Correlation-based Analytics in Big Data

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Nooshin Hamidian

May 2020

To Milad, mom, dad, and Nasim.

Acknowledgments

I would first like to express my special appreciation and thanks to my advisor, Dr. Rapinder Sawhney, for the precious opportunity to be part of his research group, and for his never-ending help and support, and invaluable guidance.

My sincere thanks also goes to Dr. Wenjun Zhou, for all the helps, insights, and support without which this dissertation would not be possible.

To my dissertation committee members, Dr. Anahita Khojandi and Dr. Oleg Shylo. Thank you for your generously offered time and your insights on this work. Your inputs in the dissertation are highly appreciated.

I also would like to thank my fellow graduate students. Thank them for sharing the graduate school life together here and the supportive team environment.

Abstract

In this study, we aim to investigate the application of correlation-based analytics in three main areas, including marketing, healthcare, and manufacturing through three separate, but inherently related studies, and utilize data analytics and optimization methods to develop novel solutions or improve the currently available methods.

First, we demonstrate the significance of items correlation in a product bundling problem. We address a product bundling problem from a data-driven perspective and model bundling utility from a retailer with a wide range of items perspective (such as Walmart, Kroger, Amazon, and so on so forth.) Utilizing the utility model, we mathematically show under which conditions bundling is not profitable in the form of a few theories and lemmas. Applying the mathematically proven theories and lemmas, we develop a pruning algorithm to solve problems with a wide range of products. We use the Dunnhumby dataset to illustrate the effectiveness and efficiency of our algorithm.

Second, we study the correlation between drugs and Adverse Drug Reaction (ADR) to detect groups of patients who are at risk of ADRs. We focus on the application of detecting adverse drug reactions that may not be visible at the global level, and we need to identify local segments in which the association between drugs and adverse reactions are strong. To effectively measure the association between drugs and ADRs per patient group in data-intensive applications, we develop a pruning algorithm. We apply our algorithm on two datasets, including FDA Adverse Events Reporting System and Vaccine Adverse Events Reporting System.

Lastly, we illustrate the importance of considering product correlation in flexibility and production planning. We estimate product demand as a function of its price and supply concerning product correlation. Then, we model the problem of flexibility and

capacity planning for four flexibility strategies. Using the developed models, we investigate conditions under which each flexibility strategy is efficient and develop a two-stage stochastic programming model to optimally detect a flexibility strategy, capacity of facilities, and production quantities. We calibrate the model and conduct sensitivity analysis to show the impact of correlation, price-sensitivity, price change, and demand variation on production revenue.

Table of Contents

1 Introduction **1**

2 Utility-Based Product Bundling:
A Price Sensitivity Perspective **4**

2.1 Introduction 5
2.2 Related Work 7
2.2.1 Itemset Mining 7
2.2.2 Mixed Bundling 8
2.3 Preliminaries and Problem Statement 10
2.3.1 Price-Demand Trade-Off: A Logit Formulation 10
2.3.2 Problem Statement 11
2.3.3 Scope of Study 11
2.4 Bundling Utility 12
2.4.1 Customer Segmentation by Historical Data 12
2.4.2 Response to Bundling by Segment 13
2.4.3 Overall Bundling Utility 15
2.5 Searching for Profitable Bundles 17
2.5.1 Removing Non-Profitable Pairs 17
2.5.2 Screening Candidate Pairs 18
2.6 Algorithm Design 19
2.6.1 The Brute-force Algorithm 20
2.6.2 The PBQ Algorithm 20
2.7 Experimental Results 21

2.7.1 Data Description . . . 21
2.7.2 A Proof-of-Concept Analysis . . . 21
2.7.3 Comparison to a Benchmark Method . . . 23
2.7.4 Efficiency . . . 24
2.7.5 Sample Results . . . 25
2.8 A BOGO Case Study . . . 28
2.9 Conclusions . . . 31

3 Simultaneous Detection of Adverse Drug Reaction and At-Risk Groups 33
3.1 Introduction . . . 33
3.2 Related Work . . . 36
3.3 Preliminaries and Problem Statement . . . 38
3.3.1 Problem Formulation . . . 38
3.3.2 Binary Data Association Measures . . . 39
3.3.3 Association Measures with Segmentation . . . 40
3.4 Problem Analysis . . . 41
3.4.1 Combination of Binary Events . . . 41
3.4.2 Initial Bounds . . . 42
3.4.3 Secondary Bounds . . . 44
3.5 Algorithm Description . . . 46
3.5.1 Subgroup Generation . . . 46
3.5.2 Brute Force Algorithms . . . 48
3.5.3 Two-level Pruning Algorithm . . . 49
3.5.4 Save and Look up Pruning Algorithm . . . 49
3.6 Experiments . . . 57
3.6.1 Data Description . . . 57
3.6.2 Effectiveness . . . 59
3.6.3 Efficiency . . . 61
3.7 Conclusion . . . 61

4 An Empirical Analysis of Capacity and Flexibility Planning Under Demand Uncertainty **64**

4.1 Introduction . . . 65

4.2 Literature Review . . . 66

4.2.1 Volume Flexibility . . . 66

4.2.2 Product Flexibility . . . 67

4.2.3 Multiple Flexibility . . . 67

4.2.4 Summary . . . 67

4.3 Model . . . 68

4.3.1 Preliminary . . . 69

4.3.2 Volume and Product Flexibility Model . . . 70

4.3.3 Volume Flexibility Model . . . 72

4.3.4 Product Flexibility Model . . . 73

4.3.5 Without Flexibility Model . . . 74

4.4 Empirical Investigation . . . 75

4.4.1 Solution . . . 76

4.4.2 Experimental set-up . . . 76

4.4.3 Scenario Generation . . . 77

4.4.4 The impact of correlation . . . 77

4.4.5 The impact of variation . . . 79

4.4.6 The impact of price-sensitivity . . . 81

4.4.7 The impact of price . . . 81

4.4.8 Managerial Implications . . . 84

4.4.9 Discussion and Future Work . . . 87

5 Conclusions **88**

Bibliography **90**

Appendices **98**

A Summary of Equations . . . 99

A.1 Proof of Theorem 2.2 99
A.2 Proof of Theorem 2.3 100
A.3 Proof of Theorem 2.4 102
A.4 Proof of Theorem 2.6 103
A.5 Proof of Theorem 2.7 103
B Search of Profitable Bundles 104

Vita **106**

List of Tables

2.1 Segmentation of n historical transactions regarding the (co-)purchase outcome of two items when neither is on promotion. . . . 13
2.2 Segmentation of N transactions regarding the (co-)purchase outcome of two items without promotion. . . . 13
2.3 POC Method . . . 22
2.4 Comparing Effectiveness . . . 24
2.5 Bundles Selected by PBQ . . . 26
2.6 Item Specification . . . 26
2.7 Comparing Selected Bundles of Two Store Sets . . . 27
2.8 A sample results of BOGO offer with three levels of promotion . . . 31

3.1 The Contingency Table. . . . 39
3.2 The Stratified Contingency Table . . . 40
3.3 The most Frequent Subgroups of Patients in VAERS . . . 60
3.4 The most Frequent Subgroups of Patients in FAERS . . . 60
3.5 Examples of at-risk patient groups . . . 60

4.1 Notations . . . 71
4.2 Parameters' range in numerical examples . . . 76
4.3 Optimality gap for four cases of flexibility . . . 77
4.4 Three level parameters range . . . 84
4.5 Ranking flexibility strategies for different correlation, variation, and price conditions . . . 85
4.6 Impact of Cost on Flexibility Selection . . . 86

List of Figures

2.1 Correlation vs. Utility. 17
2.2 Comparison of training utility and testing utility 23
2.3 Comparison of Running Time 25
2.4 Comparison of Customer Demographics. 28
2.5 The distribution of order quantities 29

3.1 Correlation vs. Utility. 58
3.2 Comparison use of space: x-axis: Theta ; y-axis: the number triplet combinations of drug, ADR, and patients group 62
3.3 Running time comparisons: x-axis: the number triplet combinations of drug, ADR, and patients group; y-axis: Running time 63

4.1 Impact of correlation on profit 78
4.2 Impact of demand variation on profit 80
4.3 Impact of price-sensitivity on profit 82
4.4 Impact of price-change on profit 83

Chapter 1

Introduction

In this dissertation, we address correlation-based problems in three different domains, including marketing, healthcare, and production through the utilization of analytics and optimization techniques in large-scale datasets. Enormous datasets have been collected from many domains such as retailing, healthcare, and manufacturing industries which provides a great opportunity to solve the most challenging problems in these areas with a data-driven approach. Measuring correlation plays a fundamental role in the effectiveness of proposed methodologies. However, effectively measure the association between elements makes challenging big data problems more perplexing. Accordingly, efficient methodologies are required to discover significant associations in big data.

The critical role of correlation in decision making has been illustrated in a large number of previous studies. However, due to the complexity of correlation measurement, the majority of previous research in various domains simplified the problem regarding the correlation. Danaher et al. [16] employed simplified correlation structures to model a mixed bundling problem in the music industry. Chao and Derdenger [12] studied mixed bundling problem in a two-sided market, assuming that the platform and the integrated content were independent. Additionally, most papers in itemset mining literature focused on discovering co-purchasing patterns without accounting for items correlation which may cause spurious patterns [65, 54, 47]. Furthermore, after the availability of post-marketing surveillance ADR records, a number of previous studies explore the correlation between drugs and ADRs. Most studies in literature limited the number of ADRs and drugs and

focused more on the accuracy of determining the significant association between drugs and ADRs [36, 17, 2]. However, by increasing the number of existing drugs, the number of required correlations increases exponentially. Moreover, correlation plays an essential role in demand forecasting. Considering increasing demand uncertainty, accounting for correlations in demand forecasting has a significant impact on the reduction of misalignment between assigned capacity and demand. Since the complexity of correlation measurement increase by increasing the number of products, many previous studies limited their flexibility and production planning problems into a two-product setting [13, 25, 8]. Therefore, novel methods are required to find the association measurements effectively in all three domains of marketing, healthcare, and manufacturing.

In this dissertation, we develop innovative solutions to detect correlation measures in three separate but relates studies. In the first study, we address a bundle discovery problem in large-scale datasets. By considering price-sensitivity and correlation, we develop a probability model to estimate the utility of bundling, which is unprecedented in multiple product settings. Furthermore, we establish mathematical theorems to prune unnecessary calculations and propose a pruning algorithm based on these theorems. We utilize Dunnhumby dataset to present the efficiency and effectiveness of our model by comparing it with the brute force algorithm and benchmark method, respectively. In the second study, we detect significant adverse drug reactions associated with drugs accounting for patients segment (including demographic and medical history features.) We efficiently discover segments of patients who are at risk of ADR occurrence by a drug exposure while the drug is not generally associated with the specific ADR. To this end, we computationally reduce the size of the problem and develop pruning algorithms for three correlation measurements. To the best of our knowledge, it is the first study exploring this problem. Moreover, selecting a useful correlation measure for an application strongly depends on several application properties Tan et al. [51]. Accordingly, we extend our method to three correlation measurements considering their specific conditions. In the third study, we develop a framework for combined capacity, production, and flexibility planning in a multiproduct setting, exploring product correlation which is unprecedented in literature. We model various

flexibility strategies with two-stage stochastic programming and solve the problem with an L-shaped algorithm.

Chapter 2

Utility-Based Product Bundling: A Price Sensitivity Perspective

While bundling has been widely adopted as a promotion strategy in many industries, determining which items to bundle has been an infrequent, coarse (e.g., at the brand or category level), and/or non-automatic practice that heavily relies on domain knowledge and business acumen. Retailers equipped with scanner data from their stores have opportunities to leverage the co-purchase correlation to develop customized, novel product bundles in a real-time fashion, while this task is non-trivial given the large number of possible items to bundle, especially at the fine-grained level. Many existing methods that search for co-purchase patterns reveal spurious associations. In particular, if we identify correlated item pairs solely based upon observed co-purchase behaviors, we ignore confounding factors such as promotions. In this paper, we consider a more realistic scenario: the price changes, and when there is a discount, consumers willingness to buy changes accordingly. In this context, we have developed a utility function for assessing the profitability of product bundles and propose an efficient algorithm to search among all possible item pairs. This utility function is able to predict the most profitable bundles and expose the predicted behaviors of customers based on historical data. To improve the efficiency and reduce the number of itemsets, a pruning algorithm has also been developed. Several experiments are conducted to show the performance of the algorithm. A real case study is then conducted to evaluate the proposed approach.

2.1 Introduction

Selling two or more items as a promitional package with a promotion is known as bundling. This strategy has been widely applied in many industries, such as tourism [49], retailing [38], telecommunications [63], technology [12], and information goods [7]. Several advantages of bundling have been determined in previous studies including cost reduction in production and transaction, complementaries among bundle components, increasing profits, and price discrimination [1, 3, 23]. However, which items to bundle has typically relied on trial and error through traditional campaigns and gauging customers responses.

Availability of microlevel data provides an opportunity to study customers' behavior directly and allows for a more targeted, personalized marketing response. and take personalized actions in response [16]. Microlevel data from traditional or digital channels, such as point-of-sales scanner data and online shopping transactions, have been used to enhance bundling strategies, which provides adaptive, automatic, and personalized bundling suggestions to customers in real time. For instance, recommender systems are now prevalent and a critical component of e-commerce because of their ability to detect and recommend personalized bundles to individual shoppers [5]. During recent years, the availability of extensive datasets has provided opportunities for developing data-driven methods considering actual customers' behavior in a specific store, area, or time period.

Data-driven product bundling usually relies on identifying products that are co-purchased. Co-purchase correlation has been a challenging computational problem because the problem complexity can become exponentially large. Traditional correlation computing does not consider the monetary value of product bundles. It is possible that a likely co-purchase is not the most profitable bundle, which leads marketers to pursue those high-utility patterns [54, 47, 39]. Studies in determing high utility itemsets primarily focused on identifying itemsets based on known, static utility as input data, and did not consider price sensitivity, making it hard to discern new promotion effects for optimizing future possible promotions. Therefore, these studies in utility mining have rarely explored price-sensitivity or promotions effects from historical data. In contrast, our study aims to develop a data-driven methodology for identifying the most profit-making product bundles while

consideration market response. We argue that to identify the utility or profitability of a bundle, we need to understand the "what-if" scenario of when the bundle is actually being offered. This will allow us to more accurately assess the bundling effect. Our setting is in a retail store that offers numerous products simultaneously. We focus on item pairs because it is prevalent technique and analytically less complex. Even when limiting the scope of our research to item pairs, this problem remains quadratic, which is prohibitive for large a number of items.

We propose a model to estimate the utility considering discount rates and items' price-sensitivity. We explore some analytical results corresponding to the developed formulation. Although we pre-compute item-level parameters using historical transactional datasets (i.e., base demand and price-sensitivity), the developed pairwise utility depends on the pair, their parameters, and their correlation. Additionally, we combine our model with an empirical dataset. To handle the quadratic number of pairs efficiently, we propose a pruning algorithm. Pruning conditions are developed in subsection 2.5.2 and subsection 2.5.1. subsection 2.5.2 investigates conditions under which bundling is not profitable. subsection 2.5.2 develops an upper bound condition to avoid the exact correlation computation as much as possible.

Hence, we only estimate exact correlations for pairs with a correlation upper bound $\overline{\phi}_{AC}$ greater than $\phi^*(\delta)$ in which δ is a user-specific threshold determining the minimum expected utility of a beneficial item pair.

Our paper makes several contributions to the existing literature. First, we utilize price-sensitivity and probability modeling to estimate the utility of each pair after discount, which is unprecedented in literature.

Second, the proposed utility model is a function of items correlation. Third, we determine conditions under which bundling is not profitable Forth, we propose a pruning algorithm to solve large problems more efficiently using the profitability conditions previously defined. Finally, we combine our model with an empirical dataset. There are no limitations associated with the application of our model. It is applicable to all transactional data sets including online shopping datasets.

The rest of this paper is organized as follows: section 2.2 overviews the relevant literature. section 2.3 delineates the problem. section 2.4 provides the problem formulation. section 2.5

provides some critical theoretical results. section 2.6 describes two algorithms, including a straightforward algorithm, and a designed algorithm. section 2.7 presents the experimental results. section 2.8 demonstrates a special case study of Buy One Get One with discount. section 2.9 concludes the study and discusses future works.

2.2 Related Work

There are two streams of literature related to our study. One main stream of literature investigates frequent itemsets and utility mining which has been extensively studied in the literature and we review the most related studies in subsection 2.2.1. The other relevant literature stream studies mixed bundling and those related studies are summarized the most related papers in subsection 2.2.2.

2.2.1 Itemset Mining

To discover shopping patterns and customers behavior, numerous studies focused on associate rule mining and utility mining techniques. Many data-driven studies concentrated on detecting the most frequent itemsets or itemsets with the highest utility, while many others focused on recommender systems to offer single or multiple items to customers specifically.

[64] and Yang et al. [65] applied the association rule mining technique to discover online shopping patterns. Hwang and Yang [28] proposed an approach based on association rule mining for new products without any transaction records. These studies rarely considered price-sensitivity in itemsets mining. Tseng et al. [54] developed three algorithms to provide a compact and lossless representation of high utility itemsets. The primary purpose of their research was to reduce the number of high utility itemsets. Mai et al. [39] proposed an algorithm to find high-utility itemsets more efficiently. In utility mining literature, some studies detected patterns of items by considering the utility of each item before utility mining implementation. [34] developed a utility-based association rule mining for a cross-selling problem. Sahoo et al. [47] discovered association rules with the aid of the utility-confidence framework. Whereas utility association mining is a promising procedure to address the itemsets' utility, in many cases it results in a large number of rules. This problem can affect

the efficiency and applicability of the algorithm. The last high utility mining studies mainly concentrated on two areas: efficient reduction of the number of itemsets and applications of utility mining. The key in any utility mining study is knowing the utility of each item as input of the utility mining algorithm; however, in our study, we needed to estimate or learn the utility of bundling for each item pair.

Recommender systems was another area of data-driven analytics have been studied extensively, but bundle recommendations or multiple items recommendations have recently gained more attention. Beladev et al. [5] introduced a model for bundle recommendations. They used the collaborative filtering technique in which correlation between customers was calculated while correlation between items was not investigated. The complexity of the problem and the absence of pruning methods, meant their proposed algorithm was not applicable in large-scale problems. Ghoshal and Sarkar [22] applied rule mining techniques for multiple items recommendations. In order to make rules tractable, they developed some pruning techniques. Ghoshal et al. [21] proposed a method based on association rules to improve the quality of recommendations. They translated the rule mining problem to an information theoretic context, and selected those rules which led to the highest mutual information. Wan et al. [59] considered three behavioral patterns, including complementarity, compatibility, and loyalty when recommending products for grocery shopping. Recommendations studies rarely considered correlation and price-sensitivity. Additionally, by applying the multi items recommendation systems, they considered the bundling solution in a personalized context as opposed to our study, which ... .

2.2.2 Mixed Bundling

A majority of research on bundling focused on optimal pricing. Bhargava [7] simplified the mixed bundle pricing problem as a univariate nonlinear optimization problem and developed a near-exact closed-form solution. Bulut et al. [10] studied bundle pricing of two perishable products with limited inventory to maximize the expected revenue of bundling. There were at least two limitations in these types of studies. First, items correlation was not investigated, and the number of items was limited to two given items. Second, they did not use real-world

datasets so many assumptions might not be sufficiently realistic. Little research in this area investigated the impact of bundling on customer behavior [23].

Empirical or real-world data-driven studies have also investigated the mixed bundling problem. Danaher et al. [16] conducted a study in the music industry using real market data. They estimated the distribution of items' value and their price elasticities, and combined their developed structure with the market level data to detect optimal price levels for single songs and albums. However, due to data limitations, they only investigated customers' behavior with two price points, which assumed those points were representative of all price ranges. Additionally, they employed simplified correlation structures to estimate the model because they could not estimate the true correlation. Chao and Derdenger [12] studied mixed bundling problem in two sided markets. They assumed that the platform and the integrated content were independent. They showed mixed bundling as a price discrimination tool caused better coordination between two sides of the market. To empirically present their results, they used the portable video game console market. Benisch and Sandholm [6] developed a framework to detect high profit bundle discounts using customer purchase data. They applied different pricing algorithms and a customer validation model considering normally distributed valuations on each item. They only considered item pairs which were not directly or indirectly related to any other items. [62] studied the problem of pricing personalized bundles based on customer requests for quotes in the computer hardware industry. Their problem was very different in the sense that each quote consisted of any number of items in a complex product category hierarchy, and the main objective being reasonably priced products. They built a utility model for each personalized bundle by considering bundle features and customers attributes. Ferreira et al. [18] conducted an empirical analytics for an online retailer. They developed an algorithm to solve the price optimization problem for Rue La La, an online fashion retailer. Although they did not consider bundling at all, their study is related to ours because they addressed the predicting demand challenges on new items not released to customer markets. These previous empirical studies discussed mainly aimed at price optimization. However, in our study, we are attempting to discover profitable bundles among a large array of products.

2.3 Preliminaries and Problem Statement

section 2.3 is devoted to preliminaries, including the demand formulation for which a logit formulation of purchasing probability depending on discount value is presented in subsection 2.3.1. subsection 2.3.2 describes the problem statement, and subsection 2.3.3 determines the scope of this study.

2.3.1 Price-Demand Trade-Off: A Logit Formulation

Microeconomic theories have proposed that product demand changes by price. Specifically, the demand of goods decreases when the price increases, and the extent of change in demand by a unit change in price is known as price elasticity or price-sensitivity. Individual items may vary drastically in price elasticity. We can reasonably assume that the price elasticity of each item may be estimated from the existing data.

Suppose that demand $\mathbb{P}_A^d$, represents the probability of purchasing an item, A, for any customer when there is a discount, d. When there is no discount (i.e., $d = 0$), we have a special case of base demand, $\mathbb{P}_A^0$. In the scope of this paper, we limit the consideration of promotion to price reduction. In practice, the effects of different formats of promotions could be converted into an equivalent price reduction scenario as needed.

Assuming a logit function, for any item, A, we have

$$\log\left(\frac{\mathbb{P}_A^d}{1-\mathbb{P}_A^d}\right) = \log\left(\frac{\mathbb{P}_A^0}{1-\mathbb{P}_A^0}\right) + \beta_A \cdot d + \epsilon \tag{2.1}$$

where β_A, represents the item's price-elasticity. In a rational market, we assume that $d \geq 0$, $\beta_A \geq 0$, and $\mathbb{P}_A^d \geq \mathbb{P}_A^0 \geq 0$.

The left-hand side of Equation 2.1 is the log odds of purchase under discount d, which can be calculated as its price-sensitivity multiplied by the discount amount, offset by its base demand (i.e., log odds of purchase under normal price).

2.3.2 Problem Statement

Since product bundling is a business decision made by sellers, we formulate the bundling utility problem from the sellers perspective. The company should select bundles of products which optimize its objective. The goal in our study is to detect bundles which maximize revenue.

Definition 2.1 (Profitable Bundle Discovery)**.** *Given a minimum gain threshold, δ (a monetary quantity) and a discount rate, s, the goal is to find all profitable two-item bundles in the form of "get s% discount if you buy A and C together."*

When considering two items, A and C, for bundling, we want to determine the change in utility, ΔU_{AC} after offering bundling discount, d. In our problem formulation, this discount is calculated as $d = s\% \times (\pi_A^0 + \pi_C^0)$, where π_A^0 and π_C^0 are normal prices for items. Normal price is the regular price of an item without any discount, and s is discount ratio on both items for co-purchasing.

2.3.3 Scope of Study

Mixed bundling, a special case of bundling, is the practice of selling products individually as well as selling them in bundles, where bundle purchases are incentivized with a promotion. In general, mixed bundling is shown to outperform component selling and pure bundling because it facilitates price discrimination and extracts customer surplus [1, 10, 11, 57]. It has been reported that mixed bundling is more effective than pure bundling (in which products are only sold in bundles) and component selling (in which products are sold individually with no promotion for co-purchasing) [40, 45]. Mixed bundling has been the most common form of bundling, especially in the retail market. For these reasons our study focuses on the mixed bundling scenario.

In this study, we consider a large retailer selling a wide range of products, such as Kroger or Walmart, or online stores such as Amazon. Our goal is to detect profitable pairs of items among all possible combinations. For simplicity, we assume that the bundling mechanism does not impose additional costs to the retailer.

It is also assumed that a customer is either interested in a specific item at a specific quantity and purchases it or is not interested at all in this type of purchase. Although exploring the order quantities is not in the scope of this research, we investigate a Buy One Get One (BOGO) case study in which quantity of products purchased matters.

To illustrate the changes in purchasing probability by offering a promotion, we utilize logit formulation as described in subsection 2.3.1. The use of logit formulation to estimate the odds of purchasing is a function of promotion. Other functions for purchasing probability, which depend on variables other than price or discount, can be eexplored in a future study.

Large retailers use market basket analysis to detect the associations between items frequently purchased. In other words, they look for items that appear in customer baskets together. Accordingly, our retailer analytics study focuses on positively correlated items.

2.4 Bundling Utility

This section includes the notations and computational utility formulations. subsection 2.4.1 describes the segments of customers using a contingency table. subsection 2.4.2 formulates customer responses to a bundling promotion which differs by customer segments and develops a utility function for each identified segment. subsection 2.4.3 shows the overall bundling utility which is a summation of each segments utility.

2.4.1 Customer Segmentation by Historical Data

We segment customers by purchase intention when no discount is offered. We find quantities in the two-way contingency table as shown in Table 2.1 from the historical data. In particular, n_{AC}, is the number of transactions with both A and C, $n_{\bar{A}C}$, is the number of transactions with A but not C, and so on. We use lower case n to indicate that these quantities are fixed numbers found from transactional records collected in the past. These quantities help us project expected segmentation of a future period for which we would like to determine profitable bundles.

When expressed with purchase probabilities (see subsection 2.3.1), we have $\mathbb{P}^0_A = \frac{n_A}{n}$ and $\mathbb{P}^0_C = \frac{n_C}{n}$. Similarly, we have the co-purchase probability $\mathbb{P}^{0,0}_{AC} = \frac{n_{AC}}{n}$.

Table 2.1: Segmentation of n historical transactions regarding the (co-)purchase outcome of two items when neither is on promotion.

	$C=1$	$C=0$	Total
$A=1$	$I : n_{AC}$	$II : n_{A\bar{C}}$	n_A
$A=0$	$III : n_{\bar{A}C}$	$IV : n_{\bar{A}\bar{C}}$	$n_{\bar{A}}$
Total	n_C	$n_{\bar{C}}$	n

Given these quantities, the item pair's co-purchase correlation may be calculated as

$$\phi_{AC} = \frac{\mathbb{P}^{0,0}_{AC} - \mathbb{P}^{0}_{A}\mathbb{P}^{0}_{C}}{\sqrt{\mathbb{P}^{0}_{A}\left[1-\mathbb{P}^{0}_{A}\right]\mathbb{P}^{0}_{C}\left[1-\mathbb{P}^{0}_{C}\right]}}. \tag{2.2}$$

Note that, in Equation 2.2, we implicitly assume that the correlation between two items does not change by offering a discount on the bundle. This assumption is reasonable because the correlation we are interested in is the consistent correlation based on customer needs, *irrespective of promotion*. In fact, this is a strength of our study since it has been reported that spurious correlations are common in traditional pattern mining literature.

2.4.2 Response to Bundling by Segment

Suppose that now we are planning for product bundling for the next period, during which N transactions has been projected (i.e., customer visits). These transactions will be classified into the four quadrants of purchase outcomes (see Table 2.2), similar to the historical data in Table 2.1: There will be N_{AC} transactions with both A and C, $N_{\bar{A}C}$ transactions with A but not C, and so on. We use capital letters to represent random variables, since these are subject to change when discounts are offered.

Table 2.2: Segmentation of N transactions regarding the (co-)purchase outcome of two items without promotion.

	$C=1$	$C=0$	Total
$A=1$	$I : N_{AC}$	$II : N_{A\bar{C}}$	N_A
$A=0$	$III : N_{\bar{A}C}$	$IV : N_{\bar{A}\bar{C}}$	$N_{\bar{A}}$
Total	N_C	$N_{\bar{C}}$	N

As illustrated in Table 2.2, the number of transactions projected for each segment (without promotion) is as follows

$$
\begin{aligned}
N_{AC} &= N \times \mathbb{P}^{0,0}_{AC}, && (2.3)\\
N_{A\bar{C}} &= N \times \left[\mathbb{P}^{0}_{A} - \mathbb{P}^{0,0}_{AC}\right], && (2.4)\\
N_{\bar{A}C} &= N \times \left[\mathbb{P}^{0}_{C} - \mathbb{P}^{0,0}_{AC}\right], \text{ and} && (2.5)\\
N_{\bar{A}\bar{C}} &= N \times \left[1 - \mathbb{P}^{0}_{A} - \mathbb{P}^{0}_{C} + \mathbb{P}^{0,0}_{AC}\right]. && (2.6)
\end{aligned}
$$

After product bundling (with bundling discount $d > 0$), the classification of the N transactions into the four quadrants is subject to change. In the following, we first discuss them by segment and then aggregate them to calculate the bundling utility/sales. Besides, the joint probability of purchasing with x discount on A, and $d-x$ on C ($\mathbb{P}^{x,d-x}_{AC}$) changes as Equation 2.7 in which $\mathbb{P}^{x}_{A}$ is the probability of purchasing item A with x discount, and $\mathbb{P}^{d-x}_{C}$ is the probability of purchasing item C with $d-x$ discount.

$$
\mathbb{P}^{x,d-x}_{AC} = \mathbb{P}^{x}_{A}\mathbb{P}^{d-x}_{C} + \phi_{AC}\sqrt{\mathbb{P}^{x}_{A}(1-\mathbb{P}^{x}_{A})\mathbb{P}^{d-x}_{C}(1-\mathbb{P}^{d-x}_{C})} \tag{2.7}
$$

In Segment I, there are N_{AC} customers interested in both items and will buy both of them even without a promotion. If we offer discount, d, for purchasing both items together, the utility after bundling for Segment I is

$$
U_I(d) = N_{AC} \times \pi^{d}_{AC} \tag{2.8}
$$

In Equation 2.8, π^{d}_{AC} presents the bundle price which is the summation of the included item prices minus offered discount as $\pi^{d}_{AC} = \pi^{0}_{A} + \pi^{0}_{C} - d$.

Segment II contains $N_{A\bar{C}}$ customers interested in A but not C. For customers in this segment, if we offer a discount for the bundle, the customers will perceive that discount is applicable to item C. In other words, they will buy A irregardless of a discount, and if the bundle discount is significant enough, they will not mind that C is part of the purchase. The

probability of purchasing both items, considering a zero discount on A and d discount on C, is $\mathbb{P}_{AC}^{0,d}$. This segment of customers are already intended to buy A for which the associated utility is $N_{A\bar{C}} \cdot \pi_A^0$. There remains a chance that a part of these customers will be interested in purchasing the bundle after discount. The probability of this segment to be interested in the bundle is $(\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0})$, which is the difference between the probability of making a purchase after receiving a discount on C, and the probability of purchasing both items before discount is applied. Accordingly, the utility after bundling for this segment is:

$$U_{II} = N_{A\bar{C}} \cdot [(\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_C^d + \pi_A^0]. \tag{2.9}$$

Segment III includes $N_{\bar{A}C}$ customers who are interested in C but not A. Following a logic similar to Segment II, we express the utility as:

$$U_{III} = N_{\bar{A}C} \cdot [(\mathbb{P}_{AC}^{d,0} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_A^d + \pi_C^0]. \tag{2.10}$$

In this case, the joint probability will be

$$\mathbb{P}_{AC}^{x^*,d-x^*} = \begin{cases} \mathbb{P}_{AC}^{d,0} & \text{if } \beta_A \geq \beta_C; \\ \mathbb{P}_{AC}^{0,d} & \text{if } \beta_A < \beta_C; \end{cases} \tag{2.11}$$

The utility in this segment is

$$U_{IV} = N_{\bar{A}\bar{C}} \cdot [\mathbb{P}_{AC}^{x^*,d-x^*} - \mathbb{P}_{AC}^{0,0}] \cdot \pi_{AC}^d. \tag{2.12}$$

2.4.3 Overall Bundling Utility

The overall bundling utility is summarized in Theorem 2.2.

Theorem 2.2 (Bundling Utility)**.** *Increased utility after bundling items A and C by offering discount amount, d, is*

$$\begin{aligned} \Delta U_{AC} = {} & N_{A\bar{C}} \cdot \pi_C^d \cdot \mathbb{P}_{AC}^{0,d} + N_{\bar{A}C} \cdot \pi_A^d \cdot \mathbb{P}_{AC}^{d,0} + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^d \cdot \mathbb{P}_{AC}^{x^*,d-x^*} \\ & -[N_{\bar{A}} \cdot \pi_A^0 + N_{\bar{C}} \cdot \pi_C^0 + N_{AC} \cdot d] \cdot \mathbb{P}_{AC}^{0,0}. \end{aligned} \tag{2.13}$$

Proof. Proof: The proof of Theorem 2.2 can be found in subsection A.1. □

From Equation 2.13, we can see that there are several terms that involve joint probabilities. To eliminate them, we leverage Equation 2.14, which depends only on the marginals, as well as the correlation coefficient. In this way, we can reduce the ΔU_{AC} into a function that just depends on ϕ, since finding the marginals is considered trivial. The result is outline in Theorem 2.3.

Lemma 2.3. *Bundling utility is a quadratic function of correlation. Specifically, we have*

$$\Delta U_{AC}(\phi) = \gamma_0 + \gamma_1 \phi + \gamma_2 \phi^2, \tag{2.14}$$

where

$$\begin{aligned}
\gamma_0 &= N_A \pi_C^d \mathbb{P}_A^0 \mathbb{P}_C^d + N_C \pi_A^d \mathbb{P}_A^d \mathbb{P}_C^0 + (N - N_A - N_C) \pi_{AC}^d \mathbb{P}_A^x \mathbb{P}_C^{d-x} - (N_{\bar{A}} \pi_A^0 + N_{\bar{C}} \pi_C^0) \mathbb{P}_A^0 \mathbb{P}_C^0 \\
&\quad + N \mathbb{P}_A^0 \mathbb{P}_C^0 [-\pi_C^d \mathbb{P}_A^0 \mathbb{P}_C^d - \pi_A^d \mathbb{P}_A^d \mathbb{P}_C^0 + \pi_{AC}^d \mathbb{P}_A^x \mathbb{P}_C^{d-x} - d \mathbb{P}_A^0 \mathbb{P}_C^0] \\
\gamma_1 &= N \mathbb{S}_A^0 \mathbb{S}_C^0 (-\pi_C^d \mathbb{P}_A^0 \mathbb{P}_C^d - \pi_A^d \mathbb{P}_A^d \mathbb{P}_C^0 + \pi_{AC}^d \mathbb{P}_A^x \mathbb{P}_C^{d-x} - d \mathbb{P}_A^0 \mathbb{P}_C^0) \\
&\quad + N \mathbb{P}_A^0 \mathbb{P}_C^0 (-\pi_C^d \mathbb{S}_A^0 \mathbb{S}_C^d - \pi_A^d \mathbb{S}_A^d \mathbb{S}_C^0 + \pi_{AC}^d \mathbb{S}_A^x \mathbb{S}_C^{d-x} - d \mathbb{S}_A^0 \mathbb{S}_C^0) \qquad (2.15) \\
&\quad + N_A \pi_C^d \mathbb{S}_A^0 \mathbb{S}_C^d + N_C \pi_A^d \mathbb{S}_A^d \mathbb{S}_C^0 + (N - N_A - N_C) \pi_{AC}^d \mathbb{S}_A^x \mathbb{S}_C^{d-x} - (N_{\bar{A}} \pi_A^0 + N_{\bar{C}} \pi_C^0) \mathbb{S}_A^0 \mathbb{S}_C^0 \qquad (2.16) \\
\gamma_2 &= N \mathbb{S}_A^0 \mathbb{S}_C^0 [-\pi_C^d \mathbb{S}_A^0 \mathbb{S}_C^d - \pi_A^d \mathbb{S}_A^d \mathbb{S}_C^0 + \pi_{AC}^d \mathbb{S}_A^x \mathbb{S}_C^{d-x} - d \mathbb{S}_A^0 \mathbb{S}_C^0] \qquad (2.17)
\end{aligned}$$

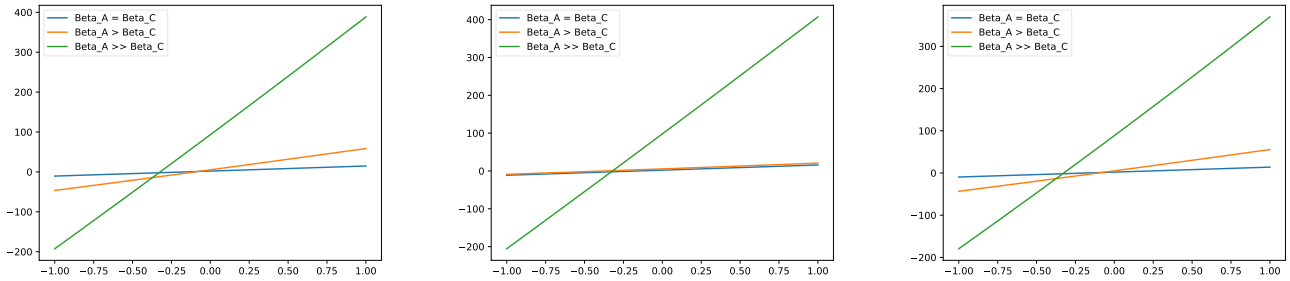
For simplicity of notations, let

$$\mathbb{S}_A^{x^*} = \sqrt{\mathbb{P}_A^{x^*} [1 - \mathbb{P}_A^{x^*}]} \tag{2.18}$$

Proof. Proof: The proof of Theorem 2.3 can be found in subsection A.2. □

Consequently, for any given item pairs, if we know its exact correlation value, we can directly calculate the ΔU using their exact ϕ values.

Figure 2.1 shows the bundling utility which is a quadratic function of correlation. It shows the bundling utility under different conditions of items price and price-sensitivity. We



(a) $Price_A = Price_C$ **(b)** $Price_A > Price_C$ **(c)** $Price_A < Price_C$

Figure 2.1: Correlation vs. Utility.

know ϕ is in the range of $[-1, 1]$, as illustrated in Figure 2.1. The bundling utility increases by increasing ϕ in $[-1, 1]$ under different conditions.

Theorem 2.4 (Monotonicity of ΔU relative to ϕ)**.** *Assuming $\phi \geq 0$, $\Delta U(\phi)$ in Equation 2.14 is an increasing function of ϕ.*

Proof. Proof: The proof of Theorem 2.4 can be found in subsection A.3. □

2.5 Searching for Profitable Bundles

Suppose that base demand and price-sensitivity have been extracted from the historical data in a preprocessing step. To detect profitable bundles (i.e., an item pair that satisfies a user-specified minimum gain threshold when bundled), we need to explore all possible item pairs. This process is computationally prohibitive due to the quadratic complexity of the problem. In this section, we identify possible ways to reduce the search space. Specifically, we identify an upper bound condition, which may help avoid unnecessary pairwise estimations such as joint probabilities.

2.5.1 Removing Non-Profitable Pairs

First, we identify a useful property for excluding a pair from the proposed set because of the theoretical guarantee that they would not generate a profit.

Lemma 2.5 (The Minimum Correlation Requirement)**.** *For a given item pair* $\{A, C\}$*, suppose that we know their respective base demand, price sensitivity, and discount, a correlation of* $\frac{-\gamma_1+\sqrt{\gamma_1^2-4\gamma_2(\gamma_0-\delta)}}{2\gamma_2}$ *is required for any item pair to worth bundling.*

Proof. Proof: According to Theorem 2.4, we can find the minimal correlation by solving the following equation:

$$\Delta U(\phi) = \delta \tag{2.19}$$

where δ is the user-specified minimum utility threshold, from the above we can find

$$\phi^*(\delta) = \frac{-\gamma_1 + \sqrt{\gamma_1^2 - 4\gamma_2(\gamma_0 - \delta)}}{2\gamma_2} \tag{2.20}$$

□

As it is shown in Equation 2.14, ΔU is a quadratic function of ϕ. By solving Equation 2.14 when the $\Delta U(\phi)$ is equal to the minimum gain threshold δ (Equation 2.19), we can find the solution as shown in Equation 2.20. Based on Theorem 2.4, we know ΔU is an increasing function of correlation when $\phi \geq 0$. Hence, $\phi^*(\delta)$ is the minimum correlation coefficient which can satisfy the minimum utility gain δ. In other words, in order for an item pair to increase utility after bundling with δ, the correlation between items should be equal or greater than $\phi^*(\delta)$.

Corollary 2.6 (Price Insensitive Pair)**.** *Bundling two price insensitive items is not profitable.*

Proof. Proof: The proof of Theorem 2.6 can be found in subsection A.4. □

2.5.2 Screening Candidate Pairs

After excluding the non-profitable pairs, our task is to screen the remaining pairs to determine if each is profitable. The primary computation bottleneck in discoverung all such bundles is the estimation of joint probability for each pair and to iterate all the possible pairs, which is a quadratic problem. Since this process is extremely time-consuming, a commonly exploited computational trick is to establish the bounds of the bundling utility, and use these bounds for pre-filtering pairs, which effectively reduces the search space. Obviously, these

bounds need to be much more computationally friendly than the pairwise computing as well as to justify the computing overhead.

Therefore, we aim to find such bounds that only depend on the marginals, so that the assessment of a pair could eliminate unpromising pairs. Specifically, we identify an upper bound condition as follows, which is independent of joint probabilities.

Based on Theorem 2.4, and Theorem 2.5, we know the minimum correlation requirement for an item pair to increase profit at least by δ is Equation 2.20. To avoid computing the exact correlation for all item pairs, we propose to compare an upper bound of correlation ($\overline{\phi}$) with $\phi^*(\delta)$ in Theorem 2.7. The correlation upper bound used in this study is based on TAPER algorithm developed by Xiong et al. [60] which can be found in Equation 2.21 assuming that $N_A \geq N_C$.

$$\overline{\phi}_{AC} = \sqrt{\frac{N_C}{N_A}}\sqrt{\frac{N - N_A}{N - N_C}} \tag{2.21}$$

In particular, for any user-specified threshold $\theta \in [0, 1]$, we can employ the TAPER algorithm developed by Xiong et al. [60] to efficiently find all pairs whose correlation is $\phi \geq \theta$.

Theorem 2.7 (Correlation Upper Bound Condition)**.** *For a given item pair $\{A, C\}$, according to Theorem 2.5, we know the item pair is profitable ($\Delta U_{AC} \geq \delta$) only if $\phi_{AC} \geq \phi^*(\delta)$. In order for ϕ_{AC} to be larger than $\phi^*(\delta)$, its upper bound ($\overline{\phi}_{AC}$) should be larger than $\phi^*(\delta)$ as well. Accordingly, a necessary condition for a profitable item pair is $\overline{\phi}_{AC} \geq \phi^*(\delta)$.*

Proof. Proof: The proof of Theorem 2.7 is shown in subsection A.5. □

2.6 Algorithm Design

In this section, we describe algorithms to solve the bundling problem. First, we represent a straightforward algorithm in subsection 2.6.1, then we describe the designed algorithm in subsection 2.6.2 which is based on results in subsection 2.5.1, and subsection 2.5.2.

2.6.1 The Brute-force Algorithm

The brute-force approach can be found in section B. As a pre-processing step, we extract per item A the required features, price-sensitivity (β_A), normal price (P_A^0), and base demand ($\mathbb{P}_A^0$) (and place them in a table for later use). This pre-processing step allows for a more efficient to make the look-up process in the brute-force approach. Instead of finding each item's features from the main transactional dataset multiple times, we extract them up from the pre-processed table. The brute-force approach begins with two loops in line 2 and line 3 to iterate all possible item pairs. The algorithm computes exact correlations for each item pair (line 4) and compare them with the required minimum correlation for each pair (calculated in line 5). Only item pairs with exact correlations larger than the minimum correlation required are extracted as selected item pairs in $\mathcal{S}$ (line 11).

2.6.2 The PBQ Algorithm

In this section, we describe a Profitable Bundle Query (PBQ) algorithm for detecting profitable item pairs which is developed based on results presented in the previous section. Its pseudocode can be found in B.

Like the brute-force algorithm, as a pre-processing step, we extract the required features for each item and save those in a table for later use as an input in the algorithm. Input variables, the transactional dataset, and δ, the user-specified minimum gain, and s, a discount rate. The output includes item-pairs with a bundling utility greater than δ.

The PBQ algorithm starts with two loops in line 2, and line 3, iterating on all possible pairs included in the pre-processed items table. In this algorithm, our goal is to avoid calculation of the exact correlation as much as possible. To achieve, based on Theorem 2.6, PBQ iterates on pairs with at least one price-sensitive item (line 4). In addition, another pruning is performed in line 7 which is based on Theorem 2.7. In this step, an upper bound of ϕ is calculated based on the TAPER algorithm developed by Xiong et al. [60] which is computationally friendly. Then, the estimated upper bound of ϕ is compared with the minimum correlation required for each item pair that will increase profits by δ, which is calculated using Equation 2.20. Hence, as shown in line 7, PBQ only iterates on item pairs

when $Upper(\phi_{AC}) \geq \phi^*(\delta)$. Then, the algorithm continues to compute the exact correlation for the remaining pairs (line 8), and identifies item pairs output for which $\phi_{AC} \geq \phi^*(\delta)$.

2.7 Experimental Results

Several experiments were conducted to demonstrate the performance of the PBQ algorithm by applying it to a large transactional dataset. A brief description of the dataset can be found in subsection 2.7.1. To illustrate the effectiveness of the algorithm, a proof of concept method is used in subsection 2.7.2, and the comparison of the proposed method with a benchmark method is shown in subsection 2.7.3. Additionally, the performance of the algorithm with the various number of products is compared with the brute force algorithm in subsection 2.7.4. A sample of these results is shown in subsection 2.7.5. Moreover, the selected bundles of two sets of stores are compared in this section. We show that the selected bundles can vary based upon customers demographics.

2.7.1 Data Description

To investigate the impacts of our theoretical analysis empirically, we consider an advanced dataset called Dunnhumby. This dataset contains transactions from 117 weeks representing 2,500 households which shopped a total of 92,339 products. It includes 8 data tables. Among all the transactions, there were some scarce products which can be disregarded in this study.

Platform. All of the experiments were accomplished on a MacBook Pro, 2.7 GHz Intel Core i5 and 8 GB of RAM. All of the programs were implemented in R.

2.7.2 A Proof-of-Concept Analysis

Following the process presented by Xue et al. [62], we present our proof-of-concept analysis in this section. Their idea of a proof-of-concept analysis is to divide the data into two periods. The first period is used for model training (i.e., identifying the most profitable bundles) and the second period is used for model testing (i.e., checking the performance of the identified bundles in a new period).

To better illustrate the results, we limited the number of products examined to the 50 most popular items, which sold in the most transaction. We used 80 weeks of the Dunnhumby dataset for training and the last 22 weeks for testing. Using the training dataset, we estimated components of the utility function such as price-sensitivity (β) and base demand ($\mathbb{P}_A^0$) for each item A and correlation of each potential item pair (ϕ_{AC}). Using the components of the utility function, the most profitable itemsets in the training dataset were identified by the algorithm. We then used the testing dataset to estimate the expected utility of the most profitable itemsets previously detected by training. To present the performance of the proposed algorithm, we compared the average utility of itemsets before bundling and the average expected utility after bundling. In this analysis, the algorithm identified the 30 most profitable item pairs using the training dataset. As presented in Table 2.3, the average utility in segment I decreased after bundling, which is expected, because customers who were already interested in both items had an opportunity to purchase them at a discount; therefore, the utility decreased. However, the expected utility in other segments increased after bundling. The total utility of each bundle also increased roughly by 300 dollars.

In the proposed algorithm, item pairs are selected based on their utility. The higher the utility is, the greater the selection probability will be. In Figure 2.2, we present the utility of item pairs using the training dataset versus their utility in the testing dataset. For this purpose, the top 150 of the most popular items were selected. Within these 150 items, a total of 11,175 potential item pairs existed. The utility of item pairs were computed using both training and testing data are compared in Figure 2.2. The number of transactions in training and testing data were different; hence, the utility of a specific pair using the testing dataset

Table 2.3: POC Method

Segment	Average actual utility (testing dataset without bundling)	Average expected utility (testing dataset with 10% discount on bundles)
I	5317.463	4785.717
II	2399.799	2448.304
III	2063.611	2158.854
IV	0	851.1999
Total	9780.873	10244.0749

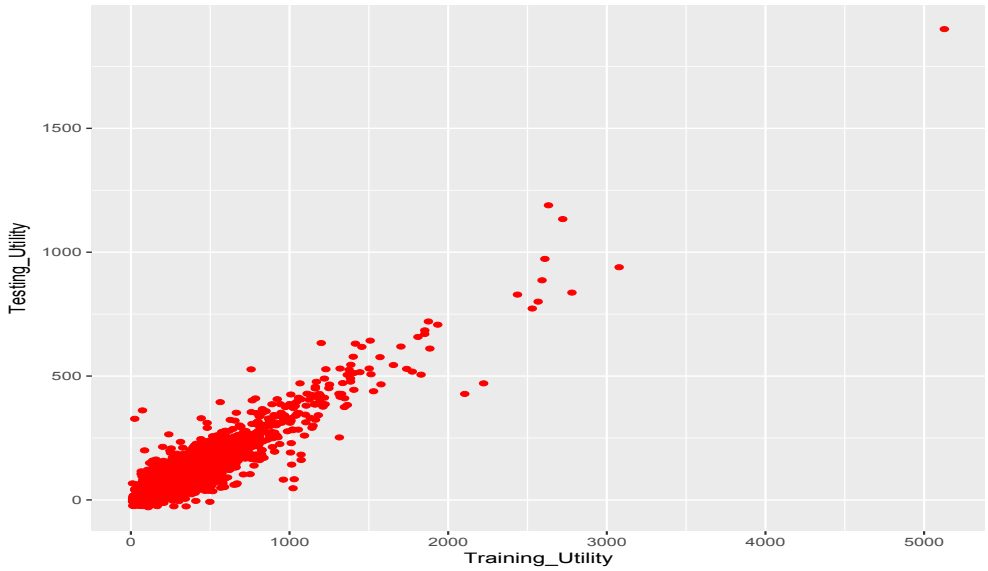


Figure 2.2: Comparison of training utility and testing utility

was not equal to its utility using the training one. The number of transactions included in the testing dataset was less than the training data; hence we expected the utility of each pair in the testing data to be smaller than that found in the training data. Irregardless, irrespective of the datasets, we expected the algorithm to select the same bundles. As presented in Figure 2.2, all points converge along the 45-degree line. Accordingly, the selected pairs in each dataset were identical.

2.7.3 Comparison to a Benchmark Method

The association rule mining technique has been extensively applied to many areas, especially in market basket analysis, such as shelf allocation and product promotion [65]. We practiced the widely applied association rule mining algorithm as the benchmark algorithm in this study. We selected two metrics, the fraction of bundle sales introduced by [7] and the average transaction value, to evaluate the performance of our algorithm and association rule mining. The fraction of bundle sales indicates the attractiveness of a bundle ($\frac{N_{AC}}{N_A+N_C}$) by comparing the number of times two items were purchased together with the number of time at least one of these items was purchased. On the other hand, the rule mining algorithm finds item pairs which have been purchased together most frequently. Accordingly, if two items are often co-purchased, they are more likely to be selected by the association rule mining algorithm

as a frequent association rule, which produces a greater fraction of bundle sale. Hence, comparing our algorithm with association rule mining using the fraction of bundle sales as a metric is fair to the association rule mining method. However, evaluating both methods by their average transaction value may not be fair to the rule mining method because the rule mining method does not evaluate item pairs by their price or utility. However, we aim to present the performance of our algorithm in increasing the fraction of bundle sales and average transaction value at the same time. Our goal is to demonstrate the importance of considering customer reaction in price reduction, rather than just their past reaction toward the co-purchase. We show that our algorithm can detect item pairs with not only a high fraction of bundle sales but also with a higher transactional average sale.

We divided the dataset into the training and testing subsets. Using 20 weeks of Dunnhumby dataset, we utilize the first 15 weeks for training and the last 5 weeks for testing. Then, the PBQ algorithm and the association rule mining approach were separately applied to the training subset. For each algorithm, the identified item pairs were saved. Accordingly, using the testing subset, both metrics were calculated for the identified item pairs by PBQ and the rule mining method.

The results are presented in Table 2.4, which shows, the proposed method outperforms the baseline method by both metrics.

2.7.4 Efficiency

We illustrate the efficiency of our algorithm by choosing the brute force for comparison as shown in Figure 2.3. The brute force algorithm was chosen because there is not an algorithm in the literature with the same function as ours. The running of the brute force and our proposed approach was compared by randomly selecting various numbers of items from all

Table 2.4: Comparing Effectiveness

	Bundle Fraction	Average Sale
Rule Mining	0.0498	\$1.692
PBQ	0.0657	\$4.174

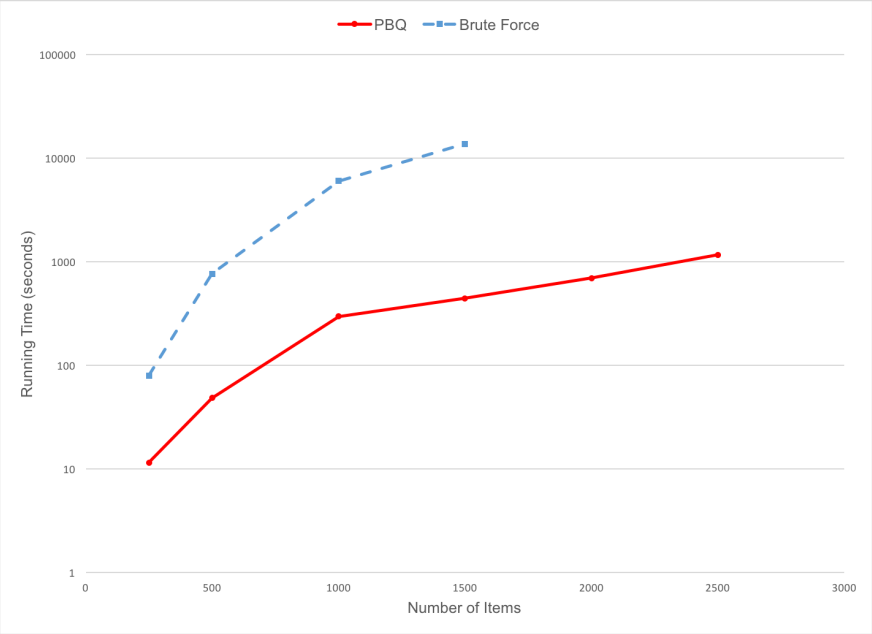


Figure 2.3: Comparison of Running Time

items included in the Dunnhumby dataset. Figure 2.3 shows the median of five runs per number of items. When the number of items increases to 1500, the running time of the brute force increases to more than 6 hours.

2.7.5 Sample Results

Looking at 10 products, we identified the most profitable bundles in PBQ. In this experiment, we selected ten of the most popular or the fastest-moving products from Dunnhumby dataset. The Table 2.6 detailes the selected ten items, their corresponding price without promotion and their price-sensitivity.

There are forty-five possible pairs for ten items. We applied our algorithm to the transactional Dunnhumby dataset containing the items presented in Table 2.6. In this example, we wanted to determine the change in utility when $N = 1,000$, $s = 0.3$ and $\delta = 2,000$. Accordingly, we looked for bundles which yielded to $2,000$ \$ or more revenue with 30 percent discount on each bundle in the next $1,000$ transactions.

Table 2.5 presents PBQ's selected item pairs. Three key factors price-sensitivity, normal price, correlation, contributed to the bundle utility.

Table 2.5: Bundles Selected by PBQ

Item A	Item C	β_A	β_C	π_A	π_C	ϕ
Butter	White bread	0.093	0.261	\$2.50	\$0.99	0.876
Butter	Chocolate milk	0.093	0.421	\$2.50	\$1.00	0.951
Cottage cheese	Juice	0.059	0.163	\$1.99	\$1.50	0.963
Butter	White milk	0.093	0.016	\$2.50	\$2.49	0.740
Cottage cheese	White milk	0.059	0.016	\$ 1.99	\$2.49	0.781

Table 2.6: Item Specification

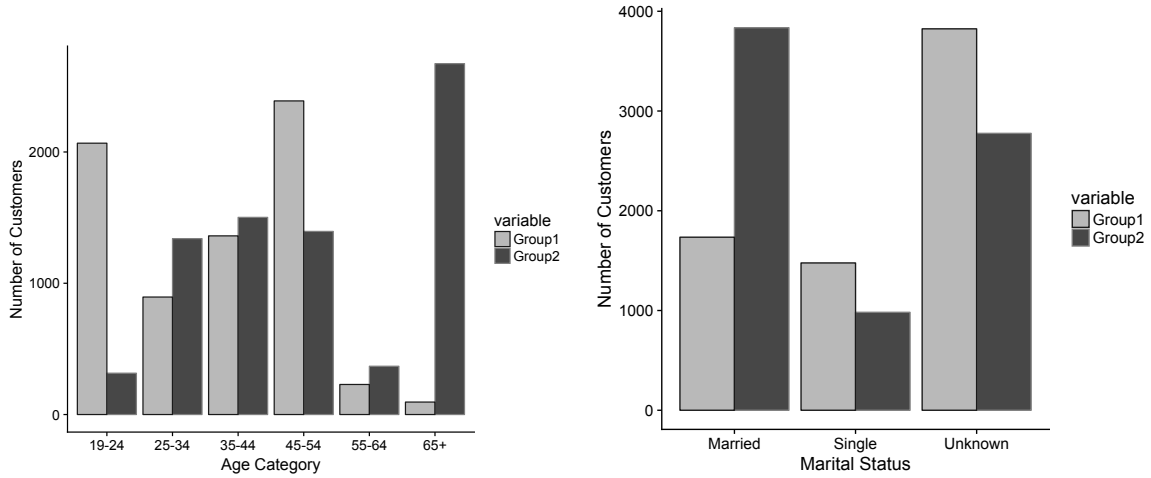
Product ID	Description	Price	Price-sensitivity
981760	Eggs	\$1.09	0.196
976199	Butter	\$2.50	0.093
1071939	Cream cheese	\$1.00	0.069
1029743	Fluid milk	\$2.49	0.016
1056509	Cottage cheese	\$1.99	0.059
883404	White bread	\$0.99	0.261
896369	Peanut butter	\$1.29	0.000
908531	Chocolate milk	\$1.00	0.421
909894	Juice	\$1.50	0.162
907014	Muffin & corn bread	\$0.34	2.657

The selected item pairs in Table 2.5 had great correlation coefficients, which resulted in larger joint probabilities. In this experiment, bundles including muffin were not among profitable ones even though muffin was a price-sensitive item, because the muffin was not strongly correlated with other items. Also, muffin's price was lower compared to other items. Additionally, the item prices are also critical in the determination of utility. For example, butter and white milk had lower correlation and price-sensitivity compared to other items, but their prices were higher which made their utility great enough to be selected as a profitable bundle.

The most profitable bundles varied by stores depending on the distribution of customers' age and their marital status. To show this difference, we selected two sets of stores in which the distribution of customers' age and their marital status varied. Selecting two sets of stores, revealed the first set of stores contained more customers in their later adolescence (19-25 years old), while the second set of stores had markedly more customers in their later adulthood (65+ years old). 2.4a compares the distribution of customers' age category in two groups, while 2.4b compares the distribution of customers' marital status. Table 2.7 shows the top selected bundles by the PBQ algorithm for each set of stores. As illustrated in Table 2.7, the most profitable bundles differed by store sets. For instance, bundles including cigarette were more profitable in the group of younger customers, while bundles including berries were more favorable in the group of older customers. This set of results indicates that finding profitable bundles at a refined level (e.g., specific stores and/customer profiles) will provide more accurate and varied results.

Table 2.7: Comparing Selected Bundles of Two Store Sets

Group 1		Group 2	
Item A	Item C	Item A	Item C
CIGARETTES	VEGETABLES SALAD	POTATOES	VEGETABLES SALAD
CIGARETTES	BAG SNACKS	POTATOES	CANNED JUICES
CIGARETTES	FLUID MILK PRODUCTS	SOFT DRINKS	BERRIES
BAG SNACKS	DELI MEATS	BAG SNACKS	DELI MEATS
CHEESE	SOFT DRINKS	BERRIES	DELI MEATS
FLUID MILK PRODUCTS	BAG SNACKS	VEGETABLES SALAD	CHICKEN



(a) Age Group **(b)** Marital Status

Figure 2.4: Comparison of Customer Demographics.

2.8 A BOGO Case Study

Oftentimes stores run Buy One Get One s% discount promotions (25%, 50% and free (100%) discount on the second item). If we wish to consider BOGO cases, as a bundle problem, specific case must be created in which we identify customers interested in purchasing one unit of an item versus customers interested in purchasing multiple units.

The distribution of order quantities is presented in Figure 2.5. In this study, the top $1,000$ popular items were selected. The number of transactions with various item order quantities that included 1 to 5 units per item was calculated and is presented in Figure 2.5.

As illustrated in Figure 2.5, the number of transactions with more than two units of an item was almost six percent of all transactions. Therefore, because of the data sparsity of order quantities more than two units, we limited our case study to order quantities of one or two units.

In this case study, we aimed to estimate the increase in purchasing multiple units of an item after a BOGO offer. This estimate indicated the extent to which an item was profitable for a BOGO offer. To this aim, we explored the change in utility for each segment of customers. Due to the similarity of two items in the BOGO offer, customers were divided into three segments in this case.

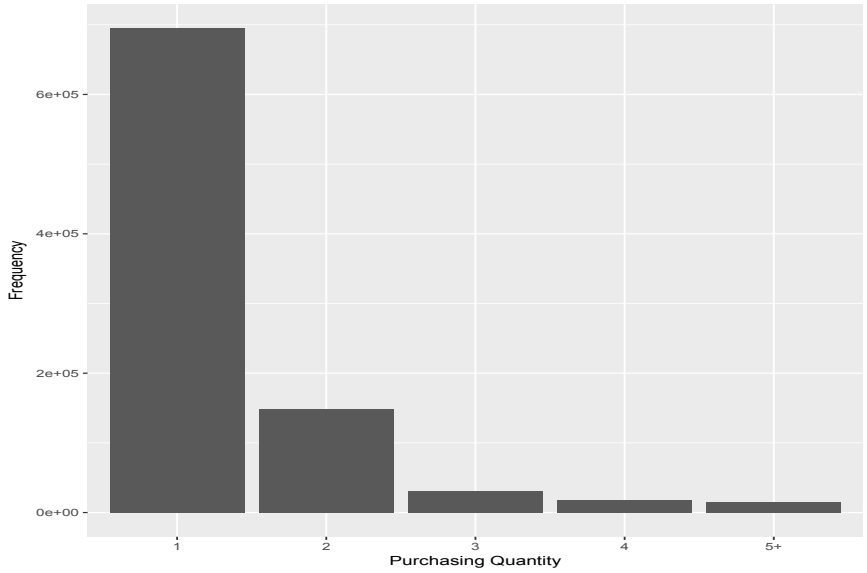


Figure 2.5: The distribution of order quantities

Segment I in the BOGO case contains all customers interested in purchasing two units of an item. Offering BOGO causes a loss of profit in this segment. The change in utility is expressed by Equation 2.22.

$$\Delta U_I \quad = \quad -N_{\{A=2\}} \times d \tag{2.22}$$

In Equation 2.22, $N_{\{A=2\}}$ is the number of customers who purchase two units of an item, and d, is the discount offered for the second item which is $d = \pi_A^0 \cdot s\%$.

Segment II and III in our previous formulation are combined into segment II in this case because two items are the same. Since customers are already interested in purchasing one unit of an item in this segment, we assume they perceive the discount on the second unit of the item. Therefore, the probability of purchasing after a BOGO offer for this segment is equal to the probability of purchasing the second unit after receiving the discount. Segment II includes customers who are only interested in purchasing one unit of an item. The number of the customers in this segment is $N_{\{A=1\}}$. The change in utility after the BOGO offer is

$$\Delta U_{II} \quad = \quad N_{\{A=1\}} \cdot (\mathbb{P}_A^d - \mathbb{P}_A^0) \cdot (\pi_A^d). \tag{2.23}$$

In Equation 2.23, $\mathbb{P}_A^d$ is the probability of purchasing one unit of item A after discount while $\mathbb{P}_A^0$ is the probability of purchasing one unit of an item before any discount.

Segment III includes customers who are not interested in purchasing A at all. Customers in segment III are not interested in purchasing neither a single unit of an item nor multiple. The portion of customers in segment III is $N_{\bar{A}} = N - N_{A=1} - N_{A=2}$. Offering an $s\%$ discount for the second unit of an item is equal to offering two units of an item with $\frac{s\%}{2}$ discount per unit. Moreover, $\mathbb{P}_{A=2}^{d/2}$ presents the probability of purchasing two units with $d/2$ discount. The probability of purchasing with the BOGO offer in this segment would be

$$\Delta U_{III} = N_{\bar{A}} \cdot (\mathbb{P}_{\{A=2\}}^{d/2} - \mathbb{P}_{\{A=2\}}^{0}) \cdot (\pi_A^{d/2} + \pi_A^{d/2}). \tag{2.24}$$

To estimate the probability of purchasing multiple units of an item, we only included the transactions with multiple order quantities and a similar way to estimate the probability of such a paper is shown in Equation 2.1. We follow the same logic to find the probability of purchasing a single unit after d discount which is $\mathbb{P}_A^d$.

We conducted several experiments in this section to detect the most favorable items for the BOGO offer. We randomly selected 200 items from top $1,000$ most popular purchased items and solved the problem to determine the most profitable candidates for the BOGO offer with three levels of promotions (25%, 50% and 100% discount on the second item). Our analysis showed by increasing the discount level, the number of items with a positive utility increased, while the utility itself decreased. The selected items for each discount level, were not necessarily the same, which shows that one item can be a proper candidate with an offer of Buy One Get One 25% off but it may not be a favorable candidate for a Buy One Get One free offer. In this case study, we detected items with positive ΔU for each discount level. Then, the average ΔU over selected items was calculated. These average ΔU along with the number of selected items are shown in Table 2.8.

Table 2.8: A sample results of BOGO offer with three levels of promotion

Discount	Average ΔU	# Items Selected	Total ΔU
20 %	40.93	43	1759.91
25 %	38.73	43	1665.23
50 %	26.59	45	1196.63
100 %	6.55	89	583.08

2.9 Conclusions

In this study, we solved the mixed-bundling problem considering a realistic scenario by exploring items' price-sensitivity, correlation, and effects of promotion. We formulated the bundling utility function per customer segment. Then, we provided a couple of theoretical conditions under which bundling is not profitable. We also developed an upper bound condition to avoid unnecessary pairwise calculations as much as possible. We applied theoretical results and the upper bound condition to develop the pruning algorithm (PBQ). The efficiency and effectiveness of the PBQ algorithm are illustrated by comparing the algorithm with the brute force and association rule mining algorithms respectively. The proposed algorithm outperformed the rule mining algorithm by identifying bundles with higher fractions of bundle sale and average sales of bundles per transaction.

More fine-grained (e.g., UPC level and store specific) bundle discovery should and could be used in practice. This is especially meaningful for retailers who carry a wide variety of brands and product categories.

Our study has a few limitations which can be extended in future studies. First, we assumed that customers either purchase an item or do not at all. We did not discriminate between order quantities. However, in a case study we extended our formulation to investigate a BOGO offer and show under which conditions a BOGO offer would be recommended. Although the BOGO case study is a specific case of bundling which investigates order quantities, studying actual order quantities in bundling is not in the scope of this study. Second, our study is limited to bundles of two elements. An interesting extension for this study would be to explore the orders quantities in bundling problems as well. Another possible study would be to investigate bundles of three or more items.

Moreover, discount optimization for profitable bundles is another attractive extension. Despite these limitations, our research develops a data-driven method for discovering bundles considering their price-sensitivity and correlation which is unprecedented to the best of our knowledge. Moreover, this method is applicable to large transactional data sets without any limitation.

Chapter 3

Simultaneous Detection of Adverse Drug Reaction and At-Risk Groups

Association mining in high-dimensional data has been a core problem in data mining. Existing work has extensively studied computation properties of a number of association measures and their application in data-intensive applications. While effectively measuring the association between item pairs has been challenging, studying the interplay among multiple factors is even more perplexing. In particular, association patterns discovered at the global level may not be consistent with its counterpart in a local segment. In this paper, we focus on the application of detecting adverse drug reactions that may not be visible at the global level, and we need to identify local segments in which the association between drugs and adverse reactions are strong. This will help identify conditions under which those patients are facing elevated risk, as well as detecting adverse reactions as early as possible.

3.1 Introduction

Post-market monitoring to detect adverse drug reactions (ADRs) [37] is required for medication safety. ADR is any unexpected and undesired effect of a drug being used at normal doses in humans [44]. Since 5% of all deaths are certainly or probably drug-related, ADR has been a significant cause of death [30]. Using data in medical records and ADR

reporting systems has become a valuable source for post-marketing studies, which provides an opportunity for ADR detection [37].

The accuracy of the correlation detection between drugs and ADRs has been investigated in several studies [37, 35, 36]. Many studies have employed correlation analysis [51, 17, 61, 20], among those there are several studies more focused on increasing accuracy of ADR detection.

Previous studies mainly focused on identifying strong associations between drugs and ADRs at the global level [37, 35, 36]. In other words, they concentrated on finding the associations at upper level without considering risk factors. However, little emphasis has been given to identifying the association between drugs and ADRs in smaller local segments. The well-known discrepancy in association measures between global level and the local level, known as the Simpson's paradox, indicates that ignoring local levels in association computing may cause false negative or false positive problems [66]. In the context of ADR detection, false negatives happen when the association between an ADR and a drug is not significant in general while it is significant in a particular subgroup; and false positives are the opposite of false negatives that happen when an association is significant globally, but insignificant locally or by considering subgroups. While some of the previous studies explored the role of confounding or false positives in association mining [2, 2, 36, 56, 66], false negatives have rarely been investigated.

In this study, we explore the problem of detecting false negatives in which drugs and ADRs are not significantly associated at a global level while there are associated at patient subgroups. Subgroups are groups of patients with one or more similar demographic and previous medication features. To discover the correlation between drugs and ADRs locally, we should find the association between each drug and ADR at each subgroup.

In this study, we attempt to meet a few challenges. It is challenging to effectively and efficiently detect false negatives in large datasets. Checking all triplets of drugs, ADRs, and subgroups make the problem as complex as $O(n^3)$, which is not scalable in large datasets. Also, we need to specify both pair and triplet co-occurrences multiple times to identify the correlation between drugs and ADRs controlling for patient subgroups. Although we may save pair co-occurrences for further lookup, it requires $O(n^2)$ memory space and is

not practical for larger scaled problems. Additionally, considering patient demographic and medical history features and various combinations of these features may result in numerous patient subgroups. Not only increasing subgroups raises the complexity of the problem, but also considering all features and their combinations is not practical in the real world. The reason is that previous records may not contain several of the possible subgroups. Accordingly, we need a powerful method to discover a practical number of most significant patient subgroups.

To discover false negatives in large datasets efficiently, we analyze properties of the three most popular association measures, including Pearson correlation, odds ratio, and relative risk. We first found three initial upper bounds for our three measurements using only marginal values. These upper bounds help in pruning unnecessary pair and triplet co-occurrence calculations. Then, we provide secondary bounds using marginal values and pair co-occurrences. These bounds are more powerful pruning than the initial bounds and reduce the number of triplet calculations further. We propose a pruning algorithm, which has two steps of pruning based on initial and secondary bounds. To enhance the efficiency of the pruning algorithm, we develop a save and lookup algorithm. First, we define necessary conditions for pair co-occurrences per association measure. These necessary conditions specify the required setting of pair co-occurrences for a triplet to have a significant association. Then, we develop a save and lookup algorithm based on the necessary conditions. The save and lookup algorithm has two phases: saving strong co-occurrences and searching false negatives using secondary bounds. Additionally, to discover the most frequent patient subgroups, we apply the association rule mining technique to practically select patient segments. We show the effectiveness and efficiency of our algorithms using two large datasets, including the FDA Adverse Events Reporting System and Vaccine Adverse Events Reporting System.

We contribute to the previous ADR detection studies in several ways:

- We study false negatives in ADR detection, which has been rarely investigated before.
- To accurately detect ADRs, we find the association between drugs and ADRs at a local level by considering patient specifications, including demographic and medical history.

- We apply the association rule mining technique to discover the most practical patient segments.
- We propose a framework for search space pruning to make false negatives detection in large datasets scalable.
- We extend our search space pruning framework to three association measurements, including Pearson correlation, odds ratio, and relative risk.

The rest of this paper is organized as follow: section 3.2 summarizes the most relevant previous studies to this research. section 3.3 describes the preliminaries, the problem statement and the scope of this study. subsection 3.3.2 includes descriptions of three association measures applied in this study. section 3.4 develops the computational conditions to reduce the number of triplets including drugs, ADRs, and segments for checking using each correlation measure. section 3.5 describes the algorithms including the brute force algorithm and pruning algorithms.subsection 3.5.1 is segment generation using association rule mining. subsection 3.5.2 describes the brute force algorithm. subsection 3.5.3 develops the pruning algorithms for all correlation measurements. section 3.6 summarizes the experiments and presents to what extend the proposed method is effective and efficient. Finally, section 3.7 draws the conclusions and future studies.

3.2 Related Work

Adverse drug reactions (ADRs) are still among the leading causes of mortality and impose billion dollars of cost every year[36, 52]. Postmarketing surveillance provides a vast collection of adverse events reports which contribute a lot to ADR studies.

Several previous studies focused on improving ADR detection accuracy by using adverse events reports and Electrical Medical Records (EMR.) Liu et al. [37] explored using EMR to increase the accuracy of ADR detection. They applied various measurements containing odds ratio, Chi-squared, and Yule. Varallo et al. [56] studied the performance of triggers in detecting ADRs to optimize risk management in hospitals and safety care. They conducted a 6-month study to illustrate the underestimate rate of tool and health professionals resulted

by risk factors. They evaluated the positive predictive value (PPV), odds ratio, and Chi-squared test to reveal statistical significance in their study.

However, the occurrence of risk factors challenges the accuracy of ADR detection. Ignoring impacts of ADR at local levels of patients results in losing critical information. Patient subgroups with specific risk factors may be more subjected to ADRs, and the occurrence of risk factors challenges the accuracy of ADR detection despite accurate association measures.

Many previous studies focused on detecting false positives accounting for confounders. All factors which are associated with both drug and ADR are potential confounders. Li et al. [36] developed a data-driven method to identify ADR signals considering confounders. They applied penalized logistic regressions to estimate confounder-adjusted ADR associations. In their proposed method, they explored two associations, including the association between the risk factors and ADR and the association between the medication and ADR. They applied their method to rhabdomyolysis and pancreatitis, which are severe ADRs. Some other studies explored the confounding effects as well [55, 52, 42]

Some of the previous studies focused on drug-drug interactions without exploring the impact of patients segmentation. Guthrie et al. [26] examined the change in rates of polypharmacy and potentially drug-drug severe interactions using multilevel logistic regression. Köhler et al. [33] used odds ratio and adjusted odds ratio to determine the association between elderly patients with drug toxicities and drug interactions. Some other studies explored the impact of risk factors as well to detect drug-drug interactions more accurate. Tatonetti et al. [52] adopted a propensity score match(PSM) method to use only co-reported drugs hypothesizing many risk factors correlate with them, and they do not need to be modeled.

Most studies in literature concentrated on improving effectiveness rather than efficiency. Moreover, the majority of existing researches focused on strongly correlated drugs and ADRs to evaluate the effectiveness of measurements in the presence of risk factors and detect false positives. Nevertheless, exploring false negatives is equally critical to specify the significant patient subgroups who are at elevated risk of ADRs.

In this study, we detect false negatives to discover the subgroups of patients who are at risk of ADRs. To this end, we computationally reduce the size of the problem and develop pruning algorithms for three correlation measurements.

3.3 Preliminaries and Problem Statement

We introduce concepts, notations, and formulation in this section to define the problem more accurately.

3.3.1 Problem Formulation

In this study, we use a significant segment for a group of patients in whom a drug has an ADR impact while in general, the drug and ADR are not correlated. Our goal is to discover false negative patterns. For example, one drug and an adverse drug reaction may not be significantly correlated while the drug has the significant adverse drug reaction in one subpopulation. Existing significant segments may cause false negative patterns.

Formally, let us assume that we have

- $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, each A_i $(i = 1, 2, \ldots, N)$ is a binary indicator of a *known treatment or behavior* (e.g., whether the patient took a given medicine/vaccine).
- $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$, each B_j $(j = 1, 2, \ldots, M)$ is a binary indicator of an *interested outcome* (e.g., whether the patient had a given symptom).
- $\mathcal{C} = \{C_1, C_2, \ldots, C_P\}$, each C_k $(k = 1, 2, \ldots, P)$ is a binary indicator of a *segmentation factor* (e.g., whether the patient belongs to a given demographic or history segment).

Suppose that r is an association measure, where

- r_{AB} measures the association between items A and B using all data and
- $r_{AB|C}$ measures its counterpart within a segment defined by a conjunction of conditions C.

Definition 3.1 (Significant Segments Discovery)**.** *Our goal is to find all patterns like* $\{A_iB_j|C\}$ *such that while* $r_{A_iB_j}$ *is not significant,* $r_{A_iB_j|C}$ *is significant.*

We aim to find false negative patterns in drugs and ADRs associations. We discover all non strong associated drugs and ADRs which are strongly associated considering a significant segment or a group of segments.

3.3.2 Binary Data Association Measures

To study the association between two binary variables, we often can lay out a 2×2 contingency table like Table 3.1. In this table, A typically represents some treatment group and B represents the outcome. N_{AB} represents the number of cases such that both $A = 1$ and $B = 1$ hold true, whereas $N_{A\bar{B}}$ represents the number of cases such that both $A = 1$ and $B = 0$, and so on.

Based on quantities listed in this table, the Pearson's correlation, the odds ratio, and the relative risk can be calculated by following , respectively. To unify our associations' notations in this study, we use γ^p_{AB}, γ^r_{AB}, and γ^o_{AB} to show Pearson's correlation, Relative Risk, and Odds Ratio respectively.

Definition 3.2 (Pearson's Correlation)**.** *Using the contingency table as shown in Table 3.1, the* γ^p_{AB} *correlation will be*

$$\gamma^p_{AB} = \frac{NN_{AB} - N_A N_B}{\sqrt{N_A(N - N_A)N_B(N - N_B)}}. \tag{3.1}$$

Table 3.1: The Contingency Table.

	B = 1	B = 0	Total
A = 1	N_{AB}	$N_{A\bar{B}}$	N_A
A = 0	$N_{\bar{A}B}$	$N_{\bar{A}\bar{B}}$	$N_{\bar{A}}$
Total	N_B	$N_{\bar{B}}$	N

Definition 3.3 (Relative Risk)**.** *Using the contingency table as shown in Table 3.1, where A represents a treatment and B represents an outcome, the relative risk is calculated as*

$$\gamma^r_{AB} = \frac{N_{AB}/N_A}{N_{\bar{A}B}/N_{\bar{A}}} = \frac{N_{AB}N_{\bar{A}}}{N_{\bar{A}B}N_A}. \tag{3.2}$$

Definition 3.4 (Odds Ratio)**.** *Using the contingency table as shown in Table 3.1, where A represents a treatment and B represents an outcome, the odds ratio is calculated as*

$$\gamma^o_{AB} = \frac{N_{AB}/N_{A\bar{B}}}{N_{\bar{A}B}/N_{\bar{A}\bar{B}}} = \frac{N_{AB}N_{\bar{A}\bar{B}}}{N_{\bar{A}B}N_{A\bar{B}}}. \tag{3.3}$$

3.3.3 Association Measures with Segmentation

When considering a segmenting condition C, we can come up with a stratified contingency table like Table 3.2.

Definition 3.5 (Segment Correlation)**.** *The correlation between factors A and B in and out of segment C will be*

$$\gamma^p_{AB|C} = \frac{N_C N_{ABC} - N_{AC}N_{BC}}{\sqrt{N_{AC}(N_C - N_{AC})N_{BC}(N_C - N_{BC})}}. \tag{3.4}$$

Definition 3.6 (Conditioned Relative Risk)**.** *The relative risk between treatment A and outcome B in segment C will be*

$$\gamma^r_{AB|C} = \frac{N_{ABC}(N_{\bar{A}BC} + N_{\bar{A}\bar{B}C})}{N_{\bar{A}BC}(N_{ABC} + N_{A\bar{B}C})} = \frac{N_{ABC}N_{\bar{A}C}}{N_{\bar{A}BC}N_{AC}}. \tag{3.5}$$

Table 3.2: The Stratified Contingency Table

Segment	Treatment	B = 1	B = 0	Total
$C = 1$	A = 1	N_{ABC}	$N_{A\bar{B}C}$	N_{AC}
	A = 0	$N_{\bar{A}BC}$	$N_{\bar{A}\bar{B}C}$	$N_{\bar{A}C}$
	Total	N_{BC}	$N_{\bar{B}C}$	N_C
$C = 0$	A = 1	$N_{AB\bar{C}}$	$N_{A\bar{B}\bar{C}}$	$N_{A\bar{C}}$
	A = 0	$N_{\bar{A}B\bar{C}}$	$N_{\bar{A}\bar{B}\bar{C}}$	$N_{\bar{A}\bar{C}}$
	Total	$N_{B\bar{C}}$	$N_{\bar{B}\bar{C}}$	$N_{\bar{C}}$
Total		N_B	$N_{\bar{B}}$	N

Definition 3.7 (Conditioned Odds Ratio)**.** *The odds ratio between treatment A and outcome B in segment C will be*

$$\gamma^{o}_{AB|C} = \frac{N_{ABC}N_{\bar{A}\bar{B}C}}{N_{A\bar{B}C}N_{\bar{A}BC}}. \tag{3.6}$$

3.4 Problem Analysis

Our objective is to specify subgroups of patients with a significant correlation between ADRs and drugs. To this aim, we need to find correlations of drugs (A) and ADRs (B) in all subgroups (C). In order to avoid triplet calculations as much as possible and limit the number of potential triplets, we propose two upper bounds for three correlation measures.

When the upper bound of a triplet is below the threshold (θ), the actual correlation is definitely below the threshold. Accordingly, we can drop the triplet from further exploration without finding the actual correlation and by limiting the calculations to marginal values.

3.4.1 Combination of Binary Events

Since counting coocurrences can be major bottleneck for computing time and memory, upper bounds for the association measures may be developed using only marginal values.

Specifically, let

$$\tilde{N}_{XY} = \min\{N_X, N_Y\}, \tag{3.7}$$

where individual events $\forall X, Y \in \{A, B, C, \bar{A}, \bar{B}, \bar{C}\}$. We have $\tilde{N}_{AC} = \min\{N_A, N_C\}$, $\tilde{N}_{\bar{B}C} = \min\{N_{\bar{B}}, N_C\}$, etc. Then $\tilde{N}_{XY}$ can serve as an upper bound for N_{XY}.

Similarly, let

$$\tilde{N}_{XYZ} = \min\{N_{XY}, N_{XZ}, N_{YZ}\}, \tag{3.8}$$

where individual events $\forall X, Y, Z \in \{A, B, C, \bar{A}, \bar{B}, \bar{C}\}$. Then we have

$$\begin{aligned} N_{XYZ} &\leq \tilde{N}_{XYZ} = \min\{N_{XY}, N_{XZ}, N_{YZ}\} && (3.9)\\ &\leq \min\{N_X, N_Y, N_Z\}. && (3.10) \end{aligned}$$

3.4.2 Initial Bounds

The initial upper bound for Pearson's correlation, Relative Risk, and Odds Ratio can be found in Theorem 3.8, Theorem 3.9, and Theorem 3.10, respectively.

Lemma 3.8 (The Initial Upper Bound for Pearson's Correlation)**.** *$\gamma^p_{AB|C}$ is not significant if:*

$$\min\{s_1, s_2\} \leq \theta \tag{3.11}$$

where

$$s_1 = \sqrt{\frac{\tilde{N}_{\bar{A}C}\tilde{N}_{BC}}{(N_C - \tilde{N}_{\bar{A}C})(N_C - \tilde{N}_{BC})}}; \tag{3.12}$$

$$s_2 = \sqrt{\frac{\tilde{N}_{AC}\tilde{N}_{\bar{B}C}}{(N_C - \tilde{N}_{\bar{B}C})(N_C - \tilde{N}_{AC})}}. \tag{3.13}$$

Proof. Since $N_{ABC} \leq N_{AC}$, according to Equation 3.4, we have

$$\begin{aligned} \gamma^p_{AB|C} &\leq \frac{N_C N_{AC} - N_{AC}N_{BC}}{\sqrt{N_{AC}(N_C - N_{AC})N_{BC}(N_C - N_{BC})}} \\ &= \sqrt{\frac{N_{AC}(N_C - N_{BC})}{N_{BC}(N_C - N_{AC})}} \\ &= \sqrt{\frac{N_{AC}N_{\bar{B}C}}{(N_C - N_{\bar{B}C})(N_C - N_{AC})}}, \end{aligned} \tag{3.14}$$

the right hand side of which is an increasing function of both N_{AC} and $N_{\bar{B}C}$. Therefore, knowing that $N_{AC} \leq \min\{N_A, N_C\}$ and $N_{\bar{B}C} \leq \min\{N_{\bar{B}}, N_C\}$, we have

$$\gamma^p_{AB|C} \leq \sqrt{\frac{\tilde{N}_{AC}\tilde{N}_{\bar{B}C}}{(N_C - \tilde{N}_{\bar{B}C})(N_C - \tilde{N}_{AC})}} \equiv s_1, \tag{3.15}$$

where $\tilde{N}_{AC} = \min\{N_A, N_C\}$ and $\tilde{N}_{\bar{B}C} = \min\{N_{\bar{B}}, N_C\}$. Similarly, we can also find that

$$\gamma^p_{AB|C} \leq \sqrt{\frac{\tilde{N}_{BC}\tilde{N}_{\bar{A}C}}{(N_C - \tilde{N}_{\bar{A}C})(N_C - \tilde{N}_{BC})}} \equiv s_2, \tag{3.16}$$

where $\tilde{N}_{AC} = \min\{N_A, N_C\}$ and $\tilde{N}_{\bar{B}C} = \min\{N_{\bar{B}}, N_C\}$. Hence, an upper bound for γ^p_{AB} would be

$$upper_I(\gamma^p_{AB|C}) = \min(s_1, s_2) \tag{3.17}$$

□

Lemma 3.9 (The Initial Upper Bound for Relative Risk)**.** *$\gamma^r_{AB|C}$ is not significant if:*

$$\frac{\tilde{N}_{\bar{A}C}}{N_{\bar{A}B} - \min(N_{\bar{A}B}, N_{\bar{C}})} \leq \theta \tag{3.18}$$

Proof. Based on Equation 3.5, we have:

$$\frac{N_{ABC}N_{\bar{A}C}}{N_{\bar{A}BC}N_{AC}}$$

We know $N_{ABC} \leq N_{AC}$, by replacing N_{AC} with N_{ABC} we find an upper bound as

$$\frac{N_{ABC}N_{\bar{A}C}}{N_{\bar{A}BC}N_{AC}} \leq \frac{N_{\bar{A}C}}{N_{\bar{A}BC}} = \frac{N_{\bar{A}C}}{N_{\bar{A}B} - N_{\bar{A}B\bar{C}}}. \tag{3.19}$$

We can show Equation 3.19 is an increasing function of $N_{\bar{A}C}$, and $N_{\bar{A}B\bar{C}}$. Therefore, we replace them with their upper bounds ($\tilde{N}_{\bar{A}C} = \min(N_{\bar{A}}, N_C)$, $\tilde{N}_{\bar{A}B\bar{C}} = \min(N_{\bar{A}B}, N_{\bar{C}})$). The upper bound is :

$$upper_I(\gamma^r_{AB|C}) = \frac{\tilde{N}_{\bar{A}C}}{N_{\bar{A}B} - \tilde{N}_{\bar{A}B\bar{C}}}$$

□

Lemma 3.10 (The Initial Upper Bound for Odds Ratio)**.** *γ^o_{AB} is not significant if:*

$$\frac{\tilde{N}_{ABC}\tilde{N}_{\bar{A}\bar{B}C}}{(N_{A\bar{B}} - \tilde{N}_{A\bar{B}\bar{C}})(N_{\bar{A}B} - \tilde{N}_{\bar{A}B\bar{C}})} < \theta \tag{3.20}$$

where $\tilde{N}_{ABC} = \min(N_{AB}, N_C)$, $\tilde{N}_{AC} = \min(N_A, N_C)$, $\tilde{N}_{B\bar{C}} = \min(N_B, N_{\bar{C}}), \tilde{N}_{\bar{A}\bar{B}C} = \min(N_{\bar{A}\bar{B}}, N_C)$, $\tilde{N}_{\bar{A}\bar{B}C} = \min(N_{\bar{A}\bar{B}}, N_C)$, $\tilde{N}_{A\bar{B}\bar{C}} = \min(N_{A\bar{B}}, N\bar{C})$, $\tilde{N}_{\bar{A}B\bar{C}} = \min(N_{\bar{A}B}N_{\bar{C}})$ *in above Equations.*

Proof. Using Equation 3.6, we have:

$$\frac{N_{ABC}N_{\bar{A}\bar{B}C}}{N_{A\bar{B}C}N_{\bar{A}BC}} = \frac{N_{ABC}N_{\bar{A}\bar{B}C}}{(N_{A\bar{B}} - N_{A\bar{B}\bar{C}})(N_{\bar{A}B} - N_{\bar{A}B\bar{C}})} \tag{3.21}$$

Equation 3.21 is an increasing function of N_{ABC} and $N_{\bar{A}\bar{B}C}$, $N_{\bar{A}B\bar{C}}$ and $N_{A\bar{B}\bar{C}}$. Hence, the upper bound is gained by replacing their upper bounds :

$$upper_I(\gamma^o_{AB|C}) = \frac{\tilde{N}_{ABC}\tilde{N}_{\bar{A}\bar{B}C}}{(N_{A\bar{B}} - \tilde{N}_{A\bar{B}\bar{C}})(N_{\bar{A}B} - \tilde{N}_{\bar{A}B\bar{C}})} \tag{3.22}$$

□

3.4.3 Secondary Bounds

In order to find a tighter bound to limit the number of potential triplets further, we develop a secondary upper bound in this section. These upper bounds are identified using marginal values (N_A, N_B, N_C), and pairwise associations (r_{AB}, r_{AC}, r_{BC}).

Let $\tilde{N}_{ABC} = \min(N_{AC}, N_{AB}, N_{BC})$, $\tilde{N}_{\bar{A}\bar{B}C} = \min(N_{\bar{A}\bar{B}}, N_{\bar{A}C}, N_{\bar{A}B})$. The Secondary upper bound for Pearson correlation, Relative Risk, and Odds Ratio can be found in Theorem 3.11, Theorem 3.12, and Theorem 3.13.

Lemma 3.11 (The Secondary Upper Bound for γ^p_{AB})**.** *The upper bound of the association between $\{A, B\}$ given C depending on the association measure is as follow: An upper bound of Pearson Correlation is*

$$upper_{II}(\gamma^p_{AB|C}) = \frac{N_C\tilde{N}_{ABC} - N_{AC}N_{BC}}{\sqrt{N_{AC}(N - N_{AC})N_{BC}(N - N_{BC})}} \tag{3.23}$$

Proof. By Equation 3.4, we have:

$$\gamma^p_{AB|C} = \frac{N_C N_{ABC} - N_{AC}N_{BC}}{\sqrt{N_{AC}(N_C - N_{AC})N_{BC}(N_C - N_{BC})}} \tag{3.24}$$

Finding N_{ABC} for each triplet of $\{A, B, C\}$ is the bottleneck of estimation γ^p_{AB}. We know: $N_{ABC} \leq \min(N_{AC}, N_{AB}, N_{BC})$, and γ^p_{AB} is an increasing function of N_{ABC}. Accordingly, by

replacing N_{ABC} with its upper bound $\min(N_{AC}, N_{AB}, N_{BC})$, we gain an upper bound for γ^p_{AB}. □

Lemma 3.12 (The Secondary Upper Bound for RR)**.** *An upper bound of the relative risk (RR) is*

$$upper_{II}(\gamma^r_{AB|C}) = \frac{\tilde{N}_{ABC} N_{\bar{A}C}}{N_{AC}(N_{BC} - \tilde{N}_{ABC})} \tag{3.25}$$

Proof.

$$\begin{aligned} \gamma^r_{AB|C} &= \frac{N_{ABC} N_{\bar{A}C}}{N_{\bar{A}BC} N_{AC}} \\ &= \frac{N_{ABC} N_{\bar{A}C}}{(N_{BC} - N_{ABC}) N_{AC}} \end{aligned} \tag{3.26}$$

To avoid unnecessary triplet calculation, we need to replace N_{ABC} in Equation 3.26 by marginal values. We can show Equation 3.26 is an increasing function of N_{ABC}. Hence, an upper bound of γ^r_{AB} can be found by replacing N_{ABC} by its upper bound. Knowing $N_{ABC} \leq \min\{N_{AB}, N_{AC}, N_{BC}\}$, we have:

$$upper_{II}(\gamma^p_{AB|C}) = \frac{\tilde{N}_{ABC} N_{\bar{A}C}}{(N_{BC} - \tilde{N}_{ABC}) N_{AC}}, \tag{3.27}$$

□

Lemma 3.13 (The Secondary Upper Bound for OR)**.** *An upper bound of the Odds Ratio (OR) is*

$$upper_{II}(\gamma^p_{AB|C}) = \frac{\tilde{N}_{ABC} \tilde{N}_{\bar{A}\bar{B}C}}{(N_{AC} - \tilde{N}_{ABC})(N_{BC} - \tilde{N}_{ABC})} \tag{3.28}$$

Proof. We have

$$\frac{N_{ABC} N_{\bar{A}\bar{B}C}}{N_{A\bar{B}C} N_{\bar{A}BC}} = \frac{N_{ABC} N_{\bar{A}\bar{B}C}}{(N_{AC} - N_{ABC})(N_{BC} - N_{ABC})} \tag{3.29}$$

We can show that Equation 3.29 is an increasing function of both N_{ABC} and $N_{\bar{A}\bar{B}C}$. By replacing them with their upper bounds we gain the following upper bound in which $\tilde{N}_{ABC} = \min(N_{AB}, N_{AC}, N_{BC})$ and $\tilde{N}_{\bar{A}\bar{B}C} = \min(N_{\bar{A}\bar{B}}, N_{\bar{A}C}, N_{\bar{B}C})$

$$upper_{II}(\gamma^{p}_{AB|C}) = \frac{\tilde{N}_{ABC}\tilde{N}_{\bar{A}\bar{B}C}}{(N_{AC} - \tilde{N}_{ABC})(N_{BC} - \tilde{N}_{ABC})}$$

□

3.5 Algorithm Description

The goal of our algorithms is to find significant subgroups (by drug-demographic combination) in which the ADR and a given drug is strongly correlated. A subgroup may be defined as a conjunction of several conditions. Many potential conditions can be considered to group patients such as age, gender, medical history, other medications, and so on. However, several of subgroups are rare and not informative enough. Accordingly, we need to specify more frequent subgroups of patients. We generate patients subgroups using the Association Rule Mining technique as described in subsection 3.5.1. We explain the most straight forward algorithm in subsection 3.5.2, and our developed pruning algorithms in subsection 3.5.3.

3.5.1 Subgroup Generation

In this section, we specify all subgroups or segments of customers with significant frequency in the dataset; accordingly, these subgroups are candidates for substantial segments. To discover all such segments, we apply the Apriori algorithm as illustrated in Algorithm 1. The Apriori algorithm identifies the most frequent subgroups of patients. It generates all combinations of conditions and returns those combinations with supports (the frequency of a subgroup in the dataset) larger than a pre-defined threshold (*minsup*). The output of Algorithm 1 is ζ the most frequent subgroups of patients which serves as the input of the next two pruning algorithms.

Input : $\mathcal{R}$: a database of patient records where each row represents a patient and the columns are treatments $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, symptoms $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$, and segmentation factors $\mathcal{C} = \{C_1, C_2, \ldots, C_P\}$

Parameters: *minsup*: a user-specified minimum frequency threshold.

Output : ζ: all $\{c\}$ groups of patients with sufficient frequency.

```
1  k = 1
2  F_k = {i|i ∈ I ⋀ σ({i}) ≥ N × minsup}
3  while F_k ≠ ∅ do
4  |   k ← k + 1
5  |   C_k ← apriori-gen(F_{k−1})
6  |   foreach transaction t ∈ T do
7  |   |   C_t ← subset(C_k, t)
8  |   |   foreach candidate itemset c ∈ C_t do
9  |   |   |   σ(c) = σ(c) + 1
10 |   |   end
11 |   end
12 |   F_k = {c|c ∈ C_k ⋀ σ(c) ≥ N × minsup}
13 |   ζ = ∪F_k
14 end
```

Algorithm 1: Segment Generation

Input : $\mathcal{R}$: a database of patient records where each row represents a patient and the columns are treatments $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, symptoms $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$, and segmentation factors $\mathcal{C} = \{C_1, C_2, \ldots, C_P\}$.
Parameters: θ: a user-specified minimum correlation threshold.
Output : All $\{A, B|C\}$ patterns that represent localized ADRs.

```
1  ζ ← apriori-gen(C)
2  for a ← 1 to N do
3  |  for b ← 1 to M do
4  |  |  Find N_AB and calculate r_AB
5  |  |  if r_AB ≥ θ then continue
6  |  |  for each c ∈ ζ do
7  |  |  |  Find r_AB|C
8  |  |  |  if r_AB|C ≥ θ then output {A, B|C}
9  |  |  end
10 |  end
11 end
```

Algorithm 2: The Brute-Force Approach

3.5.2 Brute Force Algorithms

The most straight forward algorithm to detect subgroups of patients with a significant correlation between drugs and ADRs is to estimate the correlation between drugs and ADRs for all subgroups. This brute force algorithm requires $N * M * P$ correlation calculations in which N is the number of treatments, M is the number of symptoms, and P is the number of subgroups. In Algorithm 6 the input variable is R, a database of patient records including treatments, symptoms, and segmentation factors. The input parameter is θ, the minimum correlation threshold. The outputs are patterns $\{A, B|C\}$, which indicate drug A, and ADR B are strongly correlated in subgroup C. The algorithm starts with detecting all frequent subgroups of patients by calling Algorithm 1 in line 1. Then it continues with two loops in line 2 and line 3 to iterate on each possible pair of drug and ADR. In this algorithm, we are interested in exploring pairs of drug and ADRs which are not strongly correlated; accordingly, line 4 finds a correlation between drug and ADR. If the correlation is below the threshold (line 4), the algorithm iterates over all subgroups of patients in ζ (line 6), and finds the correlation between drug and ADR in each subgroup (line 7). Then the algorithm checks the correlation (line 8) in subgroups to see if in the subgroup drug and ADR are strongly correlated.

3.5.3 Two-level Pruning Algorithm

In this section, the developed pruning algorithms are described. Algorithm 3 is based on the initial upper bounds (shown in Theorem 3.8, Theorem 3.9, and Theorem 3.10 for γ^p_{AB}, γ^r_{AB}, and γ^o_{AB} respectively) and secondary upper bounds (shown in Theorem 3.11, Theorem 3.12, and Theorem 3.13 for γ^p_{AB}, γ^r_{AB}, and γ^o_{AB} respectively). The input variables in Algorithm 3 are $\mathcal{R}$, a database including patients records. The parameter is θ, the minimum correlation threshold. The outputs are patterns $\{A, B|C\}$ presenting localized ADRs. Algorithm 3 starts with detecting the most frequent subgroups of patients using Algorithm 1 in line 1. Then it continues with two loops in line 2 and line 3 to iterate on each possible pair of drug and ADR. line 4 finds a correlation between drug and ADR. If the correlation is below the threshold (line 5), the algorithm iterates on all subgroups of patients in ζ (line 6), finds an initial upper bound for correlation in subgroup C using Theorem 3.8, Theorem 3.9, and Theorem 3.10 in line 7.

If the initial upper bound stands above the threshold (line 8), we continue to find a tighter upper bound using the secondary bounds Theorem 3.11, Theorem 3.12, and Theorem 3.13 for γ^p_{AB}, γ^r_{AB}, and γ^o_{AB} respectively in line 9. line 10 checks to see whether the tighter upper bound is also above the threshold. In case the second upper bound is larger than the threshold, the algorithm finds the exact correlation between A and B in subgroup C (line 11).

3.5.4 Save and Look up Pruning Algorithm

Pairwise Necessary Conditions

In this section, we specify necessary pairwise conditions for strong associations. The following lemmas introduce necessary conditions for pair frequencies (N_{AC} and N_{BC}) per association measure.

Input : $\mathcal{R}$: a database of patient records where each row represents a patient and the columns are treatments $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, symptoms $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$, and segmentation factors $\mathcal{C} = \{C_1, C_2, \ldots, C_P\}$.
Parameters: θ: a user-specified minimum correlation threshold.
Output : All $\{A, B|C\}$ patterns that represent localized ADRs.

$\zeta \leftarrow$ apriori-gen(C)
for $a \leftarrow 1$ *to* N **do**
  for $b \leftarrow 1$ *to* M **do**
    Find N_{AB} and calculate r_{AB}
    if $r_{AB} \geq \theta$ **then** continue
    for *each* $c \in \zeta$ **do**
      Find the initial upper bound ($upperI(r_{AB|C})$) using initial bounds for the association measure in subsection 3.4.2
      if $upperI(r_{AB|C}) < \theta$ **then** continue
      Find the secondary upper bound ($upperII(r_{AB|C=1})$) for the association measure using secondary bounds in subsection 3.4.3
      if $upperII(r_{AB|C}) \leq \theta$ **then** continue
      Find $r_{AB|C}$
      if $r_{AB|C} \geq \theta$ **then** output $\{A, B|C\}$
    end
  end
end

Algorithm 3: The Pruning Approach

Lemma 3.14 (Pair Wise Bound for γ^p_{AB})**.** *$\gamma^p_{AB|C}$ is significant only if the following two conditions are satisfied:*

$$N_{AC} \geq \frac{\theta^2 N_{BC} N_C}{N_C + (\theta^2 - 1) N_{BC}} \tag{3.30}$$

$$N_{BC} \geq \frac{\theta^2 N_{AC} N_C}{N_C + (\theta^2 - 1) N_{AC}} \tag{3.31}$$

Proof. We know $\gamma^p_{AB|C} \leq \sqrt{\frac{N_{AC}(N_C - N_{BC})}{N_{BC}(N_C - N_{AC})}}$. Accordingly, for a strong association between $\{A, B\}$ in segment C, we have:

$$\sqrt{\frac{N_{AC}(N_C - N_{BC})}{N_{BC}(N_C - N_{AC})}} \geq \theta$$

$$\frac{N_{AC}(N_C - N_{BC})}{N_{BC}(N_C - N_{AC})} \geq \theta^2$$

$$N_{AC} \geq \frac{\theta^2 N_{BC} N_C}{N_C + (\theta^2 - 1) N_{BC}}$$

$$N_{BC} \geq \frac{\theta^2 N_{AC} N_C}{N_C + (\theta^2 - 1) N_{AC}}$$

□

Lemma 3.15 (Pair Wise Bound for γ^r_{AB})**.** *$\gamma^r_{AB|C}$ is significant only if the following conditions are hold:*

$$N_{AC} \leq \frac{N_C \tilde{N}_{ABC}}{\theta N_{BC} + (1 - \theta) \tilde{N}_{ABC}} \tag{3.32}$$

$$N_{BC} \leq \frac{\tilde{N}_{ABC}(N_C + (\theta - 1) N_{AC})}{\theta N_{AC}} \tag{3.33}$$

Proof. Based on Equation 3.5, we have:

$$\gamma^r_{AB} = \frac{N_{ABC} N_{\bar{A}C}}{N_{\bar{A}BC} N_{AC}}$$

A necessary condition for a significant pair is:

$$\frac{N_{ABC}N_{\bar{A}C}}{N_{\bar{A}BC}N_{AC}} \geq \theta$$

Considering the above condition, we can find a necessary condition for N_{AC} and N_{BC} as:

$$\begin{aligned}
&\frac{N_{ABC}N_{\bar{A}C}}{N_{\bar{A}BC}N_{AC}} \geq \theta \\
&N_{ABC}(N_C - N_{AC}) \geq \theta(N_{AC}(N_{BC} - N_{ABC})) \\
&N_{AC} \leq \frac{N_C\tilde{N}_{ABC}}{\theta N_{BC} + (1-\theta)\tilde{N}_{ABC}} \\
&N_{BC} \leq \frac{\tilde{N}_{ABC}(N_C + (\theta-1)N_{AC})}{\theta N_{AC}}
\end{aligned}$$

According to Equation 3.32, a subgroup is not significantly at risk unless N_{AC} is less than $N_C - \theta \times (N_{\bar{A}B} - N_{\bar{A}B\bar{C}})$. However, to find a condition using only marginal and pairwise values, we need to estimate N_{ABC}. To estimate an upper bound for the RHS in Equation 3.32, we replace N_{ABC} with $\tilde{N}_{ABC} = min(N_{AB}, N_{BC}, N_{AC})$. Accordingly, if the condition in Equation 3.32 is not satisfied, we are certain that the subgroup is not significantly at risk of symptom B after treatment A.

Following the similar logic, we can find a necessary condition for N_{BC} as $N_{BC} \leq \frac{\tilde{N}_{ABC}(N_C+(\theta-1)N_{AC})}{\theta N_{AC}}$. □

Lemma 3.16 (Pair Wise Bound for γ^o_{AB})**.** *$\gamma^o_{AB|C}$ is significant only if the following conditions are hold:*

$$N_{AC} \leq \tilde{N}_{ABC}[\frac{\tilde{N}_{\bar{A}\bar{B}C}}{\theta(\tilde{N}_{\bar{A}B} - \tilde{N}_{\bar{A}B\bar{C}})} + 1] \tag{3.34}$$

$$N_{BC} \leq \tilde{N}_{ABC}[\frac{\tilde{N}_{\bar{A}\bar{B}C}}{\theta(\tilde{N}_{A\bar{B}} - \tilde{N}_{A\bar{B}\bar{C}})} + 1] \tag{3.35}$$

Proof. According to Theorem 3.10, we have

$$\gamma^o_{AB|C} \leq \frac{N_{ABC}N_{\bar{A}\bar{B}C}}{N_{A\bar{B}C}N_{\bar{A}BC}}$$

Therefore, the necessary condition for a significant subgroup is:

$$\begin{aligned}
&\frac{N_{ABC}N_{\bar{A}\bar{B}C}}{N_{A\bar{B}C}N_{\bar{A}BC}} \geq \theta \\
&N_{AC} \leq \frac{N_{ABC}N_{\bar{A}\bar{B}C}}{\theta N_{\bar{A}BC}} + N_{ABC} \\
&N_{AC} \leq N_{ABC}[\frac{N_{\bar{A}\bar{B}C}}{\theta N_{\bar{A}BC}} + 1] \\
&N_{AC} \leq N_{ABC}[\frac{N_{\bar{A}\bar{B}C}}{\theta(N_{\bar{A}B} - N_{\bar{A}B\bar{C}})} + 1] \qquad (3.36)
\end{aligned}$$

Equation 3.36 is a necessary condition for N_{AC} in order for C to be a significant subgroup. Moreover, Equation 3.36 is an increasing function of N_{ABC}, $N_{\bar{A}\bar{B}C}$, and $N_{\bar{A}B\bar{C}}$. Accordingly, an upper bound is $\tilde{N}_{ABC}[\frac{\tilde{N}_{\bar{A}\bar{B}C}}{\theta(\tilde{N}_{\bar{A}B} - \tilde{N}_{\bar{A}B\bar{C}})} + 1]$ in which $\tilde{N}_{\bar{A}\bar{B}C} = min(N_{\bar{A}\bar{B}}, N_{\bar{A}C}, N_{\bar{B}C})$, $\tilde{N}_{ABC} = min(N_{AB}, N_{AC}, N_{BC})$, $\tilde{N}_{\bar{A}B\bar{C}} = min(N_{\bar{A}B}, N_{\bar{A}\bar{C}}, N_{B\bar{C}})$.

Following the same logic, we can find the similar condition for N_{BC} as $\tilde{N}_{ABC}[\frac{\tilde{N}_{\bar{A}\bar{B}C}}{\theta(\tilde{N}_{A\bar{B}} - \tilde{N}_{A\bar{B}\bar{C}})} + 1]$.

□

Save and Lookup Algorithm (Phase I)

To make the pruning algorithm more efficient, in this section, we propose a pre-processing step to save all strong pairwise frequencies (N_{AC}, N_{BC}, N_{AB}) in advance. This pre-processing step is shown in Algorithm 4. To detect the strong frequencies, we develop three lemmas based on the initial bounds in subsection 3.4.2. Accordingly, we find necessary conditions for pairwise frequencies (N_{AC}, N_{BC}) for each association measure as shown in Theorem 3.14, Theorem 3.15, and Theorem 3.16. In Algorithm 4, the input variables is $\mathcal{R}$, a database including patients records. The parameter is θ, the minimum correlation threshold. The output is Ω, a table including triplets and strong frequencies in a format of $\{A, B, C, N_{AC}, N_{BC}\}$. Algorithm 4 starts with two loops as shown in line 2 and line 3 to iterate over all treatments A, and symptoms B. Since we are only interested on pairs of treatments and symptoms that are not significant in general. Therefore, the algorithm checks the whether A and B are associated significantly in line 5. The algorithm continues for non significant $\{AB\}$. It iterates over all patients' subgroups in line 6. It checks the necessary condition for N_{AC}, if it holds, then it checks the necessary condition for N_{BC} if the second condition is also satisfied the algorithm saves the triplet and frequencies as $\{A, B, N_{AC}, N_{BC}\}$ in Ω.

Save and Look Up Algorithm (Phase II)

In Algorithm 5, the input variable is $\mathcal{R}$, a database including patients records. The parameter is θ, the minimum correlation threshold. The output is Ω, a table including triplets and strong frequencies in a format of $\{A, B, C, N_{AC}, N_{BC}\}$. Algorithm 5 starts with two loops in line 2 and line 3. If the pair of $\{A, B\}$ is not significant generally, the algorithm continues with the third loop in line 4 to iterate over all patient subgroups. Then, the algorithm searches for the triplet $\{A, B, C\}$ in Ω in line 5. The algorithm continues only if the triplet exists in Ω. Algorithm 5 continues to find the secondary upper bound according to the association measure in line 7. In case the secondary upper bound is larger than the threshold, the algorithm finds the exact correlation between A and B in subgroup C (line 9).

input : R: a database containing records of ADRs;
Parameters: θ: a user-specified minimum correlation threshold.
Output : $\Omega \leftarrow$ a table including pairs and the frequency $(A, B, C, N_{AC}, N_{BC})$.

```
ζ ← apriori-gen(C)
for a ← 1 to N do
    for b ← 1 to M do
        Find N_AB and calculate r_AB
        if r_AB ≥ θ then continue
        for each C ∈ ζ do
            check the pair wise condition for N_AC according to the association measure
             using Theorem 3.14, Theorem 3.15, and Theorem 3.16
            if the condition is hold then
                check the pair wise condition for N_BC according to the association
                 measure using Theorem 3.14, Theorem 3.15, and Theorem 3.16
                if the condition is hold then Save {A, B, C, N_AC, N_BC} to Ω
            end
        end
    end
end
```

Algorithm 4: Save and Lookup Algorithm (Phase I)

input : $\mathcal{R}$: a database containing records of ADRs; Ω: a list of $\{A, B, C\}$ triplets and frequencies N_{AC} and N_{BC}.
Parameters: θ: a user-specified minimum correlation threshold
Output : All $\{A, B|C\}$ patterns that represent localized ADRs.

```
ζ ← apriori-gen(C)
for a ← 1 to N do
    for b ← 1 to M do
        for each C ∈ ζ do
            if the triplet A, B, C not in Ω then continue
            Look up N_AC and N_BC from Ω
            Find the secondary upper bound (upperII(r_{AB|C=1})) for the association
             measure using secondary bounds in subsection 3.4.3
            if upperII(r_{AB|C}) ≤ θ then continue
            Find r_{AB|C}
            if r_{AB|C} ≥ θ then output {A, B|C}
        end
    end
end
```

Algorithm 5: Save and Lookup Algorithm (Phase II)

3.6 Experiments

In this section, the experimental setup and results are summarized. We briefly explained two main datasets in subsection 3.6.1 along with the experimental setup. The effectiveness of the developed algorithms is illustrated by several experiments in subsection 3.6.2. Also, the efficiency of Algorithm 3, and Algorithm 5 are compared with the brute force algorithm in subsection 3.6.3.

3.6.1 Data Description

VAERS

The Vaccine Adverse Event Reporting System (VAERS) is a national early warning system that aims to discover safety problems in U.S. licensed vaccines. The Center for Disease Control and Prevention (CDC), and the U.S. Food and Drug Administration (FDA) co-administer VAERS. VAERS contains reports of adverse events after vaccination. While healthcare professionals and vaccine manufacturers are required to report adverse events, anyone else can also report to VAERS. VAERS datasets can be found in https://vaers.hhs.gov/data/datasets.html. Yearly data files between 1990 to 2018 are available for download.

In our analysis, we used data files of the year 2008. We utilized VAERSDATA, VAERSVAX, and VAERSSYMPTOMS files and joined them by case ID. After joining, we had 28,227 cases involving 57 vaccines and 3336 ADRs.

The demographic information available in the dataset are age, sex, and state. All the continuous and categorical variables are coded into a series of binary indicators. Age is a numeric variable which we cut into the following age groups: $[\leq 18, [18, 34], [35, 50], \geq 50,$ Unknown]. The ratio of missing age values is almost 9% which we code them as Unknown. Sex is a nominal variable including three categories of Male, Female, and unknown without missing values. State is nominal variable coded as two characters of US states. Almost 13% of records has missing state value which we coded them as Unknown.

The distribution of vaccines, the top 100 most frequent ADRs, and demographic variables in the VAERS dataset are shown in 3.1a, 3.1b, and 3.1c respectively. As shown in **??**, vaccine, ADRs, and demographic variables follow long-tail distributions.

Three of most significant demographic features are selected for the experiments including Sex (Female, Male, Unknown), Age (Generation Z, Generation Y, Generation X, Baby boomers, Others), State (including all 50 states in US). These three are selected because they have less missing values and include clear categories. However, other features can be included as well. The distribution of these demographic features is shown in 3.1c.

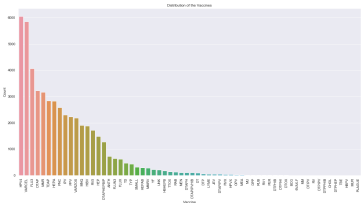
FAERS

FDA Adverse Event Reporting System contains adverse event reports FDA has received from manufacturers, consumers, and healthcare professionals. FAERS datasets can be found in https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html Quarterly data files for each year are available since 2004.

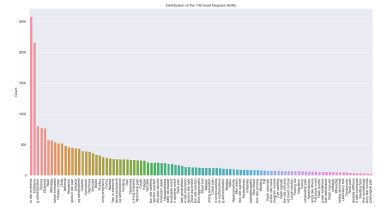
In our analysis, we used data from the second quarter of the year 2018. We used DRUG18Q2, REACT18Q2, INDI18Q2 files, and joined them by Case ID. After joining, our data include 457,170 cases[1] with 37,907 unique drugs and 11,159 unique ADRs.

The demographic information included in the dataset are age, sex, country, and medical conditions. We again code all continuous and categorical variables into a series of binary indications same as VAERS. Age is numeric variable which we cut into the following age groups: $[\leq 18, [18, 34], [35, 50], \geq 50,$ Unknown]. Sex is a nominal variable including

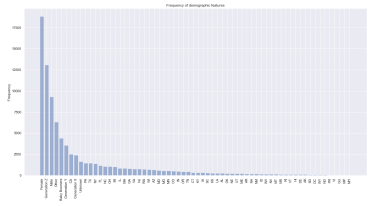
[1] Although potentially each patient could have multiple events/cases reported, here we ignore the longitudinal effect. Instead, we treat each reported case as independent.



(a) Vaccines in VAERS



(b) Frequent ADRs in VAERS



(c) Demographic features in VAERS

Figure 3.1: Correlation vs. Utility.

three categories of Male, Female, and unknown without missing values. Country is nominal variable coded as two characters of 154 countries.

The distribution of drugs, the top 100 most frequent ADRs, and demographic variables in the FAERS dataset are shown in 3.2c, 3.2b, and 3.2a respectively. As shown in **??**, vaccine, ADRs, and demographic variables follow long-tail distributions.

While the VAERS dataset is limited to vaccines in US, FAERS includes ADR reports after taking various drugs in 154 countries.

There are 11159 unique ADRs in the FAESR dataset. 3.2b shows the distribution of 100 of most frequent ADRs.

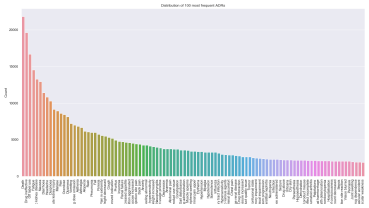
Platform. All of the experiments were accomplished on a MacBook Pro, 2.7 GHz Intel Core i5 and 8 GB of RAM. All of the programs were implemented in Python.

3.6.2 Effectiveness

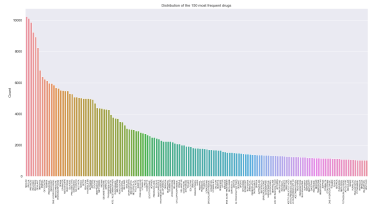
Using Algorithm 1, we detect the most frequent segments of patients considering three specifications, including sex, age, and state in VAERS. Table 3.3 shows ten most frequent segments. Additionally, Table 3.4 presents the ten most frequent segments of patients in FAERS accounting for sex, age, country, and medical condition.

Table 3.5 shows some results of the developed algorithms. It presents drugs or vaccines correlated with ADRs for some patient segments while there are not correlated at global level. This table includes a few results from each dataset. For instance, based on the FAERS data, we found that Death is associated with taking Opdivo in patients living in the US with age range between 35 and 50. Another example result according to the VAERS data, is that female patients experienced more influenza like illness after taking the LYME vaccine.

(a) Demographic features in FAERS



(b) Frequent ADRs in FAERS



(c) Frequent drugs in FAERS

Table 3.3: The most Frequent Subgroups of Patients in VAERS

Gender	Age	State	Subgroup
Male			1
		CA	2
Female			3
		NC	4
	< 18		5
		OH	6
	< 18	CA	7
Male		CA	8
		NY	9
F	¡ 18		10

Table 3.4: The most Frequent Subgroups of Patients in FAERS

Gender	Age	Country	Medical History	Subgroup
Female				1
Male				2
		US		3
	>50			4
	35-50			5
			Atrial fibrillation	6
	>50	US		7
Female	>50	US		8
	35-50	US		9
	>50		Atrial fibrillation	10

Table 3.5: Examples of at-risk patient groups

Drug/Vaccine	Drug Frequency	ADR	ADR Frequency	Patient Group	Group Frequency	Data
VARZOS	2058	Herpes zoster	744	OH	997	VAERS
LYME	51	Activities of daily living impaired	281	Male	12390	VAERS
DT	81	Ocular hyperaemia	158	<18, CA	1203	VAERS
ANTH	684	Neck pain	275	<18, CA	1203	VAERS
RUB	12	Sleep disorder	138	NC	1129	VAERS
NEXIUM	1097	Chronic kidney disease	408	US	26354	FAERS
NEXIUM	1097	End stage renal disease	130	>50	33528	FAERS
PRILOSEC	95	Tubulointerstitial nephritis	123	F, >50, US	21560	FAERS
OXYCONTIN	381	Drug abuse	67	>50, US	22151	FAERS
ACTIQ	5	Sepsis	627	F, >50, US	21560	FAERS
METHOTREXATE	2789	Treatment failure	247	US, 35-50	2849	FAERS
HUMIRA	3165	Abdominal discomfort	612	Atrial fibrillation	1971	FAERS

3.6.3 Efficiency

The number of pairs to compute and save in the pre-processed step of Algorithm 5 (as described in Algorithm 4) and the number of pairs to compute in the first pruning step of Algorithm 3 (as described in line 7 of Algorithm 3), are compared with the number of all pairs in Figure 3.2. As shown in Figure 3.2, the comparison is conducted per correlation measure (Pearson, RR, and Odds ratio) and dataset (VAERS AND FAERS). Figure 3.2 compares the number of pairs for different levels of correlation threshold. The threshold for Pearson correlation changes from 0.1 to 0.9, while it changes from 1 to 10 for Relative Risk and Odds Ratio measurements. There is a significant gap between the number of pairs in Algorithm 4 and the number of pairs in line 7 of Algorithm 3 and all pairs. By increasing the threshold, the gap also increases quickly. Accordingly, Algorithm 4 is memory efficient.

After saving pair calculation in Algorithm 4, and pruning some pairwise calculation in line 7 of Algorithm 3, the performance of the second pruning step of Algorithm 3 in line 7 is almost the same. Figure 3.3 shows a comparison of total running time of Algorithm 5 and the brute force algorithm under different threshold values per correlation measurement and dataset. As shown in Figure 3.3, Algorithm 5 is much faster than brute force algorithm.

3.7 Conclusion

In this paper, we investigated false negatives in large datasets. We identified two sets of upper bounds for pruning and developed our pruning algorithm based on those upper bounds. Additionally, we developed a save and lookup algorithm to increase the efficiency of discovering false negatives. The experimental results on two large datasets (FAERS and VAERS) demonstrate the efficiency and effectiveness of our algorithm.

There are many potential extensions for future study. It will be interesting to extend our methodology into other association measures (other than Pearson correlation, odds ratio, and relative risk.) It will be useful to extend the application of our framework to other areas such as marketing to find insights about purchasing behavior in customer subgroups.

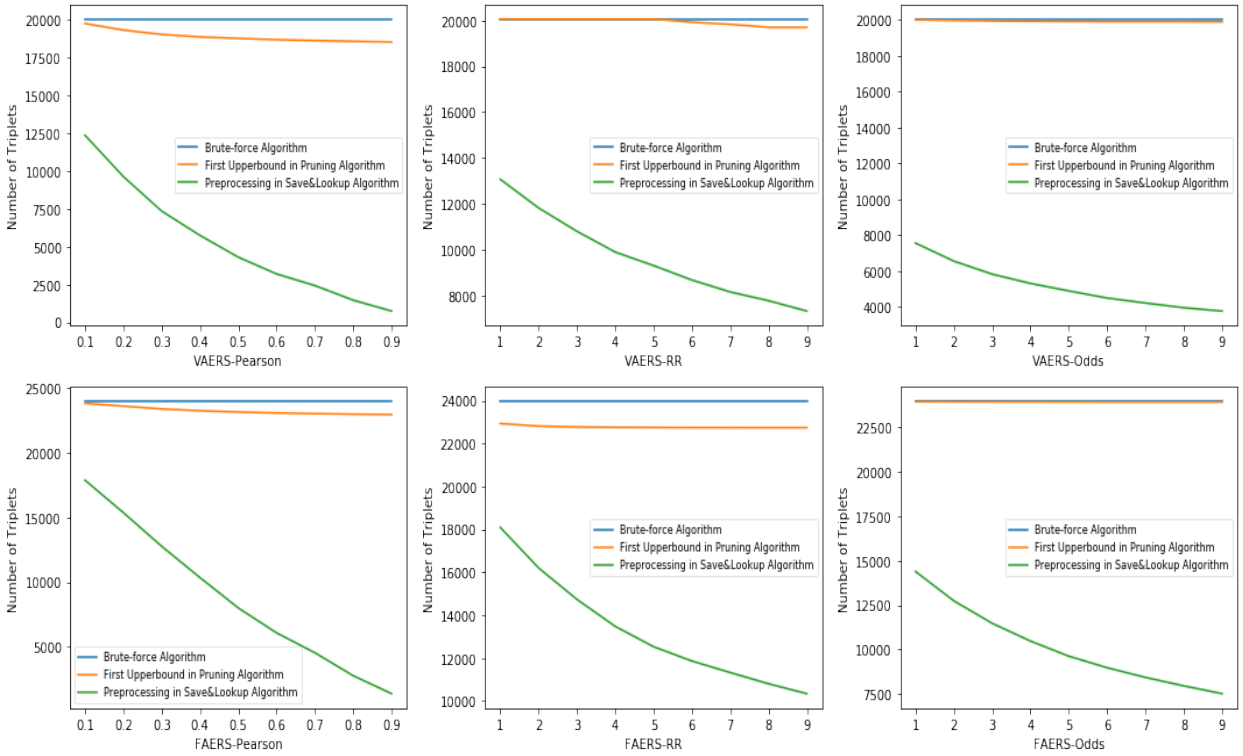


Figure 3.2: Comparison use of space: x-axis: Theta ; y-axis: the number triplet combinations of drug, ADR, and patients group

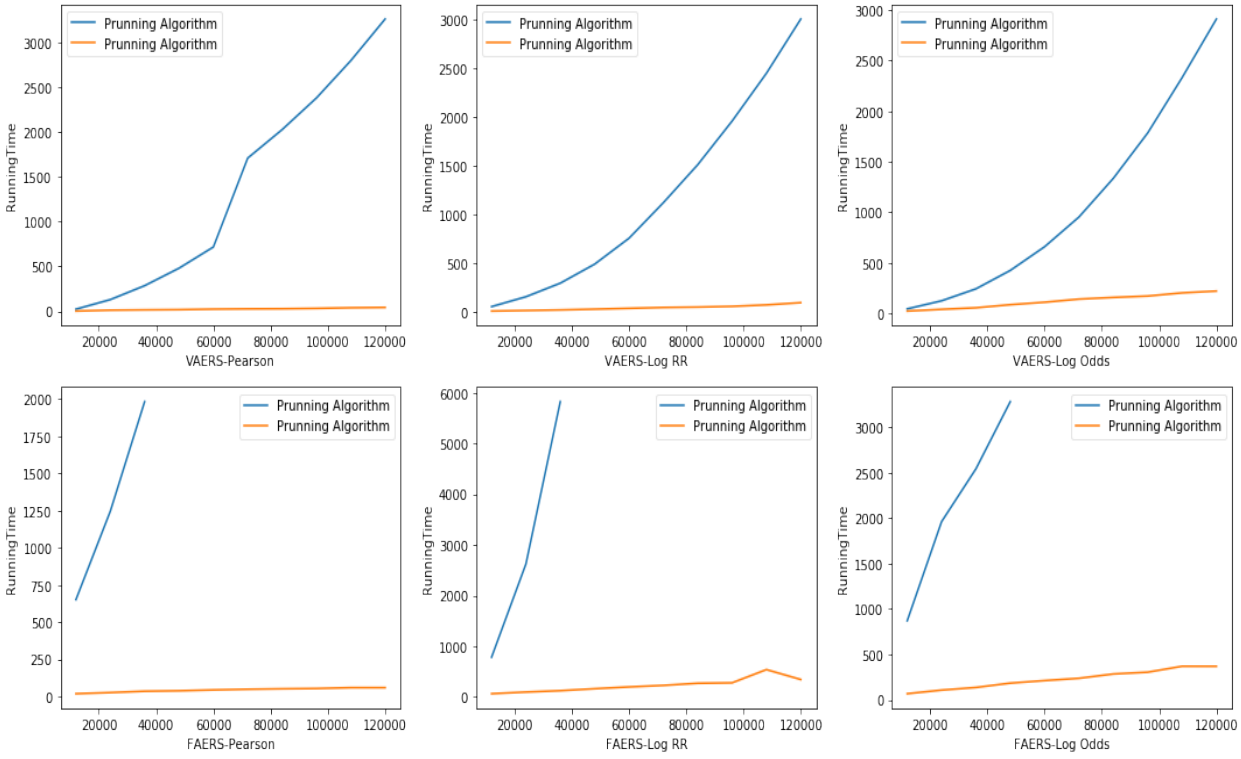


Figure 3.3: Running time comparisons: x-axis: the number triplet combinations of drug, ADR, and patients group; y-axis: Running time

Chapter 4

An Empirical Analysis of Capacity and Flexibility Planning Under Demand Uncertainty

We study the optimal capacity and production planning under stochastic demand accounting for price sensitivity and correlation impacts on demand. We analyze four flexibility strategies in a multi-products setting considering resource flexibility along with at-capacity or below-capacity production. We model each flexibility strategy as a two-stage stochastic programming problem to maximize the expected profit. In the first stage, we determine the capacity of resources. In the second stage, after demand realization, we specify the production quantity of each product group considering the capacity limits defined in the first stage. We illustrate the significant role of demand correlation and price sensitivity in production and capacity planning and flexibility strategy selection. Moreover, we show the impact of correlation, price-sensitivity, price, and demand variation on profit and flexibility strategy selection under different conditions. We emphasize that selecting an appropriate flexibility strategy depends on cost factors as well, and the model should be solved accounting for each manufacturer's parameters such as investment cost, production cost, product correlation, and price.

4.1 Introduction

On one hand, manufacturers have to expand their product portfolio to sustain in a competititve environment [50]. On the other hand, increasing product variety increases demand uncertainty, and capacity decisions are often made under demand uncertainty. This creates misalignment between assigned capacity and realized demand [25]. Manufacturing flexibility is a solution to this demand-supply mismatch. It works on the principle of reallocation or customization of capacity after demand realization [25]. We investigate two types of flexibility in this study: product (process) flexibility and volume flexibility. Product flexibility refers to feasibility to produce multiple products using the same resources or capacity, reallocating capacity wherever necessary. Volume flexibility is the ability to manufacture below capacity.

Product flexibility has been extensively studied in operations research over several decades [29, 24, 8, 50]. These efforts explore different aspects of product flexibility. For instance, Jordan and Graves [29] consider the relation between product flexibility and capacity, and Bish and Chen [8] establish product flexibility for vertically differentiated products. Unlike product flexibility, volume flexibility has not gained much attention in literature. Some analytical volume flexibility studies can be found in economics literature [43, 58] while there are some other studies in operations [32, 14]. There is an even smaller number of studies which investigate the combination of two types of flexibility in one framework [25, 8]. A majority of these studies either do not explore demand correlation effects, or limit their analysis to two-product settings due to the complexity of correlation estimation in multi-product settings.

Managing volume and product flexibility in a multiproduct firm imposes fundamental challenges. In a multiproduct setting, the difficulty of detecting product correlations increases exponentially with the number of products. On the other hand, ignoring product correlations reduces the accuracy of demand forecasts. Volume flexibility has typically been studied in a single product setting in which one product is manufactured in a given capacity. The inclusion of product correlation along with the relationship between two kinds of flexibility (product and volume) complicates the problem further.

In this study, we develop a framework for combined capacity and flexibility planning in a multiproduct setting exploring product correlation which is unprecedented in literature. We demonstrate the utility of our model using an empirical example.

4.2 Literature Review

In this section, we summarize most recent and related previous studies to our research. subsection 4.2.1 outlines papers concentrating on volume flexibility. subsection 4.2.2 describes papers focusing on product flexibility. subsection 4.2.3 includes papers accounting for both volume and product flexibility. subsection 4.2.4 summarizes the gap in the literature, and our contributions.

4.2.1 Volume Flexibility

Volume flexibility is defined as feasibility of increasing or decreasing the installed production capacity [25]. [32] survey volume flexibility for a single service using flexible labor resources and show that labor flexibility must be investigated and applied carefully to maximize profit. [46] compare three lean production methods including production leveling, flexible manufacturing system, and demand-managed milk run in terms of their prerequisites and costs. Their results showed that, while flexible manufacturing imposes more severe prerequisites to the system, it is also the most potent tool in improving volume flexibility. [41] investigate the role of employee skills on manufacturing new product, and volume and mix flexibility. They present the impact of skills such as teamwork spirit and willingness to change on manufacturing flexibility and eventually business performance. Tavaghof-Gigloo et al. [53] address a capacity planning problem considering flexibility instruments. They formulate the problem as MILP problem, and present the cost-benefit effects of flexibility instruments in manufacturing. Karakaya and Bakal [31] explore the impact of order or quantity flexibility on a retailer and a manufacturer and observed improvement in the profit of both parties. Hu et al. [27] present an approach to deal with uncertainty in capacity planning accounting for flexible production. These studies mainly focused on impacts of volume flexibility, and they did not explore the role of product flexibility simultaneously.

4.2.2 Product Flexibility

Product flexibility studies explore investment in dedicated capacity versus flexible capacity. These studies analyze the trade-off between extra cost and flexibility in a monopolist environment [24]. A number of previous studies in this area model a monopolistic firm with n products in two stages. Dedicated and flexible capacity levels are specified in the first stage while the production quantities are determined in the second stage after demand realization [19]. Other studies include the work of [24] who demonstrate the impact of competition on flexibility decision and Boyabatlı and Toktay [9] who work on the effects of imperfect capital market on flexibility decision. Unlike these studies in which prices are external parameters, Chod and Rudi [13] investigate pricing decisions in a two products setting with flexible capacity. Başak and Albayrak [4] adopted Petri nets to model the flexible automotive manufacturing system. These studies limit their scope to aspects of product flexibility.

4.2.3 Multiple Flexibility

Chod et al. [15] analyze product, volume, and new product flexibility in a single framework for a two products setting. Goyal and Netessine [25] investigate volume and product flexibility in a two-product setting under endogenous pricing, considering products correlation. Bish and Chen [8] model the flexibility problem for two vertically differentiated products as a two-stage stochastic programming problem. While these studies account for more than one flexibility, their scope is limited to a two-product setting.

4.2.4 Summary

Previous research has primarily worked on volume and product flexibility separately. Only a limited number of studies explore both types of flexibility. Moreover, there is a dearth of studies in exploring impacts of products correlation on selecting flexibility types and on production planning. Our study contributes to the literature in the following aspects:

1. Demand is stochastic and a function of price and products' supply accounting for price-sensitivity and correlation between products.

2. The following four flexibility strategies are studied:

   (a) Flexible resource and flexibility of below-capacity production (Product and Volume flexibility.)

   (b) Dedicated resources and flexibility of below-capacity production (Volume flexibility.)

   (c) Flexible resource, and at-capacity production (Product flexibility.)

   (d) Dedicated resources, and at-capacity production (Without flexibility.)

3. The impact of two flexibility types is analyzed using a single framework.

4. Flexibility, capacity, and production planning is explored in a multiproduct setting.

5. Demand is estimated more accurately by accounting for correlation between all products.

6. The impact of price-sensitivity, price change, correlation, and demand variation on flexibility, capacity, and production quantity decisions is studied.

4.3 Model

We model the problem for different flexibility settings, including volume and product flexibility in subsection 4.3.2, only volume flexibility in subsection 4.3.3, only product flexibility in subsection 4.3.4, and without flexibility in subsection 4.3.5. We model all four combinations of volume and product flexibility planning as two-stage stochastic problems. The first stage makes long-term investment decisions, which contain resources capacity determination. However, the second stage makes short-term decisions, including production quantity for each product group given available resource capacity.

4.3.1 Preliminary

Problem Statement

Consider a manufacturer that produces n different product families indexed by $i = 1, 2 \dots n$ wherein demand per product is stochastic. We determine the best flexibility plan, capacity level of resources, and production quantities per product family. We formulate the problem as a two-stage stochastic programming model. In the first stage, we detect optimal production capacity level while, after demand realization in the second stage, we determine the production quantities per product family. We study product flexibility and dedicated production strategies, and for each strategy we explore two conditions including production at capacity (i.e. utilize all available capacity) and below capacity.

Demand Function

We assume a demand function of a product is a linear function of its price and of other items supply considering the correlation between items. Equation 4.1 shows the demand function for item i in which α_i is the demand intercept and γ_i is item's price sensitivity, presenting to what extent the demand of an item i is affected by its price. The parameter $\beta_{ij} \in (-1, 1)$ presents the correlation between i and j. The correlation coefficient (β_{ij}) of complement items is positive, while the correlation coefficient is negative for substitute ones, and zero for independent items. Also, $\varepsilon_i(\omega)$ is an additive random error in scenario ω for item i.

$$d_i^\omega = \alpha_i - \gamma_i p_i + \sum_j \beta_{ij} q_j^\omega + \varepsilon_i(\omega) \tag{4.1}$$

Scope and Assumptions

The model relies on the following assumptions:

1. Demand is stochastic function of each item's price and other items' supply corresponding to their correlation. We can estimate price sensitivity and correlation coefficients using historical data.

2. The manufacturer decides the investment planning on the flexible and dedicated resource before the actual demand is known.

3. When the actual demand is realized, the manufacturer determines the production quantity for each product family.

4. We assume there are Ω possible scenarios of demand realization, each happens with probability p_ω and $\sum_{\omega=1}^{\Omega} p_\omega = 1$.

4.3.2 Volume and Product Flexibility Model

In this section, we model the problem for a manufacturer with volume and product flexibility.

In this case, the manufacturer facility has the flexibility to produce all product families. Additionally, the manufacturer can produce below the capacity due to volume flexibility. Although the manufacturer is not obliged to use all available capacity in production, there is a downtime cost associated with unused capacity. Therefore, capacities and production quantities should be specified accurately to avoid unnecessary downtime and production cost.

In the first stage, the flexible capacity (K_f) is determined considering the cost of providing flexible capacity per unit ($Cost_F$). In the second stage, given demand per product family, production quantities are specified per product group.

All notations used in the model are described in Table 4.1.

Stage 1

$$\begin{aligned} \max_{K} \quad & E[\pi^*(K, D)] - Cost_F K_f \\ \text{subject to} \quad & K_f \geq 0 \end{aligned}$$

The objective of the first stage denotes the expected total revenue of production which consists of expected revenue of production corresponding to the second stage of stochastic programming ($E[\pi^*(K, D)]$) and investment cost on flexible resource ($Cost_F K_f$).

The objective function consists of expected production profit given the feasible capacity $E[\pi^*(K, D)]$. The investment cost of a flexible resource with K_f capacity equals $Cost_F K_f$.

Table 4.1: Notations

Notation	Description
Index	
i	Product groups indexed by $i = 1 \dots n$
Ω	a set of potential scenarios $\omega \in \Omega$
Variables	
First stage variables	
K_i	Capacity of dedicated resource for manufacturing item i
K_f	Capacity of flexible resource
Scenario-based variables	
q_i^ω	Production quantity of item i
Z_i^ω	Binary variable equals 1 if $d_i^\omega < K_i^\omega$
d_i^ω	Demand of item i in scenario ω
Parameters	
$Cost_F$	Unit cost of investment on flexible resource
$Cost_i$	Unit cost of investment on item i's dedicated resource
C_{K_f}	Downtime cost of flexible facility
C_{K_d}	Downtime cost of dedicated facilities per unit
Cf_i	Unit cost of manufacturing product i using flexible resource
C_i	Unit cost of manufacturing product i using the dedicated resource
Cf_i	Unit cost of manufacturing product i using flexible resource
C_M	Unit cost of storage extra production
p_i	Unit price of item i
M	A big number

In the first stage, the demand has an additive random element; accordingly, a manufacturer determines the capacity of resource to maximize its expected profit in the second stage ($E[\pi^*(K, D)]$). The constraints ensure that the optimal flexible capacity is more extensive than or equal to zero.

Stage 2

$$\begin{aligned}
\pi^*(K, d(\omega)) = \quad & \max_q \sum_i (p_i - Cf_i) q_i^\omega - C_{K_f}(K_f - \sum_i q_i^\omega) \\
\text{subject to} \quad & q_i^\omega \leq d_i^\omega \quad i = 1 \dots n \quad \forall \omega \\
& \sum_i q_i^\omega \leq K_f \quad i = 1 \dots n, \quad \forall \omega \\
& q_i^\omega \geq 0 \quad i = 1 \dots n, \quad \forall \omega
\end{aligned}$$

In the second stage, the realization of demands d_i^ω per item are observed. The objective is to detect production quantities q_i^ω to maximize the profit under the resource capacity constraints from the first stage. In particular, the objective function in this stage contains two terms. The first term specifies the profit gained ($\sum_i (p_i - C_{f_i}) q_i^\omega$) by manufacturing q_i^ω for $i = 1, \dots n$. The second term expresses the downtime cost which includes the amount of not used capacity (the difference between the available capacity K_f, and the capacity actually used in production $\sum_i q_i^\omega$) multiplied by downtime cost C_{K_f}.

The constraint $q_i^\omega \leq d_i^\omega$ guarantees that production quantities per item do not exceed the actual demand. The next constraint ($q_{id}(\omega) \leq k_f(\omega)$) guarantees the production quantities do not exceed the available capacity. Finally, $q_i^\omega \geq 0$ ensures production quantities are larger than or equal to zero.

4.3.3 Volume Flexibility Model

In this section, we formulate the problem of a multiproduct setting with only volume flexibility strategy in which production below capacity is possible. A manufacturer with volume flexibility strategy is not limited to utilize all capacity in production; however, there is a cost associated with unused capacity. Additionally, there is a specific dedicated resource

per product family. The dedicated resource capacities per product family (K_i) are specified in the first stage.

Stage 1

$$\begin{aligned} \max_{K} \quad & E[\pi^*(K, D)] - \sum_i Cost_i K_i \\ \text{subject to} \quad & K_i \geq 0 \;\; i = 1, \ldots, n. \end{aligned}$$

The objective of the first stage is to specify the optimal dedicated capacity per product family. Other than the fact that this model determines the dedicated capacities instead of the flexible capacity, the rest of the first stage model is the same as the first stage of product and volume flexibility model in section 4.3.2.

Stage 2

$$\begin{aligned} \pi^*(K, d(\omega)) = \quad & \max_{q} \sum_i (p_i - Cf_i) q_i^\omega - C_{K_d} \sum_i (K_i - q_i^\omega) \\ \text{subject to} \quad & q_i^\omega \leq d_i^\omega \quad i = 1 \ldots n \quad \forall \omega \\ & q_i^\omega \leq K_i \quad i = 1 \ldots n \quad \forall \omega \\ & q_i^\omega \geq 0 \quad i = 1 \ldots n, \quad \forall \omega \end{aligned}$$

The second stage model for the volume flexibility strategy is close to the volume and product flexibility strategy. However, instead of flexible capacity, we deal with dedicated resource capacities per product group. Accordingly, the production quantity per item is limited to the dedicated capacity assigned to each item. Additionally, due to volume flexibility, production below capacity level is permitted. The constraint ($q_i^\omega \leq K_i$) ensures that the production quantities per product group are less than the dedicated capacities.

4.3.4 Product Flexibility Model

In this section, we model the problem with only product flexibility. In this case, a firm can assign a flexible production resource capable of manufacturing all items; however, the

manufacturer is required to be at capacity and utilize all the assigned flexible capacity in production. Therefore, the second stage model, in this case, is significantly different from section 4.3.2.

Stage 1

$$\begin{aligned} \max_{K} \quad & E[\pi^*(K, D)] - Cost_F K_f \\ \text{subject to} \quad & K_f \geq 0 \end{aligned}$$

Same as section 4.3.2, we need to determine flexible capacity in this case as well. Accordingly, the first stage does not change.

Stage 2

$$\begin{aligned} \pi^*(K, d(\omega)) = \quad & \max_{q} \sum_i (p_i - Cf_i) q_i^\omega \\ \text{subject to} \quad & q_i^\omega \leq d_i^\omega \quad i = 1 \dots n \quad \forall \omega \\ & \sum_i q_i^\omega = K_f \quad i = 1 \dots n, \quad \forall \omega \\ & q_i^\omega \geq 0 \quad i = 1 \dots n, \quad \forall \omega \end{aligned}$$

The only difference between the second stage here and section 4.3.2 is that in this case production quantities should be equal to the capacity $\sum_i q_i^\omega = K_f$.

4.3.5 Without Flexibility Model

In this section, we model the problem of capacity and production planning when a manufacturer assigns dedicated resources for each product group's production, and needs to manufacture at capacity.

Therefore, a manufacturer is required to determine dedicated capacities per product family in the first stage and utilize all dedicated capacities in the second stage.

Stage 1

This stage is similar to section 4.3.3:

$$\max_{K} \quad E[\pi^*(K, D)] - \sum_i Cost_i K_i$$

$$\text{subject to} \quad K_i \geq 0 \;\; i = 1, \ldots, n.$$

Stage 2

$$\pi^*(K, d(\omega)) = \max_{q} \sum_i p_i d_i^\omega Z_i^\omega + \sum_i p_i K_i (1 - Z_i^\omega) - \sum_i C_i K_i - C_M \sum_i (K_i - d_i^\omega) Z_i^\omega$$

$$\text{subject to} \quad d_i^\omega \leq K_i + M \times (1 - Z_i^\omega)$$

$$K_i(\omega) < d_i^\omega + M \times Z_i^\omega$$

$$Z_i^\omega \{0, 1\} \quad i = 1 \ldots n$$

This model includes a binary variable (Z_i^ω) which is one when $d_i^\omega \leq K_i$ and zero otherwise. To make this constraint linear we used big M technique. Other than Z_i^ω, the second stage model does not contain any decision variables because the production quantity per product family equals to the dedicated capacity resource associated with the product family ($q_i^\omega = K_i$). Accordingly, production cost in this case equals to $\sum_i C_i K_i$. Besides, when $d_i^\omega \leq K_i$ and $Z_i^\omega = 1$ company's sale equals to d_i^ω, and its revenue is $\sum_i p_i d_i Z_i^\omega$ while when $d_i^\omega \geq K_i$ and $Z_i^\omega = 0$, company's sale equals to K_i, and its revenue is $\sum_i p_i K_i (1 - Z_i^\omega)$.

4.4 Empirical Investigation

We demonstrate the functionality of the model using several numerical examples. The solution algorithm is described in subsection 4.4.1. Experimental set-up for problem generations is shown in subsection 4.4.2. Scenario generation is explained in subsection 4.4.3. Results of sensitivity analysis to demonstrate impact of correlation, variation in demand, and price-sensitivity are shown in subsection 4.4.4, subsection 4.4.5, and subsection 4.4.6 respectively.

4.4.1 Solution

The accuracy of the capacity and flexibility planning solution is corresponding to the accuracy of demand estimation model. Hence, the first challenge is to estimate product demand accurately. To resolve this, we develop a multi-variate demand model taking into account price-sensitivity, and correlations. Due to multi-variate setting of the problem, it is more challenging to detect correlation between all products.

To solve a two-stage stochastic integer programming problem we may develop a cutting planes method to define the optimal solution bounds. We model the problem using a two-stage stochastic programming problem with recourse, and selected L-shaped algorithm to solve the problem.

4.4.2 Experimental set-up

To generate the experiment, two correlated products are defined with the principal characteristics shown in Table 4.2. For the common parameters in our study and Bish and Chen [8], such as cost and demand, we use the defined range in Bish and Chen [8]. For the rest, we either explore the total possible range (for instance, we define the correlation range as $[-1, 1]$), or we define a range corresponding to other parameters (such as price range which is specified based on cost range.)

Table 4.2: Parameters' range in numerical examples

Parameters	Respective range
$Cost_F$	$[20, 30]$
$Cost_i$	$[10, 20]$
C_{K_f}	$[1, 5]$
C_{K_d}	$[1, 5]$
Cf_i	$[3, 5]$
C_i	$[1, 3]$
p_i	$[10, 25]$
α_i	100
$\epsilon_i(\omega)$	$[0, 40]$
γ_i	$[0, 1]$
β_{ij}	$[-1, 1]$

4.4.3 Scenario Generation

This study considers demand is a stochastic and continuous variable which is a function of item prices. The stochastic part of demand is additive as shown in Equation 4.1. To define the scenarios considering demand function $d_i^\omega = \alpha_i - \gamma_i q_i^\omega + \sum_j \beta_{ij} q_i^\omega + \varepsilon_i(\omega)$, we specify a normal distribution for $\alpha_i + \varepsilon_i(\omega)$ as Bish and Chen [8].

Our random variable follows a continuous distribution. Accordingly, a reasonable discretization of the continuous distribution is required. We need to define the number of scenarios to ensure the problem is computationally manageable. To this aim, we define scenarios using a Monte-Carlo sampling technique. Using the Sample Average Approximation (SAA) algorithm, we find the optimality gap for different number of scenarios, and select a reasonable number for sensitivity analysis. We follow Shapiro and Philpott [48] to determine an approximate $100(1-\alpha)\%$ lower bound, upper bound, and optimality gap. As illustrated in Table 4.3, we can find the optimal solution even if the number of scenarios are as low as 16. In the following experiments, we define 25 scenarios by taking a random variable from the associated normal distribution.

4.4.4 The impact of correlation

Correlation between products affects a manufacturer's decisions regarding capacity and production level. In this section, we demonstrate to what extend each production strategy is affected by product correlation. As shown in Figure 4.1, increasing the correlation between two products results in a production profit increase for all production flexibility strategies.

Table 4.3: Optimality gap for four cases of flexibility

Confidence level $(1-\alpha)$	0.9		0.95		0.99	
Number of scenarios	16	25	16	25	16	25
Optimality gap						
Volume and product flexibility	0	0	0	0	0	0
Volume flexibility	0	0	0	0	0	0
Product flexibility	0	0	0	0	0	0
Without flexibility	0	0	0	0	0	0

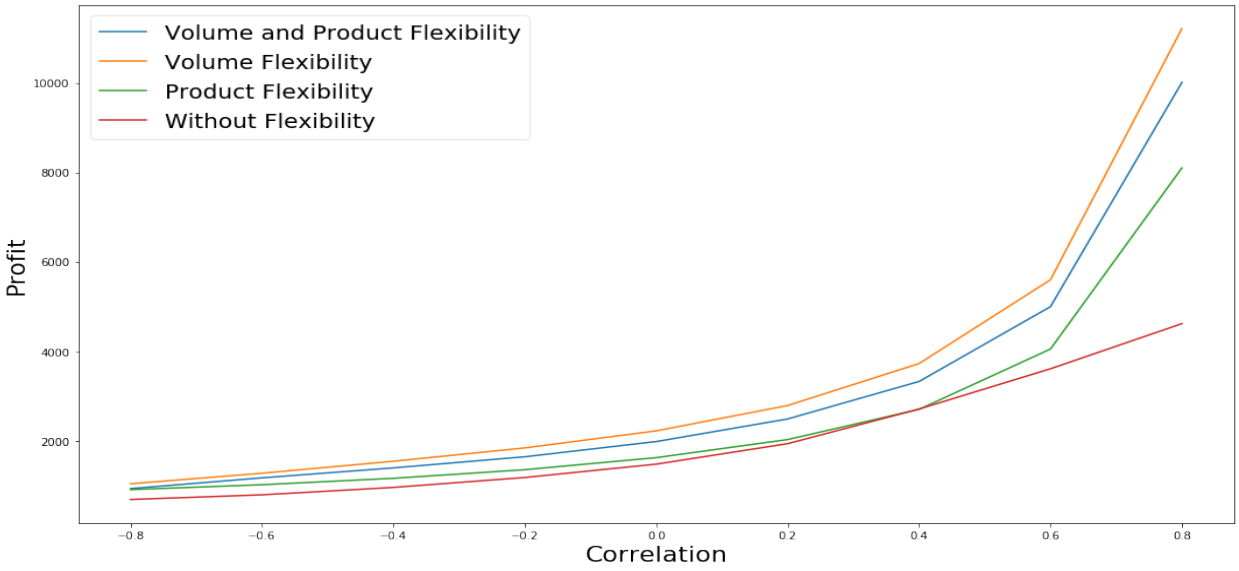


Figure 4.1: Impact of correlation on profit

4.4.5 The impact of variation

Demand variation is another variable which affects a manufacturer's production and capacity decisions, besides profit. Demand function with additive randomness is defined in Equation 4.1. The additive randomness in the demand function follows a normal distribution $Normal(\alpha_i, \epsilon_i(\omega))$ as shown in Table 4.2. Figure 4.2 shows the impact of demand variation on profit of each production flexibility strategy. As expected, a manufacturer without flexibility is more affected by variation while a manufacturer with both volume and product flexibility is barely influenced by increasing variation.

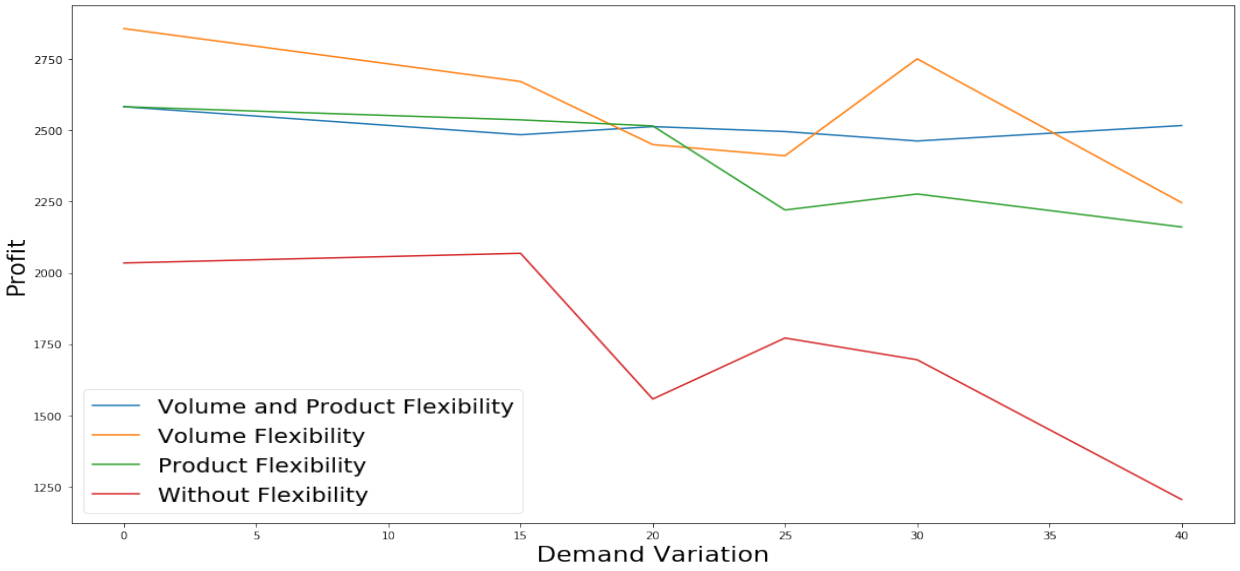


Figure 4.2: Impact of demand variation on profit

4.4.6 The impact of price-sensitivity

As shown in Equation 4.1 demand is negatively affected by price-sensitivity. In other words, higher price-sensitivity results in reduced product demand. Figure 4.3 shows negative impact of price-sensitivity which is almost the same for each production strategy.

4.4.7 The impact of price

Product price is another factor which influences profit. Increasing price increases profit up to a specific level, and then decreases the demand due to demand reduction caused by increased prices. Figure 4.4 shows the impact of price on each production strategy.

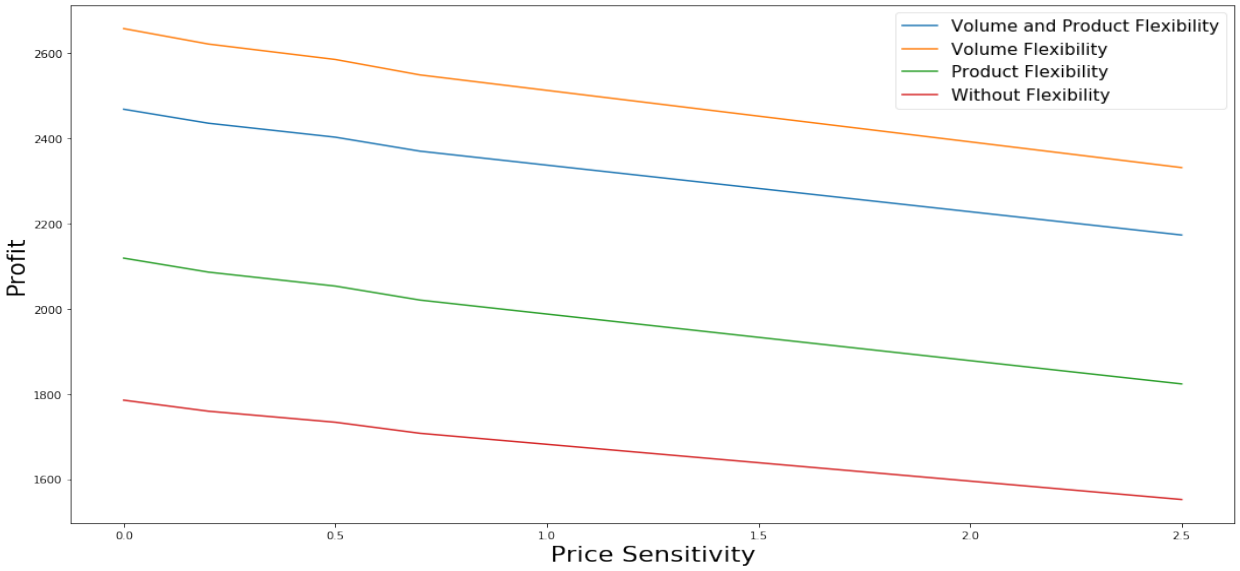


Figure 4.3: Impact of price-sensitivity on profit

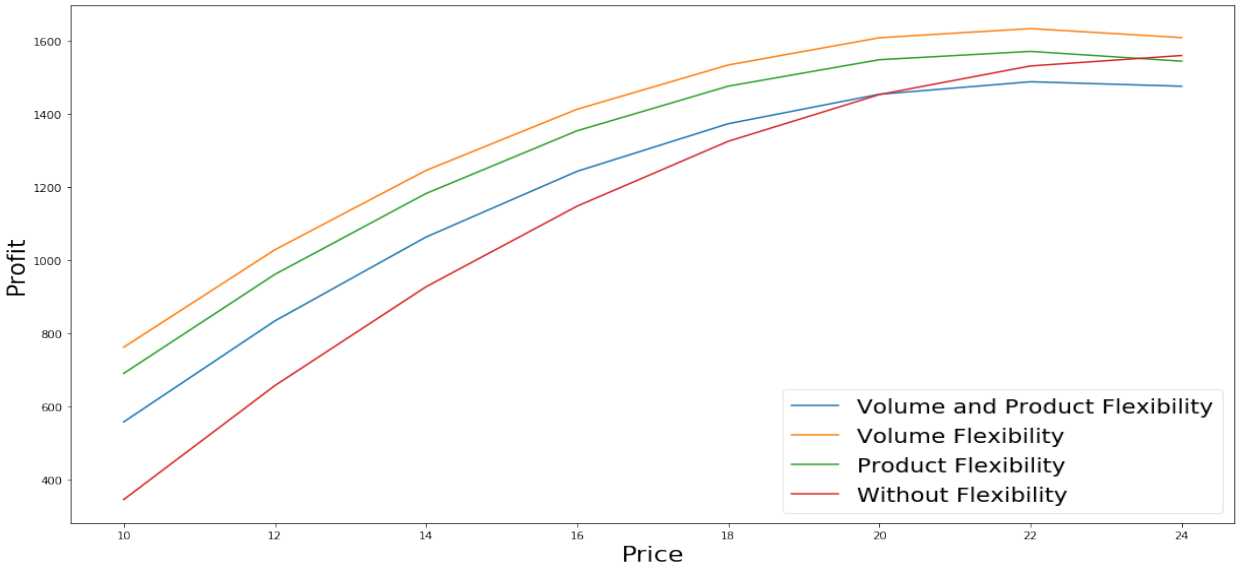


Figure 4.4: Impact of price-change on profit

4.4.8 Managerial Implications

The presented model of flexibility, capacity, and production planning helps managers to determine, in the long-term, the best flexibility and capacity plan to maximize their profit. Additionally, in the short-term, the proposed model specifies the best production plan per product family.

In this section, we analyze flexibility strategies under different conditions of correlation, variation, and price assuming these parameters are independent. To define numerical examples in this section, we use parameter ranges defined in Table 4.2. Besides, low, medium, and high levels of correlation, variation, and price in numerical examples are explained further in Table 4.4. To illustrate the impact of each variable (correlation, variation, and price) we keep other variables at medium levels. Table 4.5 ranks flexibility strategies accounting for low, medium, and high level of correlation, variation, and price. In Table 4.5, V indicates volume flexibility strategy, VP is volume and product flexibility, P is product flexibility, and W is without any flexibility.

As shown in Table 4.5 volume flexibility (V) is the best strategy in many cases and volume and product flexibility VP is a better solution in case of higher demand variation. Besides, production strategy without flexibility W results in the least revenue in most cases.

Table 4.5 provides insights for managerial decisions studying impacts of one variable at a time while keeping the rest of variables at their medium level. However, we can study the impact of two or more factors to determine the best strategy under those conditions. To further consider the interaction between correlation, variation, and price higher values, we analyze two cases:

1. High correlation and variation: when both correlation and variation are significant, the best flexibility strategy is volume and product flexibility (i.e., flexible resources, and production below capacity.)

Table 4.4: Three level parameters range

Correlation			Variation			Price		
Low	Medium	High	Low	Medium	High	Low	Medium	High
$[0, 0.25]$	$[0.25, 0.75]$	$[0.75, 1]$	$[0, 10]$	$[10, 30]$	$[30, 40]$	$[10, 15]$	$[15, 20]$	$[20, 25]$

Table 4.5: Ranking flexibility strategies for different correlation, variation, and price conditions

Rank	Correlation			Variation			Price		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
1	V	V	V	V	VP	VP	V	V	V
2	VP	VP	VP	VP	V	V	VP	VP	VP
3	P	P	P	P	P	P	P	P	W
4	W	W	W	W	W	W	W	W	p

2. High correlation and price: when both correlation and price are significant, our experiments suggest volume and product flexibility in this case as well.

Other than correlation, variation, and price, cost factors also impact flexibility strategy determination. Table 4.6 specifies the best strategies at different investment, production, and downtime cost levels. Difference between investment in one unit of product flexible resource and the average cost of investment on dedicated resources ($Cost_f - \sum_{i=1}^{n} Cost_i/n$) is a primary factor impacting the flexibility decision. The first column in Table 4.6 indicates the investment cost difference between flexible and dedicated resources. Investing in a flexible product resource is usually more expensive than investing in a dedicated resource. The difference between flexible and dedicated investment cost plays a critical role in determining whether a company requires product flexibility. A manufacturer's decision is a trade-off between the extra investment cost and increasing revenue caused by a flexible resource. When the difference between investment cost of flexible and dedicated resources is not significant, selecting flexible strategy provides an opportunity to handle demand uncertainty without imposing a considerable cost to the system comparing to a dedicated resource. The second important cost factor is production cost using product flexible and dedicated resource. Similar to the investment cost, in cases of a small difference between flexible and dedicated production cost, a flexible strategy is more favorable. The third major cost factor is the downtime cost. Downtime cost happens when a production line does not utilize its full capacity. Accordingly, downtime cost occurs in the case of production below capacity or volume flexibility. When downtime cost is zero or low, volume flexibility results in higher profit.

Table 4.6: Impact of Cost on Flexibility Selection

Flexible and Dedicated Investment Cost Difference	Flexible and Dedicated Production Cost Difference	Downtime Cost	Strategy
Low	Low	Low	VP
		Medium	VP
		High	P
	Medium	Low	VP
		Medium	VP
		High	P
	High	Low	VP
		Medium	VP
		High	P
Medium	Low	Low	VP
		Medium	VP
		High	P
	Medium	Low	VP
		Medium	VP
		High	P
	High	Low	V
		Medium	V
		High	W
High	Low	Low	VP
		Medium	VP
		High	P
	Medium	Low	V
		Medium	V
		High	W
	High	Low	V
		Medium	V
		High	W

4.4.9 Discussion and Future Work

This paper presents a model to address long-term and short-term production decisions in a single framework. An optimization model is developed to specify capacity and production quantities for different flexibility strategies. We compare flexibility strategies under various conditions. We find that the volume flexibility strategy outperforms other strategies in most cases, except in higher demand variation in which product and volume flexibility strategy is a better choice. Additionally, we show that cost factors also affect flexibility decisions. Accordingly, flexibility selection and capacity and production planning are sensitive to specific conditions of a manufacturer. To the best of our knowledge, this is the first study accounting for products' correlation, product and volume flexibility, and demand uncertainty in a multiproduct setting.

There are several potential extensions for our models. In this study, we assume the stochastic part of demand follows a normal distribution. Future studies can explore the impacts of other distribution functions as well. Additionally, we focus on stochastic demand and do not consider stochastic supply. Introducing supply uncertainty is another potential extension of this study. Moreover, by increasing the number of items, the number of demand scenarios increases exponentially; therefore, developing efficient methods for scenario reductions is necessary for future studies.

Chapter 5

Conclusions

In this dissertation, we address challenging problems in marketing chapter 2, healthcare chapter 3, and manufacturing chapter 4 accounting for correlation. Although considering correlation in already complex problems of product bundling, ADR detection, and flexibility and capacity planning makes it even more perplex, the critical role of correlation in increasing the accuracy of solutions is unavoidable.

In chapter 2, we look at the product bundling problem from a data-driven perspective. To discover the bundle of products that maximize the utility of bundling for a retailer with a wide range of products, we first develop a probabilistic utility model. The model accounts for correlation, price-sensitivity, and co-purchasing probability. To solve the utility model for all potential bundles in massive problems, we need to make the problem scalable. To this aim, we find the conditions under which the bundling is not profitable mathematically. Utilizing the proven terms, we develop an algorithm (Profitable Bundle Query (PBQ) algorithm) to solve large-scale problems. PBQ can solve a problem of 2,500 products in less than an hour while the brute force algorithm takes much longer to solve a problem of 1,500 products. Moreover, PBQ outperforms the association rule mining algorithm in detecting itemsets of two elements with a higher bundle fraction ratio and average sale. Additionally, we extend the utility model to solve Buy One Get One offers as well. Finally, we point out the importance of fine-grained bundle discovery, meaning bundle discovery at the UPC level and per store, which emphasizes the role of algorithms such as PBQ in bundling problems.

In chapter 3, we point out that detecting the correlation between drugs and ADRs is not sufficient. By limiting the correlation calculation into the general level, we lose critical information about the association between drugs and ADRs in different patient subgroups. However, accounting for patient subgroups makes the problem even more complicated for which we require a new efficient algorithm. To find patient subgroups, we consider demographic features (such as gender, age, and location) and medical history features)such as other diseases and medications). The combinations of all features result in a broad set of patient subgroups. Holding all combinations is not practical because several of these combinations may rarely occur in reality. Accordingly, to discover the most frequent patient subgroups, we apply the association rule mining algorithm. We find the patient subgroups who are at elevated risk of ADRs after using drugs. To make large problems scalable, we develop two pruning algorithms. Then, we show their efficiency by comparing their associate space usage and running time with the brute force algorithm.

Lastly, in chapter 4, we focus on the role of correlation in manufacturing decisions under demand uncertainty. Considering correlation increases the accuracy of demand prediction; accordingly, it reduces the misalignment between assigned capacity and realized demand. Accurate demand prediction, along with manufacturing flexibility provides a robust solution for the demand-supply mismatch. We model long-term and short-term manufacturing decisions in one framework. A manufacturer needs to specify the flexibility and capacity of resources in the long-term while he/she needs to detect the production quantity per family product in each production period (short-term). We model the problem as a two-stage stochastic problem for four flexibility strategies. We define several numerical examples and compare four flexibility strategies under various conditions of correlation, demand variation, price change, and price sensitivity. Our results showed that the volume flexibility strategy outperforms other strategies in most cases. However, when demand variation is significant, the product and volume flexibility strategy works better than volume flexibility. Additionally, cost factors affect flexibility selection as well. Accordingly, the problem should be solved under the specific conditions of each manufacturer, including investment and production costs.

Bibliography

[1] Adams, W. J. and Yellen, J. L. (1976). Commodity bundling and the burden of monopoly. *The Quarterly Journal of Economics*, pages 475–498. 5, 11

[2] Arah, O. A., Chiba, Y., and Greenland, S. (2008). Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Annals of epidemiology*, 18(8):637–646. 2, 34

[3] Bakos, Y. and Brynjolfsson, E. (1999). Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 45(12):1613–1630. 5

[4] Başak, Ö. and Albayrak, Y. E. (2015). Petri net based decision system modeling in real-time scheduling and control of flexible automotive manufacturing systems. *Computers & Industrial Engineering*, 86:116–126. 67

[5] Beladev, M., Rokach, L., and Shapira, B. (2016). Recommender systems for product bundling. *Knowledge-Based Systems*, 111:193–206. 5, 8

[6] Benisch, M. and Sandholm, T. (2011). A framework for automated bundling and pricing using purchase data. In *International Conference on Auctions, Market Mechanisms and Their Applications*, pages 40–52. Springer. 9

[7] Bhargava, H. K. (2013). Mixed bundling of two independently valued goods. *Management Science*, 59(9):2170–2185. 5, 8, 23

[8] Bish, E. K. and Chen, W. (2016). The optimal resource portfolio under consumer choice and demand risk for vertically differentiated products. *Journal of the Operational Research Society*, 67(1):87–97. 2, 65, 67, 76, 77

[9] Boyabatlı, O. and Toktay, L. B. (2011). Stochastic capacity investment and flexible vs. dedicated technology choice in imperfect capital markets. *Management Science*, 57(12):2163–2179. 67

[10] Bulut, Z., Gürler, Ü., and Şen, A. (2009). Bundle pricing of inventories with stochastic demand. *European Journal of Operational Research*, 197(3):897–911. 8, 11

[11] Chakravarty, A., Mild, A., and Taudes, A. (2013). Bundling decisions in supply chains. *European Journal of Operational Research*, 231(3):617–630. 11

[12] Chao, Y. and Derdenger, T. (2013). Mixed bundling in two-sided markets in the presence of installed base effects. *Management Science*, 59(8):1904–1926. 1, 5, 9

[13] Chod, J. and Rudi, N. (2005). Resource flexibility with responsive pricing. *Operations Research*, 53(3):532–548. 2, 67

[14] Chod, J., Rudi, N., and Van Mieghem, J. A. (2010). Operational flexibility and financial hedging: Complements or substitutes? *Management Science*, 56(6):1030–1045. 65

[15] Chod, J., Rudi, N., and Van Mieghem, J. A. (2012). Mix, time, and volume flexibility: Valuation and corporate diversification. *Review of Business and Economic Literature*, 57(3). 67

[16] Danaher, B., Huang, Y., Smith, M. D., and Telang, R. (2014). An empirical analysis of digital music bundling strategies. *Management Science*, 60(6):1413–1433. 1, 5, 9

[17] Duan, L., Street, W. N., Liu, Y., Xu, S., and Wu, B. (2014). Selecting the right correlation measure for binary data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(2):13. 2, 34

[18] Ferreira, K. J., Lee, B. H. A., and Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88. 9

[19] Fine, C. H. and Freund, R. M. (1990). Optimal investment in product-flexible manufacturing capacity. *Management Science*, 36(4):449–466. 67

[20] Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9. 34

[21] Ghoshal, A., Menon, S., and Sarkar, S. (2015). Recommendations using information from multiple association rules: A probabilistic approach. *Information Systems Research*, 26(3):532–551. 8

[22] Ghoshal, A. and Sarkar, S. (2014). Association rules for recommendations with multiple items. *INFORMS Journal on Computing*, 26(3):433–448. 8

[23] Goh, K. H. and Bockstedt, J. C. (2013). The framing effects of multipart pricing on consumer purchasing behavior of customized information good bundles. *Information Systems Research*, 24(2):334–351. 5, 9

[24] Goyal, M. and Netessine, S. (2007). Strategic technology choice and capacity investment under demand uncertainty. *Management Science*, 53(2):192–207. 65, 67

[25] Goyal, M. and Netessine, S. (2011). Volume flexibility, product flexibility, or both: The role of demand correlation and product substitution. *Manufacturing & service operations management*, 13(2):180–193. 2, 65, 66, 67

[26] Guthrie, B., Makubate, B., Hernandez-Santiago, V., and Dreischulte, T. (2015). The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010. *BMC medicine*, 13(1):74. 37

[27] Hu, J., Guo, P., and Poh, K. L. (2018). Flexible capacity planning for engineering systems based on decision rules and differential evolution. *Computers & Industrial Engineering*, 123:254–262. 66

[28] Hwang, S.-Y. and Yang, W.-S. (2008). Discovering generalized profile-association rules for the targeted advertising of new products. *INFORMS Journal on Computing*, 20(1):34–45. 7

[29] Jordan, W. C. and Graves, S. C. (1995). Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594. 65

[30] Juntti-Patinen, L. and Neuvonen, P. (2002). Drug-related deaths in a university central hospital. *European journal of clinical pharmacology*, 58(7):479–482. 33

[31] Karakaya, S. and Bakal, I. S. (2013). Joint quantity flexibility for multiple products in a decentralized supply chain. *Computers & Industrial Engineering*, 64(2):696–707. 66

[32] Kesavan, S., Staats, B. R., and Gilland, W. (2014). Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science*, 60(8):1884–1906. 65, 66

[33] Köhler, G., Bode-Böger, S., Busse, R., Hoopmann, M., Welte, T., and Böger, R. (2000). Drug-drug interactions in medical patients: effects of in-hospital treatment and relation to multiple drug use. *International journal of clinical pharmacology and therapeutics*, 38(11):504–513. 37

[34] Lee, D., Park, S.-H., and Moon, S. (2013). Utility-based association rule mining: A marketing solution for cross-selling. *Expert Systems with Applications*, 40(7):2715–2725. 7

[35] Li, Y., Ryan, P. B., Wei, Y., and Friedman, C. (2015). A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug safety*, 38(10):895–908. 34

[36] Li, Y., Salmasian, H., Vilar, S., Chase, H., Friedman, C., and Wei, Y. (2014). A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):308–314. 2, 34, 36, 37

[37] Liu, M., McPeek Hinz, E. R., Matheny, M. E., Denny, J. C., Schildcrout, J. S., Miller, R. A., and Xu, H. (2012). Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association*, 20(3):420–426. 33, 34, 36

[38] Ma, M. and Mallik, S. (2017). Bundling of vertically differentiated products in a supply chain. *Decision Sciences*, 48(4):625–656. 5

[39] Mai, T., Vo, B., and Nguyen, L. T. (2017). A lattice-based approach for mining high utility association rules. *Information Sciences*, 399:81–97. 5, 7

[40] McAfee, R. P., McMillan, J., and Whinston, M. D. (1989). Multiproduct monopoly, commodity bundling, and correlation of values. *The Quarterly Journal of Economics*, 104(2):371–383. 11

[41] Mendes, L. and Machado, J. (2015). Employees skills, manufacturing flexibility and performance: A structural equation modelling applied to the automotive industry. *International Journal of Production Research*, 53(13):4087–4101. 66

[42] Moghaddass, R., Rudin, C., and Madigan, D. (2016). The factorized self-controlled case series method: An approach for estimating the effects of many drugs on many outcomes. *Journal of Machine Learning Research*, 17(185):1–24. 37

[43] Oi, W. Y. (1961). The desirability of price instability under perfect competition. *Econometrica: journal of the Econometric Society*, pages 58–64. 65

[44] Organization, W. H. et al. (1969). *International drug monitoring: the role of the hospital: Report of a WHO meeting*. WHO. 33

[45] Prasad, A., Venkatesh, R., and Mahajan, V. (2010). Optimal bundling of technological products with network externality. *Management Science*, 56(12):2224–2236. 11

[46] Roessler, M. P., Wiegel, F., Abele, E., and Metternich, J. (2017). Simulation-based assessment of lean production methods: Approaches to increase volume and variant flexibility. In *Dynamic and Seamless Integration of Production, Logistics and Traffic*, pages 83–104. Springer. 66

[47] Sahoo, J., Das, A. K., and Goswami, A. (2015). An efficient approach for mining association rules from high utility itemsets. *Expert Systems with Applications*, 42(13):5754–5778. 1, 5, 7

[48] Shapiro, A. and Philpott, A. (2007). A tutorial on stochastic programming. *Manuscript. Available at www2. isye. gatech. edu/ashapiro/publications. html*, 17. 77

[49] Shugan, S. M., Moon, J., Shi, Q., and Kumar, N. S. (2016). Product line bundling: Why airlines bundle high-end while hotels bundle low-end. *Marketing Science*, 36(1):124–139. 5

[50] Simchi-Levi, D. and Wei, Y. (2015). Worst-case analysis of process flexibility designs. *Operations Research*, 63(1):166–185. 65

[51] Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313. 2, 34

[52] Tatonetti, N. P., Patrick, P. Y., Daneshjou, R., and Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31. 36, 37

[53] Tavaghof-Gigloo, D., Minner, S., and Silbermayr, L. (2016). Mixed integer linear programming formulation for flexibility instruments in capacity planning problems. *Computers & Industrial Engineering*, 97:101–110. 66

[54] Tseng, V. S., Wu, C.-W., Fournier-Viger, P., and Philip, S. Y. (2016). Efficient algorithms for mining top-k high utility itemsets. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):54–67. 1, 5, 7

[55] VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *Annals of statistics*, 41(1):196. 37

[56] Varallo, F. R., Dagli-Hernandez, C., Pagotto, C., de Nadai, T. R., Herdeiro, M. T., and de Carvalho Mastroianni, P. (2017). Confounding variables and the performance of triggers in detecting unreported adverse drug reactions. *Clinical therapeutics*, 39(4):686–696. 34, 36

[57] Venkatesh, R. and Kamakura, W. (2003). Optimal bundling and pricing under a monopoly: Contrasting complements and substitutes from independently valued products. *The Journal of Business*, 76(2):211–231. 11

[58] Vives, X. (1989). Technological competition, uncertainty, and oligopoly. *Journal of Economic Theory*, 48(2):386–415. 65

[59] Wan, M., Wang, D., Liu, J., Bennett, P., and McAuley, J. (2018). Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1133–1142. ACM. 8

[60] Xiong, H., Shekhar, S., Tan, P.-M., and Kumar, V. (2006). TAPER: A two-step approach for all-strong-pairs correlation query in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):493–508. 19, 20, 103

[61] Xiong, H., Zhou, W., Brodie, M., and Ma, S. (2008). Top-k φ correlation computation. *INFORMS Journal on Computing*, 20(4):539–552. 34

[62] Xue, Z., Wang, Z., and Ettl, M. (2015). Pricing personalized bundles: A new approach and an empirical study. *Manufacturing & Service Operations Management*, 18(1):51–68. 9, 21

[63] Yang, B. and Ng, C. (2010). Pricing problem in wireless telecommunication product and service bundling. *European Journal of Operational Research*, 207(1):473–480. 5

[64] Yang, T.-C. and Lai, H. (2007). Comparison of product bundling strategies on different online shopping behaviors. *Electronic Commerce Research and Applications*, 5(4):295–304. 7

[65] Yang, Y., Liu, H., and Cai, Y. (2013). Discovery of online shopping patterns across websites. *INFORMS Journal on Computing*, 25(1):161–176. 1, 7, 23

[66] Zhou, W., Xiong, H., Duan, L., Xiao, K., and Mee, R. (2018). Paradoxical correlation pattern mining. *IEEE Transactions on Knowledge and Data Engineering*. 34

Appendices

A Summary of Equations

A.1 Proof of Theorem 2.2

Equation 2.13 is the difference of utility after bundling (which is the summation of each segment utility $U_{AC} = U_I + U_{II} + U_{III} + U_{IV}$) and utility before bundling.

Before bundling, the utility is

$$U_{before} = N_A \cdot \pi_A^0 + N_C \cdot \pi_C^0; \tag{1}$$

and after bundling, by combining all segments described in subsection 2.4.2, we have

$$U_{after} \quad = \quad U_I + U_{II} + U_{III} + U_{IV}, \tag{2}$$

where U_I, U_{II}, U_{III}, and U_{IV} maybe found in Equation 2.8, Equation 2.9, Equation 2.10, and Equation 2.12, respectively. The change in utility after bundling would be

$$
\begin{aligned}
\Delta U_{AC} &= U_{after} - U_{before} \\
&= N_{AC} \cdot \pi_{AC}^{d} + N_{A\bar{C}} \cdot [(\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{C}^{d} + \pi_{A}^{0}] + N_{\bar{A}C} \cdot [(\mathbb{P}_{AC}^{d,0} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{A}^{d} + \pi_{C}^{0}] \\
&\quad + N_{\bar{A}\bar{C}} \cdot (\mathbb{P}_{AC}^{x^*,x^*-d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{AC}^{d} - N_A \cdot \pi_A^0 - N_C \cdot \pi_C^0 \\
&= N_{AC} \cdot \pi_{AC}^{d} + N_{A\bar{C}} \cdot (\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{C}^{d} + N_{A\bar{C}} \cdot \pi_{A}^{0} + N_{\bar{A}C} \cdot (\mathbb{P}_{AC}^{d,0} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{A}^{d} \\
&\quad + N_{\bar{A}C} \cdot \pi_{C}^{0} + N_{\bar{A}\bar{C}} \cdot (\mathbb{P}_{AC}^{x^*,d-x^*} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{AC}^{d} - N_A \cdot \pi_A^0 - N_C \cdot \pi_C^0 \\
&= N_{A\bar{C}} \cdot (\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{C}^{d} + N_{\bar{A}C} \cdot (\mathbb{P}_{AC}^{d,0} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{A}^{d} + N_{\bar{A}\bar{C}} \cdot (\mathbb{P}_{AC}^{x^*,d-x^*} \\
&\quad - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{AC}^{d} + (N_{AC} - N_A + N_{A\bar{C}}) \cdot \pi_{A}^{0} + (N_{AC} - N_C + N_{\bar{A}C}) \cdot \pi_{C}^{0} - N_{AC} \cdot d \\
&= N_{A\bar{C}} \cdot (\mathbb{P}_{AC}^{0,d} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{C}^{d} + N_{\bar{A}C} \cdot (\mathbb{P}_{AC}^{d,0} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{A}^{d} \\
&\quad + N_{\bar{A}\bar{C}} \cdot (\mathbb{P}_{AC}^{x^*,d-x^*} - \mathbb{P}_{AC}^{0,0}) \cdot \pi_{AC}^{d} - N_{AC} \cdot d \\
&= N_{A\bar{C}} \cdot \pi_{C}^{d} \cdot \mathbb{P}_{AC}^{0,d} - N_{A\bar{C}} \cdot \pi_{C}^{d} \cdot \mathbb{P}_{AC}^{0,0} + N_{\bar{A}C} \cdot \pi_{A}^{d} \cdot \mathbb{P}_{AC}^{d,0} - N_{\bar{A}C} \cdot \pi_{A}^{d} \cdot \mathbb{P}_{AC}^{0,0} \\
&\quad + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} \cdot \mathbb{P}_{AC}^{x^*,d-x^*} - N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} \cdot \mathbb{P}_{AC}^{0,0} - N \cdot \mathbb{P}_{AC}^{0,0} \cdot d \\
&= N_{A\bar{C}} \cdot \pi_{C}^{d} \cdot \mathbb{P}_{AC}^{0,d} + N_{\bar{A}C} \cdot \pi_{A}^{d} \cdot \mathbb{P}_{AC}^{d,0} + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} \cdot \mathbb{P}_{AC}^{x^*,d-x^*} \\
&\quad - [N_{A\bar{C}} \cdot \pi_{C}^{d} + N_{\bar{A}C} \cdot \pi_{A}^{d} + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} + N \cdot d] \cdot \mathbb{P}_{AC}^{0,0} \\
&= N_{A\bar{C}} \cdot \pi_{C}^{d} \cdot \mathbb{P}_{AC}^{0,d} + N_{\bar{A}C} \cdot \pi_{A}^{d} \cdot \mathbb{P}_{AC}^{d,0} + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} \cdot \mathbb{P}_{AC}^{x^*,d-x^*} \\
&\quad - \{[N_{\bar{A}C} + N_{\bar{A}\bar{C}}] \cdot \pi_{A}^{0} + [N_{A\bar{C}} + N_{\bar{A}\bar{C}}] \cdot \pi_{C}^{0} + [N - N_{A\bar{C}} - N_{\bar{A}C} - N_{\bar{A}\bar{C}}] \cdot d\} \mathbb{P}_{AC}^{0,0} \\
&= N_{A\bar{C}} \cdot \pi_{C}^{d} \cdot \mathbb{P}_{AC}^{0,d} + N_{\bar{A}C} \cdot \pi_{A}^{d} \cdot \mathbb{P}_{AC}^{d,0} + N_{\bar{A}\bar{C}} \cdot \pi_{AC}^{d} \cdot \mathbb{P}_{AC}^{x^*,d-x^*} \\
&\quad - [N_{\bar{A}} \cdot \pi_{A}^{0} + N_{\bar{C}} \cdot \pi_{C}^{0} + N_{AC} \cdot d] \cdot \mathbb{P}_{AC}^{0,0}
\end{aligned}
$$

for which the most tricky parts are the join probabilities $\mathbb{P}_{AC}^{0,0}$, $\mathbb{P}_{AC}^{0,d}$, $\mathbb{P}_{AC}^{d,0}$, and $\mathbb{P}_{AC}^{x^*,d-x^*}$.

A.2 Proof of Theorem 2.3

For simplicity of notations, let

$$\mathbb{S}_A^{x^*} = \sqrt{\mathbb{P}_A^{x^*}\left[1 - \mathbb{P}_A^{x^*}\right]} \tag{3}$$

Then the joint probability in Equation 2.7 becomes

$$\mathbb{P}_{AC}^{x^*,d-x^*} = \phi_{AC} \cdot \mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} + \mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} \tag{4}$$

If we replace all joint probabilities in Equation 2.13 by Equation 4, and all N_{AC} values by its equivalence $N \cdot \mathbb{P}_{AC}^{0,0}$, we have

$$\begin{aligned}
\Delta U_{AC} \;=\; & (N_A)\pi_C^d[\phi_{AC} \cdot \mathbb{S}_A^0\mathbb{S}_C^d + \mathbb{P}_A^0\mathbb{P}_C^d] + (N_A)\pi_A^d[\phi_{AC} \cdot \mathbb{S}_A^d\mathbb{S}_C^0 + \mathbb{P}_A^d\mathbb{P}_C^0] \\
& +(N - N_A - N_C)\pi_{AC}^d[\phi_{AC} \cdot \mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} + \mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*}] \\
& -(N_{\bar{A}}\pi_A^0 + N_{\bar{C}}\pi_C^0)[\phi_{AC} \cdot \mathbb{S}_A^0\mathbb{S}_C^0 + \mathbb{P}_A^0\mathbb{P}_C^0] \\
& +N[\phi_{AC} \cdot \mathbb{S}_A^0\mathbb{S}_C^0 + \mathbb{P}_A^0\mathbb{P}_C^0] \times [-\pi_C^d[\phi_{AC} \cdot \mathbb{S}_A^0\mathbb{S}_C^d + \mathbb{P}_A^0\mathbb{P}_C^d] - \pi_A^d[\phi_{AC} \cdot \mathbb{S}_A^d\mathbb{S}_C^0 + \mathbb{P}_A^d\mathbb{P}_C^0] \\
& +\pi_{AC}^d[\phi_{AC} \cdot \mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x} + \mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*}] - d[\phi_{AC} \cdot \mathbb{S}_A^0\mathbb{S}_C^0 + \mathbb{P}_A^0\mathbb{P}_C^0]]
\end{aligned} \tag{5}$$

Then, we can write Equation(5) by separating terms including ϕ and ϕ^2 as follow

$$\begin{aligned}
\Delta U_{AC} \;=\; & \phi_{AC}^2 N\mathbb{S}_A^0\mathbb{S}_C^0[-\pi_C^d\mathbb{S}_A^0\mathbb{S}_C^d - \pi_A^d\mathbb{S}_A^d\mathbb{S}_C^0 + \pi_{AC}^d\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - d\mathbb{S}_A^0\mathbb{S}_C^0] \\
& +\phi_{AC}[N\mathbb{S}_A^0\mathbb{S}_C^0(-\pi_C^d\mathbb{P}_A^0\mathbb{P}_C^d - \pi_A^d\mathbb{P}_A^d\mathbb{P}_C^0 + \pi_{AC}^d\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - d\mathbb{P}_A^0\mathbb{P}_C^0) \\
& +N\mathbb{P}_A^0\mathbb{P}_C^0(-\pi_C^d\mathbb{S}_A^0\mathbb{S}_C^d - \pi_A^d\mathbb{S}_A^d\mathbb{S}_C^0 + \pi_{AC}^d\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - d\mathbb{S}_A^0\mathbb{P}_C^0) \\
& +N_A\pi_C^d\mathbb{S}_A^0\mathbb{S}_C^d + N_C\pi_A^d\mathbb{S}_A^D\mathbb{S}_C^0 + (N - N_A - N_C)\pi_{AC}^d\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} \\
& -(N_{\bar{A}}\pi_A^0 + N_{\bar{C}}\pi_C^0)\mathbb{S}_A^0\mathbb{S}_C^0] + N_A\pi_C^d\mathbb{P}_A^0\mathbb{P}_C^d + N_C\pi_A^d\mathbb{P}_A^D\mathbb{P}_C^0 \\
& +(N - N_A - N_C)\pi_{AC}^d\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - (N_{\bar{A}}\pi_A^0 + N_{\bar{C}}\pi_C^0)\mathbb{P}_A^0\mathbb{P}_C^0 \\
& +N\mathbb{P}_A^0\mathbb{P}_C^0[-\pi_C^d\mathbb{P}_A^0\mathbb{P}_C^d - \pi_A^d\mathbb{P}_A^d\mathbb{P}_C^0 + \pi_{AC}^d\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - d\mathbb{P}_A^0\mathbb{P}_C^0]
\end{aligned} \tag{6}$$

Equation(6) contains coefficients of ϕ and ϕ^2 and several terms which are independent of ϕ.

A.3 Proof of Theorem 2.4

The first derivative of $\Delta U(\phi)$ is:

$$\Delta U'(\phi) = 2\gamma_2\phi + \gamma_1 \tag{7}$$

We can show that, γ_1 and γ_2 is positive; accordingly, when $\phi \geq 0$, $\Delta U'(\phi) \geq 0$. Thus, ΔU is an increasing function of ϕ. Equation 8 and Equation 2.17 show γ_1 and γ_2 are positive respectively. We know x maximizes $\mathbb{P}_{AC}^{x,d-x}$; so, $\mathbb{P}_{AC}^{x,d-x} \geq \mathbb{P}_{AC}^{0,d}$ and $\mathbb{P}_{AC}^{x,d-x} \geq \mathbb{P}_{AC}^{d,0}$.

$$\begin{aligned}
\gamma_1 \;=\; & N\mathbb{S}_A^0\mathbb{S}_C^0(\pi_C^d(\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - \mathbb{P}_A^0\mathbb{P}_C^d) + \pi_A^d(\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - \mathbb{P}_A^d\mathbb{P}_C^0) + d(\mathbb{P}_A^{x^*}\mathbb{P}_C^{d-x^*} - \mathbb{P}_A^0\mathbb{P}_C^0)) \\
& + N\mathbb{P}_A^0\mathbb{P}_C^0(\pi_C^d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^0\mathbb{S}_C^d) + \pi_A^d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^d\mathbb{S}_C^0) + d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^0\mathbb{S}_C^0)) \\
& + (\pi_A^0 + \pi_C^0)(N - N_A - N_C)(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^0\mathbb{S}_C^0) + N_A\pi_C^0(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) \\
& + N_C\pi_A^0(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^0) + \pi_C^d(N - N_A - 2N_C)\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} \\
& + \pi_A^d(N - 2N_A - N_C)\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} + N_Ad(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) + N_Cd(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) \\
& + N_A\pi_A^0\mathbb{S}_A^0\mathbb{S}_C^d + N_C\pi_C^0\mathbb{S}_A^d\mathbb{S}_C^0
\end{aligned} \tag{8}$$

Without loss of generality by assuming $\beta_A \geq \beta_C$, we have $\mathbb{P}_{AC}^{x^*,d-x^*} = \mathbb{P}_{AC}^{d,0}$ and we can re-write Equation 8 as:

$$\begin{aligned}
\gamma_1 \;=\; & N\mathbb{S}_A^0\mathbb{S}_C^0(\pi_C^d(\mathbb{P}_A^d\mathbb{P}_C^0 - \mathbb{P}_A^0\mathbb{P}_C^d) + d(\mathbb{P}_A^d\mathbb{P}_C^0 - \mathbb{P}_A^0\mathbb{P}_C^0)) + N\mathbb{P}_A^0\mathbb{P}_C^0(\pi_C^d(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^d) \\
& + d(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^0)) + \pi_C^0(N - N_A - N_C)(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^0) \\
& + \pi_A^0(N - 2N_A)(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^0) + N_A\pi_A^0(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) + N_A\pi_C^0(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) \\
& + N_C\pi_A^0(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^0) + \pi_C^d(N - N_A - N_C)\mathbb{S}_A^d\mathbb{S}_C^0 + \pi_A^d(N - N_A - N_C)\mathbb{S}_A^d\mathbb{S}_C^0 \\
& + N_Ad(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) + N_Cd(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0) + d\pi_A^0\mathbb{S}_A^d\mathbb{S}_C^0 + d\pi_C^0\mathbb{S}_A^d\mathbb{S}_C^0
\end{aligned} \tag{9}$$

We know probability of purchasing each item in large datesets is less that 0.5. Also, probability of purchasing of items after discount is larger or equal to probability of purchasing before discount ($\mathbb{P}_A^d \geq \mathbb{P}_A^d$). Accordingly, Equation 9 is positive.

Besides, assuming probability of purchasing each item in large datasets is less than or equal to 0.5, we have $\mathbb{S}_A^d \geq \mathbb{S}_A^0$. Thus, Equation 10 is positive too.

$$\begin{aligned}\gamma_2 &= N\mathbb{S}_A^0\mathbb{S}_C^0[\pi_C^d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^0\mathbb{S}_C^d) + \pi_A^d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^d\mathbb{S}_C^0) + d(\mathbb{S}_A^{x^*}\mathbb{S}_C^{d-x^*} - \mathbb{S}_A^0\mathbb{S}_C^0)] \\ &= N\mathbb{S}_A^0\mathbb{S}_C^0[\pi_C^0(\mathbb{S}_A^d\mathbb{S}_C^0 - \mathbb{S}_A^0\mathbb{S}_C^d) + d(\mathbb{S}_A^0\mathbb{S}_C^d - \mathbb{S}_A^0\mathbb{S}_C^0)] \end{aligned} \tag{10}$$

A.4 Proof of Theorem 2.6

Assuming $\beta_A = 0$, we have

$$\mathbb{P}_A^d = 1 - \frac{1}{1 + e^{\log\left(\frac{\mathbb{P}_A^0}{1-\mathbb{P}_A^0}\right)}} = \mathbb{P}_A^0 \tag{11}$$

Therefore, when both items are price insensitive (i.e. $\beta_A = 0$ and $\beta_C = 0$), we know $\mathbb{P}_A^d = \mathbb{P}_A^0$ and $\mathbb{P}_C^d = \mathbb{P}_C^0$. Accordingly, regardless of what d is all joint probabilities would be

$$\mathbb{P}_{AC}^{0,0} = \phi_{AC}\sqrt{\mathbb{P}_A^0\left[1 - \mathbb{P}_A^0\right]\mathbb{P}_C^0\left[1 - \mathbb{P}_C^0\right]} + \mathbb{P}_A^0\mathbb{P}_C^0 \tag{12}$$

we have

$$\mathbb{P}_{AC}^{d,0} = \mathbb{P}_{AC}^{0,d} = \mathbb{P}_{AC}^{x^*,d-x^*} = \mathbb{P}_{AC}^{0,0} \tag{13}$$

By replacing all joint probabilities with $\mathbb{P}_{AC}^{0,0}$ the change in the utility will be

$$\Delta U_{AC} = -N_{AC} \cdot d \leq 0 \tag{14}$$

Therefore, the utility decreases after bundling.

A.5 Proof of Theorem 2.7

According to Theorem 2.5, for a profitable item pair we have $\phi_{AC} \geq \phi^*(\delta)$. From the TAPER algorithm developed by Xiong et al. [60], an upper bound of ϕ_{AC} can be found as (assuming

that $N_A \geq N_C$)

$$\overline{\phi}_{AC} = \sqrt{\frac{N_C}{N_A}}\sqrt{\frac{N - N_A}{N - N_C}}$$

We have $\overline{\phi}_{AC} \geq \phi_{AC} \geq \phi^*(\delta)$; hence, $\overline{\phi}_{AC} \geq \phi^*(\delta)$.

B Search of Profitable Bundles

A bundle is a profitable offering when its utility gain is greater than the user-specified minimum, δ. In other words, we want to find all two-item bundles $\{A, C\}$ such that

$$\Delta U_{AC} \geq \delta \text{ under discount } d = s \cdot (\pi^0_A + \pi^0_C).$$

Input : ItemsTable: items information including β_A, π^0_A, $\mathbb{P}^0_A$ and N_A for each item A;
Output : $\mathcal{S}$: a set of all selected item pairs
Parameters: δ: a user-specified minimum gain; s: discount rate ;

$M \leftarrow Size(ItemsTable)$;
for $a \leftarrow 1$ *to* $(M-1)$ **do**
  for $c \leftarrow a+1$ *to* M **do**
    `// Compute exact correlations` ϕ_{AC}
    $\phi_{AC} \leftarrow \frac{\mathbb{P}^{0,0}_{AC} - \mathbb{P}^0_A \mathbb{P}^0_C}{\sqrt{\mathbb{P}^0_A\left[1-\mathbb{P}^0_A\right]\mathbb{P}^0_C\left[1-\mathbb{P}^0_C\right]}}$ by Equation 2.2 ;
    `// Compute minimum correlations requirement`
    Calculate $\phi^*(\delta)$ by Equation 2.20 ;
    if $\phi_{AC} \geq \phi^*(\delta)$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{A, C\}$
    end
  end
end
return $\mathcal{S}$;

Algorithm 6: The Brute force Algorithm

Input : ItemsTable: items information including β_A, π_A^0, $\mathbb{P}_A^0$ and N_A for each item A;

Output : $\mathcal{S}$: a set of all selected item pairs

Parameters: δ: a user-specified minimum gain; s: discount rate;

$M \leftarrow Size(ItemsTable)$;
for $a \leftarrow 1$ *to* $(M-1)$ **do**
  for $c \leftarrow a+1$ *to* M **do**
    `// Pruning by Theorem 2.6`
    if $\beta_A \neq 0$ *or* $\beta_C \neq 0$ **then**
      Calculate $Upper(\phi_{AC})$;
      `// Compute minimum correlations requirement`
      Calculate $\phi^*(\delta)$ by Equation 2.20;
      `// Pruning by Theorem 2.7`
      if $Upper(\phi_{AC}) \geq \phi^*(\delta)$ **then**
        `// Compute exact` ϕ_{AC}
        $\phi_{AC} \leftarrow \frac{\mathbb{P}_{AC}^{0,0} - \mathbb{P}_A^0 \mathbb{P}_C^0}{\sqrt{\mathbb{P}_A^0\left[1-\mathbb{P}_A^0\right]\mathbb{P}_C^0\left[1-\mathbb{P}_C^0\right]}}$ by Equation 2.2 ;
        `// Pruning by Theorem 2.5`
        if $\phi_{AC} \geq \phi^*(\delta)$ **then**
          $\mathcal{S} \leftarrow \mathcal{S} \cup \{A, C\}$
        end
      end
    end
  end
end
return $\mathcal{S}$;

Algorithm 7: The PBQ Algorithm

Vita

Nooshin Hamidian is a Ph.D. candidate in Industrial and Systems Engineering at the University of Tennessee at Knoxville. She received a B.S. degree from K.N.T. University of Technology and a M.S. degree from Sharif University of Technology. Her Ph.D. thesis is entitled Correlation- based Analytics in Big Data, and her research interests include correlation analysis, data mining, statistical analysis, and Operations Research.