheese, Cream Cheese, Sour Cream, Greek Yogurt
	Condiments	Nut Butters, Salad Dressings, Salsa, Hummus, Mayonnaise, Jam, Jelly & Marmalades
Theme 2 (QSR)	QSR Beverages	Cola, Pepper-Style Soda, Hot Coffee, Citrus & Berry Soda
	QSR Breakfast	Hash Browns, Donuts, Sausage Breakfast Sandwich, Ham Breakfast Sandwich
	QSR Sandwiches & Wraps	Beef Burger, Crispy Chicken Sandwich
	QSR Sauces & Condiments	Ketchup, BBQ Sauce, Mild Sauce, Ranch Dip, Hot Sauce
	QSR Mexican	Chicken Quesadilla, Steak Taco, Veggie Burrito, Steak Nachos, Steak Burrito
	QSR Snack & Sides	French Fries, Potato Chips
	QSR Desserts	Milk Shake, Pie, Cookies, Ice Cream Cone
	QSR Entrees	Chicken Nuggets, Bone-In Chicken
	QSR Italian	Hand Tossed Pizza, Pan Pizza, Thin Crust Pizza, Breadsticks
	QSR Salads	Chicken Cobb Salad, Chicken Garden Salad, Southwest Chicken Salad
Theme 3 (Bakery)	Bakery Sweet Goods	Donuts, Muffins, Cakes, Cookies, Snack Pies, Pastries, Pies
	In-Store Bakery	Bread & Breadsticks, Rolls, Bagels, Croissants, Tortillas, Italian Bread, Buns
	Packaged Bakery	Tortillas, Bagels, Buns, English Muffins, Rolls, Wheat Bread, White Bread
	Baking & Cooking	Pasta & Pizza Sauces, Sugar, Barbecue Sauce, Baking Chips, Pudding, Custard & Mousse Mix, Marshmallows, Tomato Sauce, Paste & Puree, Gelatin & Jello Mix, Cake Mixes, Dessert Syrups
	Dairy	Milk, Coffee Creamers, Shredded Cheese, Cream Cheese, Sour Cream, Greek Yogurt
Theme 4 (Prepared Food)	Frozen Foods	Frozen Breakfast, Frozen Chicken, French Fries, Single Serve Meals, Packaged Ice Cream, Pizza Bites/Rolls
	Canned	Prepared Beans, Canned Tuna, Canned Tomatoes, Variety Beans, Applesauce, Fruit Cups, Canned Green Beans
	Deli & Prepared Foods	Chicken-Prepared, Deli Salad, Pork-Deli, Ready-Made Sandwiches & Wraps-Prepared, Turkey-Deli, Sushi-Prepared
	Beverages	Cola, Sports Drinks, Citrus & Berry Soda, Still Water, Seltzers & Sparkling Water
	Condiments	Nut Butters, Salad Dressings, Salsa, Hummus, Mayonnaise, Jam, Jelly & Marmalades
Theme 5 (Unprepared Food)	Produce	Bananas, Onions, Tomatoes, Apples, Grapes, Avocado, Cucumber, Potatoes, Lettuce, Strawberries
	Shelf Stable Meals	Canned Soups, Mac & Cheese, Ramen & Noodle Soups, Potato Mixes, Pasta Dishes, Rice Dishes
	Meat	Steaks, Ground Beef, Chicken, Lunch Packs, Whole Cuts & Roasts, Pork Chop, Sausage, Pork-Mainstream
	Herbs & Spices	Mexican Seasoning, Grill Seasoning, Garlic Powder & Garlic Salt, Chili Seasoning
	Condiments	Nut Butters, Salad Dressings, Salsa, Hummus, Mayonnaise, Jam, Jelly & Marmalades

Note: (1) This table displays product names without brand information, and duplicate names across different brands have been removed; (2) The top products are those that appear most frequently in both online and brick-and-mortar baskets.

Table 9 Top B&M products

Theme 1 (Pasta)	Pasta & Noodle	Spaghetti, Penne, Dry, Rotini, Macaroni
	Meat	Chicken, Lunch Packs, Pork-Mainstream, Steaks, Ground Beef, Pork Chop, Sausage-Pork, Whole Cuts & Roasts
	Produce	Cucumber, Apples, Onions, Tomatoes, Potatoes, Bananas, Strawberries, Grapes, Avocado, Lettuce
	Dairy	Greek Yogurt, Shredded Cheese, Coffee Creamers, Milk, Sour Cream, Cream Cheese
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings
Theme 2 (QSR)	QSR Beverages	Citrus & Berry Soda, Pepper-Style Soda, Frappuccino, Cola
	QSR Breakfast	Sausage Breakfast Sandwich, Hash Browns, Donuts, Ham Breakfast Sandwich
	QSR Sandwiches & Wraps	Beef Burger, Crispy Chicken Sandwich
	QSR Sauces & Condiments	Ranch Dip, Mild Sauce, BBQ Sauce, Ketchup, Hot Sauce
	QSR Mexican	Steak Taco, Veggie Burrito, Chicken Quesadilla, Steak Nachos, Chicken Taco
	QSR Snack & Sides	French Fries
	QSR Desserts	Pie, Milk Shake, Cookies
	QSR Entrees	Bone-In Chicken, Chicken Nuggets
	QSR Italian	Breadsticks, Pan Pizza, Thin Crust Pizza, Hand Tossed Pizza
	QSR Salads	Chicken Garden Salad, Chicken Cobb Salad, Southwest Chicken Salad
Theme 3 (Bakery)	Bakery Sweet Goods	Cookies, Packaged Muffins, Snack Pies, Pastries, Pies, Muffins, Donuts, Cakes
	In-Store Bakery	Tortillas, Italian Bread, Buns, Croissants, Bread & Breadsticks, Rolls, Bagels
	Packaged Bakery	Bagels, Buns, Wheat Bread, Sandwich Bread, White Bread, Tortillas, English Muffins, Rolls
	Baking & Cooking	Sugar, Cake Mixes, Gelatin & Jello Mix, Barbecue Sauce, Dessert Syrups, Pudding, Custard & Mousse Mix, Baking Chips, Marshmallows, Pasta & Pizza Sauces, Tomato Sauce, Paste & Puree
	Dairy	Greek Yogurt, Shredded Cheese, Coffee Creamers, Milk, Sour Cream, Cream Cheese
Theme 4 (Prepared Food)	Frozen Foods	French Fries, Syrup Carriers-Frozen Breakfast, All Other Single Serve Meals, Packaged Ice Cream, Nutrition Single Serve Meals, Frozen Chicken, Prepared Entrees-Frozen Breakfast
	Canned	Fruit Cups, Canned Tomatoes, Applesauce, Canned Green Beans, Canned Tuna, Variety Beans, Prepared Beans
	Deli & Prepared Foods	Turkey-Deli, Ready-Made Sandwiches & Wraps-Prepared, Pork-Deli, Chicken-Prepared, Deli Salad, Sushi-Prepared
	Beverages	Still Water, Seltzers & Sparkling Water, Pepper & Skipper Soda, Sports Drinks, Citrus & Berry Soda, Cola
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings
Theme 5 (Unprepared Food)	Produce	Cucumber, Apples, Onions, Tomatoes, Potatoes, Bananas, Strawberries, Grapes, Avocado, Lettuce
	Shelf Stable Meals	Ramen & Noodle Soups, Pasta Dishes, Canned Soups, Rice Dishes, Potato Mixes, Mac & Cheese
	Meat	Chicken, Lunch Packs, Pork-Mainstream, Steaks, Ground Beef, Pork Chop, Sausage-Pork, Whole Cuts & Roasts
	Herbs & Spices	Chili Seasoning, Garlic Powder & Garlic Salt, Mexican Seasoning, Grill Seasoning
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings

Note: (1) This table displays product names without brand information, and duplicate names across different brands have been removed; (2) The top products are those that appear most frequently in brick-and-mortar baskets.

Table 10 Top online products

Theme 1 (Pasta)	Pasta & Noodle	Spaghetti, Penne, Dry, Rotini, Macaroni, Lasagna, Bow-Tie
	Meat	Chicken, Lunch Packs, Pork-Mainstream, Pork Bacon, Steaks, Ground Beef, Sausage-Pork
	Produce	Cucumber, Salad Greens, Apples, Onions, Tomatoes, Bananas, Grapes, Avocado, Lettuce
	Dairy	Cream, Butter, Shredded Cheese, Natural Sliced Cheese, Coffee Creamers, Milk, Cream Cheese
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings
Theme 2 (QSR)	QSR Beverages	Cola, Hot Coffee, Lemonade
	QSR Breakfast	Bagels, Bakery Other, Hash Browns, Muffins
	QSR Sandwiches & Wraps	Grilled Chicken Sandwich, Crispy Chicken Sandwich, Bacon Breakfast Sandwich
	QSR Sauces & Condiments	Hot Sauce, Ketchup, Mild Sauce, Beef Burger, Jam & Jelly
	QSR Mexican	Chicken Quesadilla, Steak Taco
	QSR Snack & Sides	French Fries
	QSR Desserts	Cookies, Milk Shake, Pie
	QSR Entrees	Chicken Nuggets, Bone-In Chicken
	QSR Italian	Hand Tossed Pizza, Breadsticks, Cheese Bread, Pan Pizza
	QSR Salads	Chicken Garden Salad, Chicken Cobb Salad, Southwest Chicken Salad
Theme 3 (Bakery)	Bakery Sweet Goods	Cookies, Packaged Muffins, Snack Pies, Muffins, Donuts, Pies
	In-Store Bakery	Tortillas, Buns, Croissants, Bread & Breadsticks, Rolls, Ciabatta
	Packaged Bakery	Bagels, Buns, Wheat Bread, White Bread, Tortillas, English Muffins, Rolls
	Baking & Cooking	Distilled Vinegar, Sugar, Barbecue Sauce, Pudding, Custard & Mousse Mix, Baking Chips, Marshmallows, Pasta & Pizza Sauces, Tomato Sauce, Paste & Puree
	Dairy	Cream, Butter, Shredded Cheese, Natural Sliced Cheese, Coffee Creamers, Milk, Cream Cheese
Theme 4 (Prepared Food)	Frozen Foods	Frozen Mixed Vegetables, French Fries, Syrup Carriers-Frozen Breakfast, Frozen Broccoli, Packaged Ice Cream, Nutrition Single Serve Meals, Frozen Chicken, Prepared Entrees-Frozen Breakfast
	Canned	Fruit Cups, Canned Tomatoes, Applesauce, Canned Green Beans, Canned Tuna, Canned Corn, Variety Beans, Prepared Beans
	Deli & Prepared Foods	Turkey-Deli, American Cheese (Deli), Ready-Made Sandwiches & Wraps-Prepared, Pork-Deli, Deli Salad
	Beverages	Still Water, Seltzers & Sparkling Water, Fruit Juice, Sports Drinks, Cola
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings
Theme 5 (Unprepared Food)	Produce	Fresh Cucumber, Salad Greens, Apples, Onions, Tomatoes, Bananas, Grapes, Avocado, Lettuce
	Shelf Stable Meals	Ramen & Noodle Soups, Pasta Dishes, Canned Soups, Rice Dishes, Mac & Cheese
	Meat	Chicken, Lunch Packs, Pork-Mainstream, Pork Bacon, Steaks, Ground Beef, Sausage-Pork
	Herbs & Spices	Chili Seasoning, Onion Powder, Grill Seasoning, Cinnamon (Powder & Sticks), Mexican Seasoning, Garlic Powder & Garlic Salt
	Condiments	Salsa, Nut Butters, Jam, Jelly & Marmalades, Hummus, Mayonnaise, Salad Dressings

Note: (1) This table displays product names without brand information, and duplicate names across different brands have been removed; (2) The top products are those that appear most frequently in online baskets.

Appendix F: Statistical Significance of Model Comparisons

In this paper, we evaluate model fit using the Akaike Information Criterion (AIC) when testing Hypotheses 1 and 2. Alternatively, we can apply the Likelihood Ratio Test (LRT), which, under standard regularity conditions, yields a test statistic that follows a chi-square distribution under the null hypothesis (Burnham and Anderson 2004). Suppose we have two models:

- Model $\mathcal{M}_0$: a causal network model (model e in Sections 5.1)
- Model $\mathcal{M}_1$: the network structure with the lowest AIC score among the alternative structures or causal representation specifications.

$\mathcal{M}_0$ constrains the coefficients of non-causal relationships to zero. Let $\log L_0$ and $\log L_1$ denote the maximum log-likelihoods under models $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. The likelihood ratio test statistic is defined as: $\Lambda = -2\left(\log L_0 - \log L_1\right)$. The connection to the AIC difference is $\Delta\mathrm{AIC} = \mathrm{AIC}_0 - \mathrm{AIC}_1 = \Lambda - 2\Delta k$ or equivalently $\Lambda = \Delta\mathrm{AIC} + 2\Delta k$. Assuming $\mathcal{M}_0$ is the correct model for describing customer basket-shopping behavior, and under regularity conditions, the LRT statistic Λ approximately follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters: $\Lambda \sim \chi^2_{\Delta k}$ where $\Delta k = k_1 - k_0$.

This allows us to compute a p-value for testing whether the restrictions imposed by the causal model are consistent with the data: $p = \Pr\left(\chi^2_{\Delta k} > \Lambda\right)$. Based on the above analysis, we compute the LRT statistic between the causal product network and alternative network structures for both Hypotheses 1 and 2.

Table 11 LRT Statistic Comparison (Hypothesis 1)

Theme	Best Benchmark	AIC (Best)	AIC (Model e)	ΔAIC	Δk	LRT Statistic
Theme 1	b2	559,393	557,721	1,672	261	2,194***
Theme 2	b1	437,961	436,085	1,876	460	2,796***
Theme 3	b2	308,765	304,645	4,120	217	4,554***
Theme 4	b2	646,568	638,588	7,980	188	8,356***
Theme 5	d	857,323	822,830	34,493	-118	34,729***

Note: (1) *p<0.1; **p<0.05; ***p<0.01; (2) Models a1 and a2 represent the Lasso-regularized complete network and the Ridge-regularized complete network, respectively; models b1 and b2 represent the SVD-decomposed correlation network and the EVD-decomposed correlation network; model c represents the collaborative filtering-based network; model d represents the association network; and model e represents the causal product network; (3) The best benchmark model refers to the model with the lowest AIC score for each theme; (4) The LRT statistic Λ approximately follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters.

Table 11 presents the likelihood ratio test results comparing the causal product network (model e) with the best benchmark model for each theme. Across all five themes, the causal product network consistently outperforms the best benchmark model in terms of both model fit and statistical significance. In Theme 1 (Pasta), the causal model improves upon the best benchmark (model b2) by reducing the AIC by 1,672, and although it uses 261 fewer parameters, the resulting LRT statistic is 2,194. It indicates strong evidence of improved fit. In Theme 2 (QSR), the causal model achieves an AIC improvement of 1,876 over model b1 and uses 460 fewer parameters, resulting in an LRT statistic of 2,796. The difference is even more pronounced in Theme 3 (Bakery), where the causal model improves AIC by 4,120 with 217 fewer parameters, yielding an LRT statistic of 4,554. In Theme 4 (Prepared Food), the causal model lowers the AIC by 7,980 compared to model b2 and uses 188 fewer parameters, resulting in an LRT statistic of 8,356. Finally, in Theme 5 (Unprepared Food), the causal model achieves the largest improvement, reducing the AIC by 34,493 relative to model d, despite using 118 more parameters; the resulting LRT statistic of 34,729 still indicates overwhelming support for the causal model. In all cases, the improvements are statistically significant at the 1% level, confirming the superior performance of the causal product network.

Table 12 LRT Statistic Comparison (Hypotheses 2)

Theme	Best Benchmark	AIC (Best)	AIC (Model e)	ΔAIC	Δk	LRT Statistic
Theme 1	h	561,288	557,721	3,567	92	3,751***
Theme 2	h	440,309	436,085	4,224	98	4,420***
Theme 3	h	309,806	304,645	5,161	81	5,323***
Theme 4	h	641,099	638,588	2,511	62	2,635***
Theme 5	h	881,728	822,830	58,898	92	59,082***

Note: (1) *p<0.1; **p<0.05; ***p<0.01; (2) Model e represents the causal product network; models f, g, and h represent the category-level causal effects model, the product-to-category causal effects model, and the category-to-product causal effects model, respectively; (3) The best benchmark model refers to the model with the lowest AIC score for each theme; (4) The LRT statistic Λ approximately follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters.

As shown in Table 12, the causal product network outperform other best alternative specification in terms of both model fit (AIC) and statistical validation (LRT) across all five themes. In Theme 1 (Pasta), it lowers the AIC by 3,567 relative to model h, with 92 additional parameters, resulting in an LRT statistic of 3,751. Similarly, for Theme 2 (QSR), the causal model achieves an AIC reduction of 4,224 and includes 98 more parameters, yielding an LRT value of 4,420. The improvement is more substantial in Theme 3 (Bakery), with a 5,161 drop in AIC and 81 extra parameters, leading to an LRT statistic of 5,323. In Theme 4 (Prepared Food), the causal model surpasses model h by decreasing the AIC by 2,511 while adding 62 parameters, resulting in an LRT score of 2,635. Theme 5 (Unprepared Food) shows the most dramatic result: a reduction of 58,898 in AIC with 92 additional parameters, producing an LRT statistic of 59,082. In every case, the observed improvements are highly statistically significant at the 1% level, reinforcing the causal product network's effectiveness over other causal specifications.

Appendix G: Supplement to Assortment Optimization

In this appendix, we present algorithmic details for the assortment optimization results presented in Section 7 of the main body of the paper.

G.1. Mixed-Integer Program for Assortment Optimization Using CPN

Our mixed-integer programming model aims to select the optimal product assortment by maximizing total demand, explicitly accounting for individual product demand as well as causal relationships captured within the discovered causal product network (CPN). Although we currently use total demand as the objective due to the absence of price and cost information in our dataset, our formulation can easily incorporate these factors when available.

We define binary decision variables y_{ij}, where $y_{ij} = 1$ if product i in category j is included in the assortment, and 0 otherwise. Let d_{ij} denote the demand for product i from category j. Additionally, we introduce auxiliary binary variables z_{ij} and a sufficiently large positive constant M to ensure the non-negativity of demand rates. Let K represent the maximum number of products allowed in the assortment. Given a causal product network $\mathcal{G} = (V, E)$ with edge coefficients $\beta^{ij}_{i'j'}$ corresponding to causal relationships $(X_{i'j'} \to X_{ij}) \in E$, we define the assortment optimization problem as CPN-MIP$(\mathcal{G}, \boldsymbol{\beta})$, formulated as follows:

$$
\begin{aligned}
\text{CPN-MIP}(\mathcal{G}, \boldsymbol{\alpha}, \boldsymbol{\beta}): \quad \underset{\mathbf{y},\mathbf{d},\mathbf{z}}{\text{maximize}} \quad & \sum_{i \in I_j, j \in J} d_{ij} && \text{(10a)}\\
\text{subject to} \quad & d_{ij} \le a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'} + M \cdot (1 - z_{ij}), \quad \forall i \in I_j, j \in J, && \text{(10b)}\\
& a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'} \le M \cdot z_{ij}, \quad \forall i \in I_j, j \in J, && \text{(10c)}\\
& a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'} \ge -M \cdot (1 - z_{ij}), \quad \forall i \in I_j, j \in J, && \text{(10d)}\\
& y_{ij} \le z_{ij}, \quad \forall i \in I_j, j \in J, && \text{(10e)}\\
& d_{ij} \le M \cdot y_{ij}, \quad \forall i \in I_j, j \in J, && \text{(10f)}\\
& \sum_{i \in I_j, j \in J} y_{ij} \le K, && \text{(10g)}\\
& y_{ij} \in \{0, 1\}, \quad \forall i \in I_j, j \in J, && \text{(10h)}\\
& z_{ij} \in \{0, 1\}, \quad \forall i \in I_j, j \in J. && \text{(10i)}
\end{aligned}
$$

The objective function maximizes the total demand across all the products in the assortment. Constraint (10b) ensures that the demand d_{ij} for product i from category j does not exceed the sum of its base demand rate, i.e., $a_{ij} \cdot y_{ij}$ and the spillover demand from its parent products, i.e., $\sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'}$. Notice that $a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'}$ can be negative if $\beta^{ij}_{i'j'} \le 0$ for some $(i', j') \in pa_{ij}$. The term $M \cdot (1 - z_{ij})$ in constraint (10b) ensures that the demand rate d_{ij} remains non-negative, and enables us to capture substitution and complementarity effects in the same model. Constraint (10c) ensures that z_{ij} is 1 when $a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'}$ is positive and constraint (10d) ensures that z_{ij} is 0 when $a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'}$ is negative. Constraint (10e) ensures that if $z_{ij} = 0$, which occurs when $a_{ij} \cdot y_{ij} + \sum_{(i',j') \in pa_{ij}} \beta^{ij}_{i'j'} \cdot d_{i'j'}$ is negative, then product i from category j is not included in the assortment, i.e., $y_{ij} = 0$. Constraint (10f) ensures that if product i in category j is not selected to be in the assortment, i.e., $y_{ij} = 0$, the demand rate d_{ij} cannot be positive. Lastly, constraint (10g) ensures the total number of products selected in the assortment does not exceed K.

G.2. Greedy Algorithm for the MNL Model

The Multinomial Logit (MNL) model describes the probability that a shopper selects a particular alternative from a discrete set. Let I_j denote the set of items currently offered in category j. Then, under the Multinomial Logit (MNL) model, the probability that a consumer selects item $i \in I_j$ from category j is given by:

$$\Pr\{i \mid I_j\} = \frac{q_{ij}}{q_{0j} + \sum_{l \in I_j} q_{lj}},$$

where q_{ij} represents the latent attractiveness (utility weight) of item i, and q_{0j} is the utility associated with the no-purchase option for category j. These parameters (q_{ij} and q_{0j}) are typically inferred from historical baskets, choice experiments, or conjoint studies. In our implementation, we estimate q_{ij} for each product based on the utility function of the MNL model, separately for each category. Specifically, q_{ij} is the average purchase quantity of product i in category j across all shopping baskets. Given a total of T baskets, we compute $q_{ij} = \frac{1}{T}\sum_{t=1}^{T} X_{ijt}$, where X_{ijt} is defined as the purchase quantity of product i in category j in basket t. Similarly, $q_{0j} = \frac{1}{T}\sum_{t=1}^{T} X_{0jt}$, where X_{0jt} indicates whether no product in category j is purchased in basket t.

van Ryzin and Mahajan (1999) prove that, under the MNL with a no-purchase option, the optimal K-item subset within a single category is always the first K items when products are ordered by attractiveness ($q_{1j} \geq \cdots \geq q_{|I_j|j}$). The proof uses a majorization argument showing that the marginal increase in purchase probability is quasi-convex and monotone in the "next" highest q value. Thus, a greedy selection of the top K items is globally optimal for a single category.

When extending to multiple categories, directly comparing raw attractiveness values (q_{ij}) across categories becomes problematic. The incremental lift of adding another item to a category diminishes as more items are added due to the denominator in the choice probability formula. Therefore, we use a greedy strategy that selects the category with the largest incremental utility gain, defined as

$$IU_j(k_j+1) = \frac{\sum_{i=1}^{k_j+1} q_{ij}}{\sum_{i=1}^{k_j+1} q_{ij} + q_{0j}} - \frac{\sum_{i=1}^{k_j} q_{ij}}{\sum_{i=1}^{k_j} q_{ij} + q_{0j}}.$$

This incremental utility heuristic is optimal within each individual category but represents a practical greedy heuristic extension for allocation across multiple categories and may not guarantee global optimality.

The above considerations lead to the following greedy algorithm for assortment construction:

Algorithm F1: Greedy Algorithm for MNL Model

Input: $q_{ij}, i \in I_j, j \in J$ where $q_{1j} \geq \cdots \geq q_{|I_j|j}$.
Output: $Y = \{y_{ij} \mid i \in I_j, j \in J\}$.
Initialization: $y_{ij} = 0, \forall i \in I_j, j \in J$, $IU_j = 0$ and $k_j = 0, \forall j \in J$.
1. Initialize $U = 0$.
2. **While** $S < K$:
3. **for** each j in $\{1, ..., m\}$:
4. Compute incremental utility $IU_j = \frac{\sum_{w=0}^{k_j+1} q_{wj}}{\sum_{w=0}^{k_j+1} q_{wj}+q_{0j}} - \frac{\sum_{w=0}^{k_j} q_{wj}}{\sum_{w=0}^{k_j} q_{wj}+q_{0j}}$.
5. Select $\max\{IU_1, \ldots, IU_m\}$ and the corresponding index j'.
6. Update $k_{j'} = k_{j'} + 1$, $y_{k_{j'}j'}$=1 and $S = \sum_{j=1}^{m} \sum_{i \in I_j} y_{ij}$.
7. Return Y.

Algorithm F1 provides a greedy heuristic to construct an assortment by incrementally selecting products from multiple categories based on the Multinomial Logit (MNL) choice model. Initially, all selection indicators $y_{ij} = 0$, meaning no items are selected. For each category j, a counter k_j tracks the number of items selected from that category, initialized to zero. At each iteration, the algorithm computes the incremental utility

(IU_j) of adding the next unselected item to each category. The category j' providing the highest incremental utility is identified, and the next item from this category is selected by setting $y_{k_{j'}j'} = 1$ and incrementing $k_{j'}$. The total number of selected items, S, is updated by summing all current selections: $S = \sum_{j=1}^{m} \sum_{i \in I_j} y_{ij}$. The process repeats until exactly K items are selected. The algorithm returns selection indicators y_{ij}, where $y_{ij} = 1$ indicates that item i from category j is included in the assortment, and $y_{ij} = 0$ otherwise.