ABSTRACT

Title of thesis: AN EXPECTATION MAXIMIZATION APPROACH TO REVENUE MANAGEMENT ON RAIL TICKET DATA

Kartik Kaushik, Master of Science, 2016

Thesis directed by: Professor Cinzia Cirillo
Department of Civil Engineering

In the world of sale of perishable commodities without regulation, competition causes cut-throat pricing and heavy discounts for the commodity. Even though this is beneficial to the customers, the companies that offer the commodity have to be careful to prevent the offered discounts and cut-throat pricing from cutting into their profits. The science of managing revenues in such scenarios is loosely termed as Revenue Management (RM). RM holds its roots to the competition generated in the American airline industry after deregulation. Since then, it has spread to virtually all industries that deal with perishable commodities such as hotel and hospitality, rental vehicles, and all forms of long distance public transportation, even freight [23].

The commodities in these industries refer to the items for sale. In a hotel, it may be rooms of different classes and sizes; in vehicle rentals, cars; and in all forms of long distance transportation, seating space. Perishability of these commodities can be understood simply by the fact that after a certain date, a certain commodity

will not be available. In long distance transportation, it is easy to imagine that the seats on a vehicle (plane, bus, train or ferry) will not be available after the vehicle has departed on its way. Similarly rooms in a hotel or cars with a rental agency will loose value the longer they are kept empty or unused. The goal of modern day RM is, therefore, to ensure profitable sales of such commodities, such that they are priced at better rates than the competition.

This thesis attempts to apply the theory of Expectation Maximization (EM) to the purchase data from railway industry in a attempt to better the existing pricing logic. The EM algorithm used here was developed by Dr. Kalyan Talluri and Dr. Gareth van Ryzin in their seminal paper published in 2004 [49]. In that paper the authors develop the algorithm, derive the mathematics that powers it and apply it to test data sets to prove that it out performs the current industry standard. However, application of that method to a real dataset has never been done, which is the goal of this thesis.

We find, and document herewith, the issues that resulted from applying the EM algorithm directly to the data. Mainly, assumptions in the EM algorithm required heavy data clean up, after which it was found that the results were neither satisfactory nor useful. The reasons for the failure of the model are examined in detail, the primary reason being lack of identifiability in the data. To conclude, the EM algorithm needs substantial modification or additional data in order to lose certain debilitating assumptions and make it more general or reduce the identifiability problem of the data.

AN EXPECTATION MAXIMIZATION APPROACH TO REVENUE MANAGEMENT ON RAIL TICKET DATA

By

Kartik Kaushik

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2016

Advisory Committee:
Professor Cinzia Cirillo, Chair and Advisor
Professor Paul Schonfeld
Dr. Stanley E. Young

Dedication

Fondly dedicated to my Mother and my Advisor,
For their relentless support and inspiration.

Acknowledgments

I would like to express my deepest gratitude and appreciation for my advisor Dr. Cinzia Cirillo. She was a constant source of motivation, knowledge and support. I have learned a lot from her during the course of my Masters at University of Maryland. She inspired me to focus on the details, while not loosing sight of the bigger picture. I owe the completion of this thesis to her.

I would next like to extend my appreciation and gratitude to the Department of Civil and Environmental Engineering at the University of Maryland. The excellent Transportation Engineering program at the department helped me hone my skills and gain valuable knowledge without which completing this thesis would have been an impossibility.

Within the Department of Civil and Environmental Engineering, I would like to sincerely thank the Center for Advanced Transportation Technologies. At CATT, I was exposed to the real world of transportation engineering, with a plethora of real world tasks. Accomplishing those tasks helped me master the skills required to program a computer, handle large amounts of data efficiently, and also kindled the data scientist within me.

Dr. Partha Lahiri and Dr. Daniel Bonnery from the Joint Program in Survey Methodology will forever be honored as the teachers who taught me all that I know about Bayesian Inference, Expectation Maximization, and Probability Distributions and Statistics. I would not have been able to understand anything about my topic without their teachings.

My friends Preeti Lakhole, Darshan Pandit and others at the Department of Civil and Environmental Engineering, were always there for me, no matter the hour or the reason. I am grateful to them for everything from late night pep talks to patiently listening to my rants and showing me the silver lining.

The developers of Python, Pandas, Numpy and Matplotlib — the tools I used for all the analysis in my thesis — I thank humbly. without their dedication to the open source projects, and their awesome skills at coding, I would not have been able to do anything as efficiently as I have been able to. They are the real enablers.

Last, but certainly not the least, I would like to express my sincere thanks, love, humility and gratitude to my Mother and my sister. They have always been there for me, day or night, rain or shine. I could not have reached where I am now without them, and their help, support and understanding.

Table of Contents

List of Figures

List of Abbreviations

β	is a vector of coefficients for model parameters.
$\hat{\beta}$	is estimates of the coefficients, computed by the model.
X	are the model parameters, costs of ticket in each fare class.
t	is the time index from departure of train. $t = 0$ is considered as the train departure. t is monotonic and negatively indexed in the remaining time till departure. Only one customer arrival is permitted per t, at most.
T	is a set of all time periods t.
λ	is the mean of the arrival probability in t.
$\hat{\lambda}$	is the estimate of arrival probability, which is obtained as the total arrivals divided by the total number of periods (see Equation 5.1).
$a(t)$	is the probability of arrival in time t. It is 1 when a confirmed booking has taken place in t, otherwise it is a distribution obtained by Bayes Theorem (see Equation 5.3). When confirmed booking has not taken place, the corresponding $a(t)$ is initialized to zero.
D	is a set of periods in T where a booking has been recorded.
$\bar{D}$	is the periods in T when a booking was not recorded.
P_j	is the probability of choosing j, which can be any of the offered fare classes, or the choice to not buy.

CM	Choice Models
CML	Conditional Maximum Likelihood
DCM	Discrete Choice Models
E-Step	Expectation Step
EM	Expectation Maximization
EMSR	Expected Marginal Seat Revenue
GEV	Generalized Extreme Value
HSR	High Speed Railway
L-BFGS-B	Limited Memory Broyden–Fletcher–Goldfarb–Shanno with Box constraints
LL	Log-Likelihood
M-Step	Maximization Step
ML	Maximum Likelihood
MNL	Multinomial Logit
OR	Operations Research
pmf	Probability Mass Function
RM	Revenue Management
US	United States of America
WESML	Weighted Exogenous Sample Maximum Likelihood
YM	Yield Management

Chapter 1:   Introduction

The foundations for Revenue Management was laid by Littlewood working for British Airways (then British Overseas Airways Corporation) in late 1960s and early 1970s [23]. Prior to that, RM was mainly restricted to overbooking control, whereby companies tried to determine the exact number of seats to overbook such that the target revenues were met on the day of the flight [35]. British Airways was the first company that began to calculate discounts to offer on unsold seats to attract latent demand and improve occupancy and revenues of flights. American Airlines, following the deregulation of airline industry in the US, took the British Airways idea one step further by analytics and data driven inventory control which they called Yield Management [18]. Since the early 1980s this practice of Yield Management spread to other airlines and other industries such as the vehicle rental and hospitality industries.

Ever since British Airways implemented discounted fares for passengers booking early, two problems arose which are still being solved to this day. The first problem was that of date: how late to close the early booking period. British Airways had a 21 day policy wherein bookings occurring before 21 days from departure were eligible for the discount fares, while patrons booking after 21 days before de-

parture of flight had to pay the full fare [35]. The second issue was that of capacity: how many seats to allocate (protect) for the late booking customers who were willing to pay the full fare. If too few seats were protected, the flight would lose revenue, if too many seats were protected, the flight might fly with empty seats, again losing revenue.

In order to overcome these two problems, many of solutions have been proposed since the late 1970s. Many of them involved statistical techniques to estimate the correct breaking point so as to maximize revenues. Littlewood devised a thumb rule for solving the problems which stated that discount seats should be offered so long as the revenue obtained from sale of discounted seats exceeded the revenue obtained from sale of full fare tickets [28]. This rule had some issues in implementability as it was never known up front how many full fare paying customers, and how many cheap fare paying customers were expected.

In 1987, Belobaba proposed the Expected Marginal Seat Revenue (EMSR) algorithm that most accurately implemented Littlewood's rule [5, 7]. Belobaba further improved and revised the algorithm in 1989, calling it EMSR-b [6]. The EMSR-b algorithm became the gold standard in Revenue Management, and is still widely used today. These methods, still had some limitations, mainly that they only applied to one leg of a journey, and were not easily extensible to industries like hospitality and rental vehicles, where the horizon (time of expiry of commodities) is not sharply defined. In the intervening years between Littlewood's rulemaking, Belobaba's EMSR algorithms and today the field of RM has vastly diversified with researchers focusing various aspects and problems of the proposed methods. Attempts have been made

to expand a given method to multiple segments and fare classes, and also to develop new methods.

Most of the methods, however, do not account for the passengers. It almost seems assumed by all the methods that passengers are insensitive to price, and the demand will be constant [49]. That is never the case in reality, and accounting for passenger sensitivity and choice behavior will provide the most complete picture for managing capacity. Ultimately, the decision to purchase a given ticket lies with the customer, and understanding and factoring in the customer's choice preferences might be the final key in unlocking the revenue streams. The seminal attempt to factor in the customers' choice behavior has been made by Talluri and van Ryzin in 2004 (although some attempts have been made earlier).

Almost all research that came earlier to account for customer preferences in purchasing tickets, have focused on buy-ups and buy-downs. A buy-up is when a customer decides to buy a higher price ticket when discounts are not available, while conversely, a buy-down is when a customer buys a cheaper ticket when discounts are available. In a general discrete choice customer behavior framework, these decisions are easy to explain, but outside of a well defined utility based framework it is impossible to distinguish customers buying up or down from regular customers [49]. Work by Belobaba (1987) and Phillips (1994), although attractive because it allows for buy-ups, does not scale to a network [5, 40]. Under the pressures of competition, several airlines continued to research a robust solution, resulting in notable papers by Andersson (1998) and Algers and Besser (2001) for Scandinavian Airline Systems [2, 1]. However, it was only with the inclusion of choice behavior models that a new

3

path around the obstacles was paved.

Choice behavior science, and consequently models, are very frail and heavily dependent on conditions and assumptions. Therefore, combining choices with inventory control is much more difficult than initially apparent. Chiefly, inventory companies seldom record visits to their booking system that do not result in a confirmed purchase. The visits data is generally filled with noise made by people with no intent to buy at the time of visit, but who visit to track prices or explore options. This causes missing data in that no-purchase decisions are not recorded. Choice models in general require an exhaustive list of choices in the data. All items that can be chosen must have been recorded in the data. Failing this, it is generally not possible to fit a choice model.

The paper by Talluri and van Ryzin proposes a framework which is agnostic to choice models, and can work with any model that provides probabilities as output [49]. Within the framework, the customers' arrival are also estimated. The estimated arrival acts as a proxy to the no-purchase records and enables estimating an embedded choice model. In order to estimate both arrivals and choice model parameters together, the framework relies on Expectation Maximization, where iteratively the current choice model parameters are used to estimate expected arrivals in the expectation step, which in turn is used to fit the choice model and update the parameters in the maximization step.

The idea of the EM algorithm is to iteratively find a stable point in the data where the arrival rate and choice model parameters converge. Upon convergence, the estimated parameters and arrival rate can be used to predict customer arrival, and

ascertain customer sensitivities to price. Therefore, this algorithm affords real time control over prices, such that various revenue targets can be met taking customer choice behavior into account. It affords better overall control over the price of the inventory, and can maximize sales easily. Moreover the parameters can be estimated before hand, on past data from the same market segments operating at the same times, and improved with each new datum, offering flexibility and updated parameters.

However, as illustrated in this thesis, it is not without problems. EM algorithms in general tend to be finicky, and convergence is not always guaranteed. Further, the problem examined in this thesis has an identifiability issue. Identifiability issues arise when a given segment of missing data cannot be identified. In this case, neither the complete population size of the visitors to the system, nor the no-purchase rate for unit number of purchases is known. Therefore, there is a degree of freedom in the model that allows for multiple values as solution. This causes the EM convergence to become even more uncertain, and bounce between values far too often, or for the objective function to be infinitely decreasing with no minima.

Some of the issues identified in this thesis can be corrected, but the argument of using EM to solve the problem discussed in this thesis is challenged. When data has an issue with identification of missing values, EM is a poor choice as a solver. However, as examined in the last section of this thesis, there may be ways to still overcome the challenges and hurdles.

The following subsections clearly define the problem statement, the research scope and the contributions of this thesis.

1.1   Problem Statement

This thesis attempts to solve missing data in a consumer demand framework for managing inventory to maximize revenues. The problem deals exclusively with data from a major public long-distance train operator, and all discussions henceforth will be specific to that market. Customers deciding to not purchase a ticket on a train are not recorded in the booking transaction data, which means there are no data for a whole class of customers. Due to the exhaustive data requirements of most choice behavior models, the missing data needs to be imputed.

A generic Expectation Maximization framework for jointly estimating choice model parameters and the missing customer data was proposed by Kalyan Talluri and Gareth van Ryzin in 2004. In their seminal paper, Talluri and van Ryzin, apply their proposed method to synthetic data only. The literature that followed their paper, to improve the method or propose rival methods also mainly demonstrate results from synthetic data, or real data with additional information. There has been no attempt in literature to apply the methods to real data without additional information, and test fidelity (see Section 2.3).

The problem statement of this thesis is two-fold: to solve for the missing data problem, to apply the EM algorithm to real data, from a railway company, in the absence of additional information about the data. In the process, the goals are to understand the workings of an EM algorithm, and the issues that arise from it. If the results of the EM are logical and sensible, the next step is an attempt to solve the Bellman Equation by a dynamic program so that appropriate decisions about

6

offer price can be made. However, the results are neither correct nor rational, and hence the dynamic program has not been solved.

1.2   Research Objectives and Scope

The main scope of this thesis is to implement and fit the EM algorithm proposed by Talluri and van Ryzin. The algorithm is first fit to synthetic data created similarly to the data Talluri and van Ryzin demonstrate in their paper. Then the algorithm is fit to real data from a railway company, and the algorithm is studied to examine all the points where it fails, and propose strategies to improve the weak links.

This thesis attempts to thoroughly test the EM algorithm with real data. This has never been documented in existing literature. The two objectives of this research are to examine if the algorithm can be fit to real data in the naive form, and if it works on railway data. In their paper, Talluri and van Ryzin, build the framework with the idea that the data will be from airlines. However, the booking pattern for railways is vastly different from the booking patterns for airlines. Moreover, the classes and prices offered by railway companies are different from airline classes and fares as well.

Railways can offer many more classes, with a range of amenities to the passengers. Everything from seating location to in-seat service can be offered at various price points through out the booking period. However, the railway company that provided the data for this thesis offers a single price for each class of tickets

throughout the booking period. Nonetheless, the booking pattern of customers is much different from the booking pattern found on airlines: they cluster towards the end of the period, just before the departure of the train (see Section 4.2).

Booking patterns seen with the railway data cause problems with the assumptions made while deriving the framework. The authors of the original paper assumed that arrivals are uniformly distributed and represented by a univariate Poisson process. However, the real data bookings cluster towards the end of the booking period, which is not representative of uniform Poisson distribution. Moreover, the original authors also restrict that only one customer arrival is permitted per time step, and because bookings are clustered in the railway data, one arrival per time step requires very fine time windows, which cause overestimation of arrival rate.

Consequently, applying the EM algorithm to real data from a railway company is challenging and difficult.

1.3 Contributions

The research presented herein shows that the EM algorithm proposed by Talluri and van Ryzin does work on synthetic well-behaved data, but fails to work on real data. To enumerate, the major contributions of this thesis are as follows:

- Implement the EM algorithm on real data from a railway company, and test it, in the process understand the EM framework.

- Understand and reveal the reasons the algorithm failed to produce credible results.

- Analyze the reasons the algorithm failed, and present the issues.

- Suggest improvements to be applied in the future to the algorithm to overcome the shortcomings.

1.4   Thesis Organization

The rest of this document consists of literature review (chapter 2). Chapter 3 is an overview of the model and assumptions made, so that the remaining chapters fit together in a seamless and confusion free manner. The data used and the method are elaborated in chapters 4 and 5. Results are discussed in chapter 6, before conclusions are drawn in chapter 7. The last chapter, 8, discusses future work where some of the identified limitations can be lifted.

Chapter 2:   Literature Review

In this chapter, the existing work, background information and the relevant context is presented. This chapter is divided into subsections to discuss in detail the various aspects that compose this thesis. The first section, Yield Management presents the history of Revenue Management, and a rough overview of the existing methods before tying it to the decision theory and choice models, which make up the second section. Within the section of Choice Models, the primary issue with the choice models is illustrated, namely the missing data. Finally, modern day revenue management efforts are shown.

2.1   Yield Management

When yield management was conceived by Littlewood more than fifty years ago, it was a radical practice that transformed the way airline companies managed the seats on their flights. Prior to that, the companies mainly focused on overbooking control wherein they tried to fix the number of seats to allow to be overbooked. It only worked on segments with high enough demand to fill more than one vehicle. And, it did nothing to affect the prices, and seats remained expensive for a majority of the population.

Since British Airways began offering seats at a discount provided the booking was done more than 21 days before departure of the flight, they took advantage of a latent demand which always existed but could not afford the seats at the full price [35]. These discounted seats meant that segments of the flight which flew with empty seats now got closer to a 100% occupancy rates. However, this practice soon gave rise to two issues: the amount of discount to offer and the period or capacity for which to offer the discount.

The question of the amount of discount to offer is fairly easy to solve. The discounted fare should ideally not fall below the marginal cost of transporting that passenger. Computation of marginal costs of providing a service to one more customer is well defined and studied in microeconomics [45]. Therefore, if a base fare is fixed, the discount can drop as low as the marginal cost. Any lower value would incur a loss. Very little research is actually focused on improving the offered discounts, although the trade offs between offered discounts and additional perks is an interesting topic of research. This thesis similarly does not focus on the offered price, but instead concentrates on the effects that given prices will have on the behavior or prospective customers.

The other question is actually a two fold question of time and capacity, which are intrinsically linked to each other because of customers' willingness to pay. The problem is to define a point in time and capacity such that discounts can be either wholly or partially lifted upon meeting that point. In a 300 passenger airplane, for example, the point would be to determine the date after which the discounts will not be applied (21 days in the case of the original British Airways method) or the

capacity that needs to be protected for higher paying customers.

Although they seem disconnected, they are linked because customers arriving later in the booking period may have a higher willingness to pay. With a bit of introspection, it becomes apparent that later arriving customers have a more pressing need to travel, the reason for travel came up recently. If the capacity target has not been met, and the booking is kept open at discounted fare rates well into the later stages, the seats will be sold at lower prices to customers who are willing to pay more thus loosing revenue for the airline. On the other hand, if discounts are discontinued early in the booking period, and not a lot of customers are willing to pay the higher fare, the vehicle will remain with empty seats. A universal solution to this problem will have to be a function of time, capacity and demand.

The balancing rule for these two parameters was suggested by Littlewood, and later came to be known as Littlewood's rule [28]. Littlewood simply stated that as long as revenues from discounted fares exceeded the future expected revenue from full fare sales, discounted bookings should be accepted. Most of the literature that followed including Belobaba's EMSR algorithms, used this rule as foundation [5, 7, 6]. Attempts were made to develop better heuristic approaches that compute the optimal capacity and time for different discount levels, and to apply these approaches to more than one leg of journey.

Subsequent attempts have thoroughly researched the problem of single-leg models. In these models, demand is assumed to occur in non-overlapping periods, and each segment of a global journey is treated as a separate entity. Works by Curry (1990), Wollmer (1992), Brumelle and McGill (1993) and Robinson (1995),

are notable in this area, and show that nested allocations can result in optimal policies [19, 58, 10, 41, 49]. Similarly, a huge amount of work has been done in solving the much harder multi-leg allocations that deal with a whole network. Key papers in this area are Glover et al. (1982), Dror et al. (1988), Williamson (1988, 1992), Simpson (1989), Curry (1990), Talluri (1996), Talluri and van Ryzin (1998, 1999), and Cooper (2000) [22, 21, 57, 56, 44, 19, 50, 48, 47, 16]. See Barnhart and Talluri (1996), and McGill and van Ryzin (1999) for more information [3, 35].

These approaches, however, just approached the problem from the view point of filling up seats. The passengers who made the decision to buy or not buy were not considered in the equations. Models that aim to predict the customers' choices are well known, extensively researched and applied, but there are not many integrated approaches where inventory and customers are considered together, as outlined in the following sections.

2.2 Choice Models

Choice models have been researched since at least 40 years ago [31]. The cornerstone, however, of choice models is Dr. McFadden's Nobel prize winning work in 1974 [34]. Traditionally choice models were constructed to have a simple binary form, to which a logistic expression could be applied. This was made possible by Luce's assumption of independence from irrelevant alternatives [31]. This assumption allowed researchers to treat each choice as being independent of the others, thus enabling logistic regression to solve for the parameters. McFadden showed

that unobserved utility in a logit framework had an extreme value Gumbel type I distribution, thus deriving a closed form expression [34].

Since then, choice models have expanded to include probit, and other generalized extreme value models to cater to situations where the independence from irrelevant alternatives assumption breaks down, and a correlation exists between the alternatives or the subjects in the choice experiment. Further, to account for multiple choices, the logistic regression was replaced by multinomial logit (MNL) formulations [8].

MNL models have been extensively applied to areas where people's choices needed to be understood and forecasted. The most widespread use of choice models is in the choice and options theory as it applies to stocks and consumer goods industry. In transportation they are often used in predicting ridership of new lines and service brackets, or development avenues [13, 14]. Investments in proposed transportation options are often decided based on potential usage forecasts, which are computed by applying MNL and other choice models to stated and revealed preference surveys [30, 29].

Even though choice models are designed to predict the choices made by people usually buying something, they were not that common in revenue management, until recently. Researchers that deal with revenue management often work as if customer demand is insensitive to the prices offered, and focus only on inventory control. In reality this is never the case, with customers willing to go to different lengths to suit their individual travel requirements. Therefore incorporating choice behavior into revenue management is the next step in improving revenue streams.

Choice models, unfortunately, have a few issues before they can be used with revenue management. The first is the independence from irrelevant alternatives assumption [51]. While this assumption will help in predicting if a customer will use a personal car, or public transport to get to work, it hinders prediction when the choice is between the same service offered at different price points because the alternatives in this case are not independent of each other. Additionally, choice models require that the choice set is complete and exhaustive [51].

Exhaustive dependent variables actually are a necessity in any kind of regression analysis. In the case of Choice Models, the dependent variable are the choices, which depend on independent variables such as characteristics of the individual, including needs and wants, and offered amenities. Unless a given choice exists in the dependent variable, it cannot be modeled. Therefore, all missing data will either have to be accounted for, or the model specifications changed to exclude the requirement of missing data.

2.2.1 Missing Data Problem

Missing alternatives is a big problem to discrete choice modeling. An alternative cannot be included in the model if there is no data for that alternative. The decision maker must choose one of the available options, and nothing from outside. Conceptually, it is easy to include all options in the choice set. For example, if three fare classes are up for sale at different price points, and the customer can buy any one of them or none at all, then the choice set consists of four alternatives: the three

fare classes and the option to not buy [51].

All ticketing systems record transaction data, but they usually only record the transactions where a purchase or a cancellation was made. The visits to the system made by customers who walked out on the available options are seldom recorded [37]. This results in a missing or censored data problem, and logit models cannot be directly calibrated on this data. Logit models need all possible choices as a part of the choice set, or the dependent variable. If one or more of the available choices are not part of the choice set, the model will have a 'leak', and the coefficients fitted will not be meaningful [51].

Several attempts have been made to resolve the missing data issue. At the outset, the solutions proposed in the literature can be divided into two factions: those that require knowledge of the market shares, and those that dont. If market shares (alternative specific constants) are known then the Weighted Exogenous Sample Maximum Likelihood (WESML) method proposed by Manski and Lerman (1977) can be used [32]. Similarly, the Conditional Maximum Likelihood (CML) method developed by Manski and McFadden (1981) may also be used [33]. Note that these works are from quite long ago, almost around the time when the logit model was initially developed.

When the market shares are unknown, most of the models require an additional source of information that can be used to estimate the model coefficients. Different models need different pieces of information from which to reconstruct the coefficients based on the observed parameters and the observed and unobserved choices. The CML method from Manski and McFadden (1981) can be used for situations when

16

market shares are unknown, with the CML coefficients also estimated alongside the parametric coefficients [33].

Cases where market shares are unknown have only recently gained in popularity, however, with the first set of papers being published in 2004. Talluri and van Ryzin (2004) formulate an EM algorithm that uses Bayes Theorem to combine the probability functions of customer arrival and choices [49, 25]. They show that this method can produce better revenue than the well-established Expected Marginal Seat Revenue (EMSR) algorithm. In this case, the customer arrival probability is the external information needed for the model to work, which is estimated in the maximization step of the algorithm.

Newman et. al. (2012) showed that when market shares are unknown, but the total market size is assumed to be stable over time, then the market shares of the missing choice can be estimated along with the logit model parameters [37]. This study stems from a development on the original idea by Talluri and van Ryzin (2004) [49], conducted by Talluri (2009) [46]. Vulcano et. al. (2012) proposed a method in which knowing just the market share of the missing choice is sufficient [52].

Newman et. al. (2013a) further published papers showing that for Nested Logit (NL) class of models, no outside information was necessary for certain nesting orders [38]. Further, they showed — Newman et. al. (2013b) — that by including covariance terms, parameters influencing market shares can be identified for some general models that are a part of the Generalized Extreme Value (GEV) family, obviating the need for external data [36].

2.3 Revenue Management with Choice Models

As mentioned, pricing and inventory control are two sides of the same coin. Pricing determines the speed of sale of inventory. Cheaply priced inventory is more likely to sold faster than same inventory available at a greater cost. However very few studies were carried out with joint estimation of customer behavior and inventory control before the 2000s. Since then a few studies have taken a complete view of the problem, and attempted to combine customer behavior with sale of products.

2.3.1 Independent Demand Source

Studies outlined here use another model or dataset to source the demand. There is no coupling such that set prices on seats affects the demand. For example, if the demand were sourced from an Origin-Destination dataset on which a route choice or a mode choice model has been applied, the resultant demand will be unaffected by the prices being charged. In other words, these models have static global demand, and work to improve market capture.

Kuyumcu and Garcia-Diaz (2000) were the first to study the joint problem of pricing and seat allocation [26]. This study came directly after papers by Weatherford (1997) [53] and McGill and van Ryzin (1999) [35] wherein the argument for a combined behavior and allocation models were first presented. In their study Kuyumcu and Garcia-Diaz assumed no connection between offered fares and product characteristics [26]. Price was therefore an exogenous variable to customers, whose demand was assumed to be normally distributed with no cross talk between

fare classes, or market segments. In other words, generated customer demand only catered to a single fare class, and a single market segment.

Hood (2000) improved on a framework laid by Ciancimino, et. al. in 1999 that allocates an optimal number of seats for a railway between a pair of stations [24]. Ciancimino, et. al. modeled demand as a truncated normal distribution [12], which Hood used in a logit model to optimize pricing and timetable for the framework Ciancimino et. al. developed. Similarly, yet differently, Bertsimas and de Boer (2002) assumed demand was uncertain, but the expected demand was a function of only the offered price [9]. They formulated a model for a network, and showed that for some demand situations the model was convex and therefore had a solution for large instances, and could lead to significant revenue gains.

In a departure from integrated demand and revenue management, Cote, et. al. (2003) modeled pricing and seat allocation in the presence of a competitor [17]. The moves of the competitor were assumed known, while the demand and price were decoupled from each other. The main goal of this model was to make the best decisions to beat competition. On the other hand, Whelan with Johnson in 2004 and with various authors in 2008 formulated a nested logit to estimate the impact of prices on train overcrowding, without regard to revenues [54, 55]. The upper level consisted of the choice of taking a train or not, which fed demand to the lower level which chose between classes in a train. The demand for the upper level was generated from an activity based model. Modeling demand similarly, by using an activity based model, Li, et. al. (2006) used the nested logit to solve for optimal dynamic pricing policies and overall network performance [27].

Also using a nested logit model, but with the upper level modeling a choice between the Japanese High Speed Railway (HSR) and domestic airline, Ongprasert (2006), studied seat allocation policy for the HSR [39]. The lower level of the nested logit model consisted of fare choices, which fed a seat allocation model with choice data. Improvements in revenue were shown in the results of the study.

Chew, et. al. (2008) used a dynamic program to solve for seat allocation and pricing in the current time step by modeling expected sales in the future based on fare and customer demand [11]. Demand was modeled as a linear function of price, which was in turn steadily increased as vehicle departure drew close. The model was concave, and therefore a recursive method was used to optimize the model When extended to multiple time periods, heuristic approaches were used to solve the model which was no longer concave. Using a dynamic program, but with a nested logit choice model similar to Whelan and Johnson (2004) [54], and Whelan, et. al. (2008) [55], Sibdari, et. al. (2008) solved for dynamic pricing policy for Amtrak data [43]. They showed that the demand in Amtrak data could be modeled as an exponential function, with almost no bookings beyond 30 days from departure. The goal of the study was to model customer choice to buy up or down or to upgrade.

Cizaire (2011) developed an ensemble of methods to jointly estimate seat allocation and customer choices [15]. Demand was generated as a function of the fares predicted by the seat allocation algorithms. Heuristic and deterministic solutions to the models were also presented in this study.

In all the above studies, the no-purchase alternative was conveniently modeled as a different entity, or a known demand process was used to simulate customers.

For example, Cote, et. al. (2003), Whelan and Johnson (2004), Li, et. al. (2006), Ongprasert (2006), and Whelan, et. al. (2008), pass the customers who do not purchase to a competitor or another mode of transport respectively, avoiding imputing data [17, 54, 27, 39, 55]. A full knowledge of customer demand is obtained either by simulating customers from a known process, or from an activity based model. The other papers mentioned above, all simulate demand from a distribution, and thus avoid the identification problem tackled in this thesis.

2.3.2 Combined Estimation of Demand and Choice Parameters

Studies that attempt to estimate demand and choice parameters, with the complication of lack of no-purchase transactions are fewer and infrequent. Talluri and van Ryzin (2004) — the seed paper for this thesis — proposed the first joint optimization framework that estimated the demand, the choice parameters in the light of missing data and then proposed optimal inventory control policies [49]. An Expectation Maximization algorithm was used to estimate the expected demand (expectation step) and then find the choice parameters that achieved the function maximum (maximization step). This paper lay the foundation for other studies, most of which focused on the missing data aspect and have been discussed above.

Other studies to stem from Talluri and van Ryzin (2004) catered to the complete problem of demand, choice, and inventory control. These papers are mainly authored by van Ryzin in collaboration with other researchers. In an attempt to correct for the identification problem and the degree of freedom in the original

21

framework, Talluri (2009) presented a model with a finite total population [46]. Newman, et. al. (2012) extended Talluri (2008) study to market segments where total population size can be assumed to be constant over a large time window [37]. Stability in demand over a considerable span of time is a reasonable assumption, as the daily variations in number of traveling passengers is averaged out, and the value can be treated as a function of the location served by the transport station under consideration.

Newman, et. al. (2012) estimated the missing data accurately, provided the external information about population size was known to the model [37]. In 2013, however, Newman, et. al. also showed that for certain nesting orders in a nested logit framework, no additional information was necessary [38]. The study uses a nested logit model, within an EM framework, and recovers both generic model parameters and alternative-specific parameters at the same time. The improvement of Newman, et. al. (2013a) over Newman, et. al. (2012) is that this study does not require additional information, and any differences in observation are all attributed to the Gumbel noise parameter. Essentially, the question of number of similar cases that chose different alternatives is posed and formulated, then solved by the EM algorithm. Newman, et. al. (2013b) applies the same method for some Generalized Extreme Value (GEV) models, and show that under certain conditions, additional information is not necessary [36]. Further, they show that the results are consistent with those obtained by other means, provided the sample sizes are large for certain cases. CML is used in this study.

Note that Newman, et. al. (2012, 2013a, 2013b) are all studies that only

estimate the choice models and customer demand, but do not delve into drafting revenue management policies [37, 38, 36]. The understanding is that those studies can be incorporated into any generic inventory control algorithm such as a dynamic program — like the Bellman equations [4] — and used for drawing allocation policies. The drawback of those studies, however, is the complexity of the models. EM is finicky by nature, requiring perfect conditions to converge; in fact, even initial parameter values can be the difference between convergent and divergent solutions [59]. Nested logit and GEV models are further complicated versions of MNL, in which certain assumptions such as independence from irrelevant alternatives are removed and solving these choice models is usually more difficult than solving MNL [51]. Therefore these proposed models can be very difficult and unstable to estimate.

Most recently, in a paper selected to be published in Operations Research, van Ryzin and Vulcano (2016) formulate a demand model that consists of well defined classes of customers, each with an arrival probability given by a probability mass function, and a discrete choice model [42]. They use EM to jointly estimate the pmf of customers, and arrival rate. The assumption is that customers belonging to a particular class have a list of preferred products in the order of greatest utility. They buy the product that ranks the highest on their list, which might also be the choice to not purchase. They show that the convergence of the model is better than competing models.

2.4 Expectation Maximization

Expectation Maximization algorithm was first generalized, explained and named by Dempster et. al. (1977), although it was used before that in specialized forms [20]. EM algorithm is mainly used when a dataset used for fitting a function has missing values. The EM algorithm works in two steps: the first step creates (or updates) a log-likelihood function, and in the second step finds the maximum solution of the log-likelihood.

The initial values of the parameters are taken as the starting point. A maximum likelihood search of all values the missing data can take is then conducted, with the missing values filled by the best candidates. The best candidates are values that maximize the log-likelihood of the objective function. This data, with the imputed missing values, is then used to fit the model and update the coefficients. The coefficients are again used to update the candidates for the missing data, and the cycle repeats till convergence. The EM algorithm works best on exponential functions, where a log transformation will convert the likelihood function into a linear function for easy maximization.

The EM does not guarantee that the likelihood obtained at the end of the Maximization step is the maximum likelihood estimator of the objective function [59]. Multimodal distributions can converge to a local optima, or the function may never converge. Therefore, despite being a generic tool, the EM is not widely used in all missing data situations. A specialized version of EM is used in this thesis to fit the parameters of a choice model to data with missing values.

Chapter 3: Overview and Assumptions

This section presents a quick overview of the method used in this thesis. The EM algorithm proposed by Talluri and van Ryzin is used to fill in missing data, and obtain choice parameters that can be used in determining the optimal pricing strategies.

This study caters to the analysis of railway data. Therefore further discussion will focus solely on trains. We assume that trains in this case are run like airplanes, in that they pick up passengers at an origin and drop them at a destination. Intermittent boarding and alighting is not permitted in the current construct of the problem; however this problem will be addressed in future research. In this study it is assumed that time is backwards-indexed. That is at time = 0, the train departs, and bookings are permitted only between 31 days and four hours in advance of departure time. No sale of tickets on board is permitted.

The data collected by reservation companies often lack the information on patrons who visited the system but did not purchase the tickets. This no purchase choice is what gives rise to the missing data problem inherent in studying patron behavior related to ticketing. The EM algorithm invented by Talluri and van Ryzin is used to impute this missing data so as to enable further analysis. In its most

naive form, the algorithm requires there to be at most one arrival per time step. Therefore, in essence the number of time steps defines the maximum population of patrons that can arrive at the system to purchase tickets. Each patron makes a decision to buy one of the available fare classes, or not to buy at all.

The EM algorithm takes the data for all time steps in the dataset, and for those time steps without any confirmed booking attempts to compute an arrival probability such that the product of the estimated arrival probability and the estimated no-buy probability equal the true no-buy probability. This simple probability relation is fostered by Bayes's theorem, which sits at the heart of the EM algorithm. In a real world scenario, for each time step where a booking was not recorded, a patron may have arrived, and chosen to not buy or may not have arrived. There is no real way to recover this information, and therefore the best that can be done is the product of the estimated probabilities of arrival and no-purchase.

Once initial no-buy and arrival probabilities are estimated, new choice coefficients can be estimated using the imputed data. The new choice coefficients are then used to update the prior knowledge on arrival probability, and then the purchase decision probabilities for each offered product, including the option to not buy. This new data is then used to update the prior knowledge in an iterative manner till convergence is reached. At convergence, the final model coefficients offer the choice parameters under complete data.

3.1 One Arrival per Time Step

The model assumes that time is monotonic and decreasing. That is it moves in discrete steps of constant sizes towards zero, when the vehicle departs (or commodity expires). Further, in it's naive form, the EM algorithm's permits only one arrival per time step. This raises certain issues that are documented but unsolved in this thesis. The major issue is that it fixes the population size to the number of time steps. A fixed population size, with a fixed number of bookings in the booking period makes the probability of bookings known.

A second issue, and more relevant to this thesis, is that during periods of heavy bookings, the recording computer may not have sufficient time granularity to accord each booking a unique timestamp. This in turn causes the need to massage the data such that the assumption is not violated.

Both of these issues can be resolved by assuming a continuous function of arrivals. However, this assumption will hugely complicate the Bayesian equation at the center of the EM algorithm.

Chapter 4:   Data

This chapter describes the data used in the study. In order to ensure that the methodology proposed by Talluri and van Ryzin has been correctly reproduced, the first test was conducted with synthetic data. The algorithm was then applied as is to the real data. The first section in this chapter deals with the details of the data synthesis. The second section details the characteristics of the real data. A subsection within each section discusses the preprocessing done to make it suitable for the algorithm.

In all following discussions, note that one record in the dataset denotes one time step. The two terms may be used interchangeably.

4.1   Synthetic Data

Synthetic data was constructed in order to ensure that the algorithm was correctly reproduced. The passenger arrival rate was fixed at 0.5, meaning a passenger arrived at the system to book a ticket at least in half the time steps. This is done simply by drawing from 0 and 1 with equal probability for each time step. 1 indicates customer arrival, 0 indicates no arrival in the given time step. A total of 30,000 time steps are used, each representing a second, for a total duration of about

8.5 hours. Four fare classes, available at all time steps, but priced differently are included. An arriving customer then faces five choices: buy one of the four fare classes, or no-buy.

The prices of the four fare classes increases in well defined steps as time runs out, but not in proportion to each other. The closer the departure of the vehicle, the higher the price. Using these prices, and setting the choice parameter coefficient to -0.015, the choices of passengers are simulated. To account for unobserved characteristics of an individual, a random Gumbel distributed noise is added to each utility function, and the option that maximizes utility is selected as the choice of the customer.

In other words, the β in equation 5.5 is fixed at -0.015, while ϵ is drawn from a Gumbel distribution for each alternative. The alternative with the maximum utility is assumed to be chosen by the customer. Note that there are 5 alternatives available to the customer: the option to not buy, and the four offered classes. This is done for only those time steps in which the arrival process indicates that a customer has arrived at the system.

From the table of choices, all records when no customer arrived or when a customer chose to not buy anything were deleted. The resulting dataset had 8220 records (out of 30,000 total time steps). Figure 4.1 shows how the prices for the fare classes vary with time. Note that time (on the x-axis) runs out at 0, and the vehicle departs. Each point on the fare price line in figure 4.1 represents a confirmed booking in that class. The legend shows (in parenthesis) the total number of tickets booked in each class. The ordinate shows the price, the abscissa shows time steps.
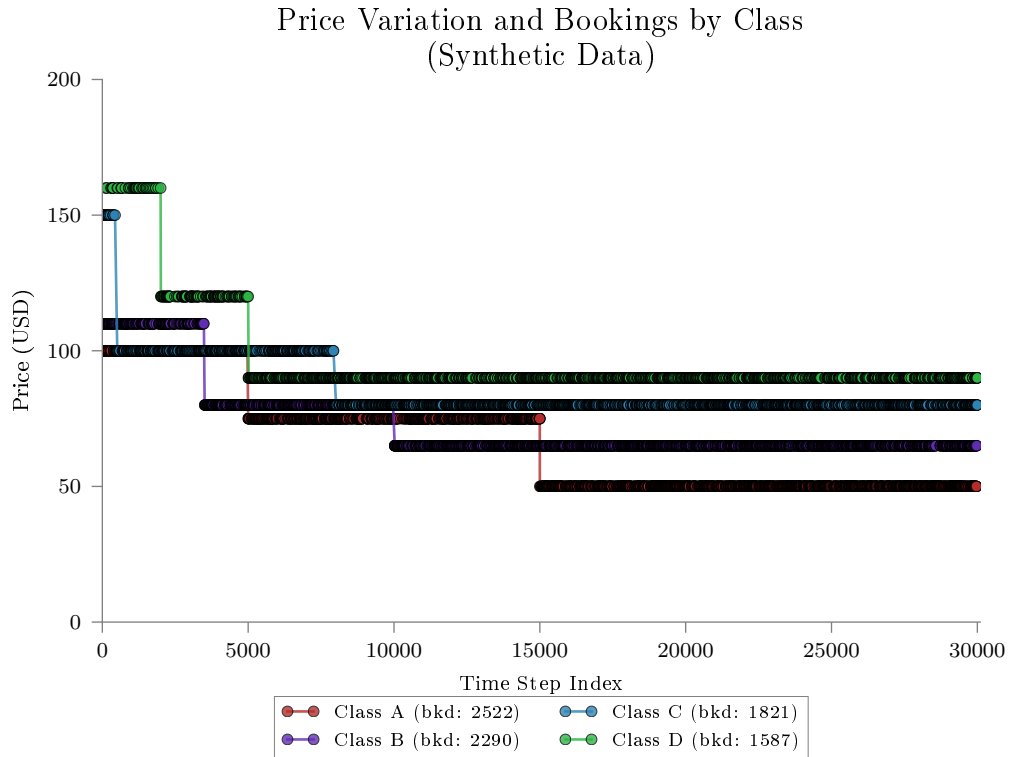
Figure 4.1: Synthetic Booking Prices by Time.

The price of a given class at a given time step can be easily read using this figure, and also the booking pattern inferred.

Table with descriptive statistics is not provided because figure 4.1 provides all information about the classes, purchases and fare prices in one place. A table would not add any more information to this document. The following subsection presents the processing done on the synthetic dataset.

4.1.1 Data Manipulation

In order to use the EM algorithm, a record per time step is necessary in the dataset. In order to reconstruct the time steps that were deleted while synthesiz-

ing the data (no purchases, and no arrivals), two approaches were used: the same number of time steps were recreated and a different number of time steps were recreated.

The second scenario, when a different number of time steps are created is representative of the real world scenario. In the real world one will never know the actual number of time steps that would most accurately capture the arrival process in the booking period. The first scenario, with the correct number of time steps reproduced, therefore represents the control case.

In order to recreate the time steps, the total number of time steps required is decided first. Since 30,000 time steps were used initially, data is recreated with 30,000 time steps for the control case, and 45,000 time steps for the second scenario. With the total number of time steps determined, time indexes are randomly and evenly drawn from a list running from zero to the maximum time step. The confirmed bookings are then placed in their rightful place in the random series of time indexes.

Figures 4.2 and 4.3 show the resultant arrivals per 60 times steps after the purchase dataset is expanded to 30,000 and 45,000 records respectively. As evident from the figures, the final distribution of arrivals is equally random, and the expansion has not affected the data in any way.

Again, since all necessary information is included in the figures, an accompanying data table is not provided.
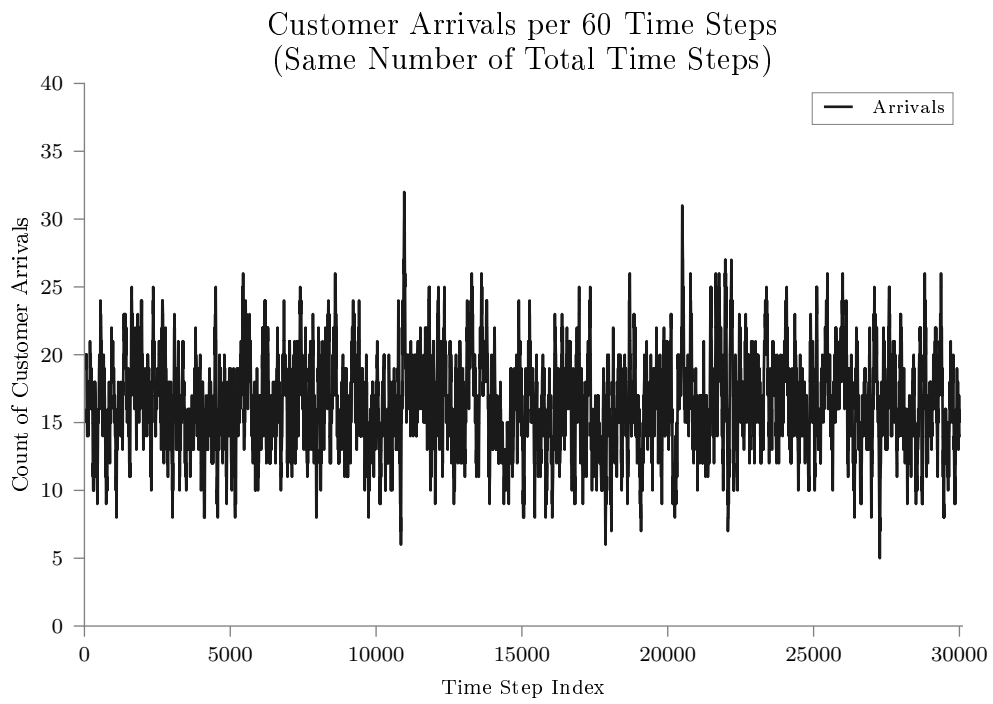
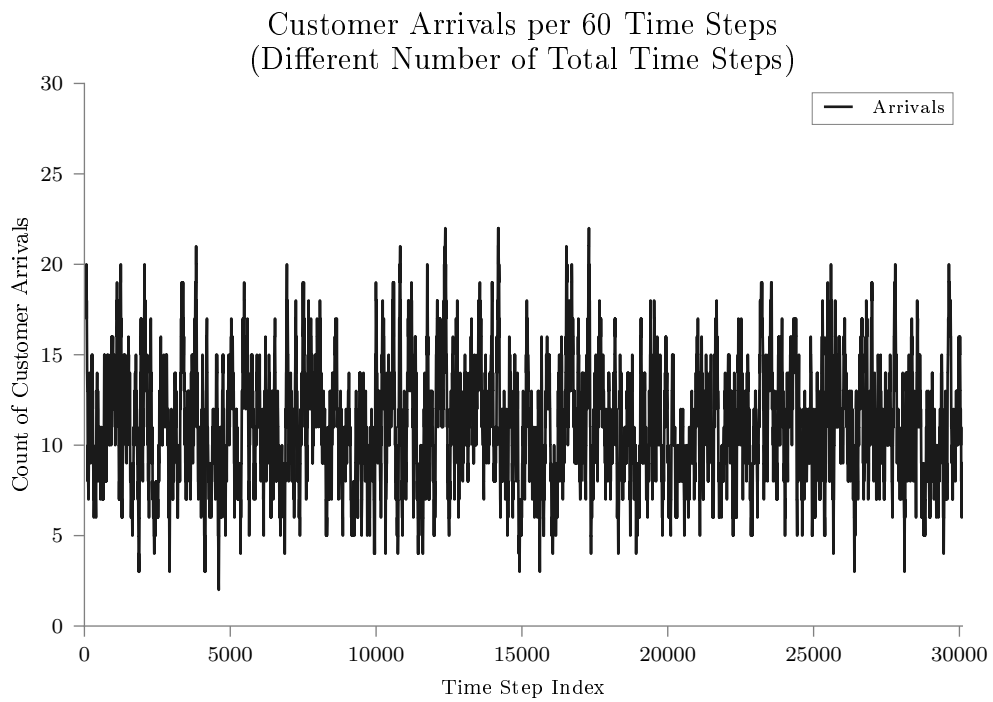Figure 4.2: Synthetic Customer Arrivals by Time (Same Time Steps).

Figure 4.3: Synthetic Customer Arrivals by Time (Different Time Steps).

4.2 Real Data

The data used in this thesis was obtained from a passenger rail operator. Nothing more about the source of the data can be disclosed, as the data is procured under a strict non-disclosure agreement. An overview of fields from the data used for this study, and the pre-processing methods used to clean the data before applying the EM algorithm are detailed in this section and other subsections below.

The data spans a period of two months and is a timestamped record of ticket purchases and the price paid for the tickets. The records contain booking details for journeys on various trains running between different combinations of origin and destination stations as well as departure times. After considerable deliberation, a well patronized origin and an equally well patronized destination serviced by a direct train were selected as the journey segment. The distance between the two stations is about 400 kilometers (250 miles), and the train travels at a high speed to cover the distance (including intermediate stops) in about 2.5 hours. The train competes with air travel between the selected origin-destination pair, and almost always runs filled to capacity.

The data is filtered to include only the journeys originating at the first selected station, and terminating at the destination station. From this filtered data, two journey dates and times are selected, one falling on a Friday, and the other falling on a Sunday. The tickets booked for departures on these dates are sub-selected from the filtered data, forming two independent tables. The following subsections will elaborate further on the data from each date. However, a few common attributes

are discussed below.

It is assumed that the train runs non-stop from the origin to the destination. No boarding or alighting is permitted, except at the origin and destination stations. This assumption is required because the EM algorithm derived by Talluri and van Ryzin applies to only a single leg of a journey. It is akin to imagining a direct non-stop flight between the origin and the destination stations. Dropping this assumption is part of planned future work. Strictly in this case, the assumption is not unrealistic, as less than 2% of the total train capacity boards or alights at intermediate stations.

The data has no socioeconomic or demographic information about the passengers. All that the data contains is time of booking, origin-destination, train number, departure and arrival times, number of tickets booked and price paid along with various company codes and auxiliary information. Even then, there are avenues to add increasing information in the choice models, and those will be dealt at a future time as well. As of now, only the class of ticket booked, the amount paid for the ticket per passenger, the time of booking and the departure time of the train are used.

The railway company uses variable number of fare classes depending on the day. The current method of pricing is not dynamic within a fare class; each of the fare classes have a fixed price, which varies by about 20 to 30 USD between classes. The availability of the classes changes per a fixed schedule, with lower priced classes closed earlier in the booking window. (Note that this system does not take advantage of any revenue management policies or algorithms developed over the years, it is just based on thumb rule.)

After filtering the data for the selected origin-destination pair and the date and time of departure (weekday and weekend both), discounted sales records from the data were removed. Discounted tickets are offered to passengers with army service records, rewards programs, disabilities or age related factors (the company policy dictates heavy discounts to students and the elderly). This was done to ensure passengers enjoying discounted tickets would not bias the results of the estimation. The records remaining are full fare transactions.

For all trains departing during weekdays, there are three main classes that customers can book tickets in. In this document those classes are called Classes A, B, and C. On weekends, two more classes are available, both cheaper and more restrictive than Class C called Classes D and E. In addition, tickets can be purchased at the station with the help of a clerk or an automated kiosk. These tickets are priced more than the most expensive class (class A), and are categorized in Class K. Except Class K, All the other fare classes are available for web-based purchase only. Note that no purchases were recorded in Class K for the the weekend departure considered in this study.

As a general rule, Classes K and A are available through out the booking period. Class B is closed 24 hours prior to departure. Class C is closed 7 days prior to departure, while the class D is closed 14 days prior to departure. In the data, Class E appears to be open through out the booking period, as there are multiple instances of very late bookings in Class E. It is not conclusive from the data whether they are the consequence of the class reappearing due to cancellations, or capacity not sold in other classes is offered in Class E, or both. Note that any class may be

36

closed earlier if capacity for that class fills up.

The most expensive class (Class A) is targeted at business travelers with amenities like in-seat dining, work tables and extra-reclining seats with more legroom. Tickets booked in Class A are also eligible for full monetary refund of the ticket amount if canceled anytime before the departure of the train, with less administrative charges levied at cancellation time. A ticket purchased at the station, under Class K also shares these amenities, but has a higher administrative charge for the services of the clerk.

Intermediate priced classes offer intermediate options, such as flexible cancellation policy with 100% refunds up to 24 hours before train departure, and reserved seats that recline less and have smaller leg room than the seats of the most expensive classes. Seats sold under the cheapest class are the same as those sold with intermediate prices, but unreserved, with seats available on a first come first serve basis, provided they are not reserved by intermediate fare passengers. The cheapest class of tickets are not eligible for a monetary refund upon cancellation, however a limited time travel voucher will be issued for the ticket amount.

During weekdays, Class C is the cheapest class ticket, Class B is the intermediate class. Over weekends, however, Classes B and C are intermediate price tickets, Class D is the cheapest class, while Class E falls in a special "web discount only" category. Class E probably is also used to sell tickets that were not sold in other classes after they are closed for bookings, and/or tickets that were canceled are offered for resale at last minute on Class E. Class E shares characteristics with Class D, and is unreserved, with a limited time travel voucher issued upon cancellation.

The following subsections discuss the findings from the cleaned and filtered data in greater detail.

4.2.1   Weekday

Weekday data is extracted from Friday April 24 for a train departing at 6 PM. This train has a good mix of business and leisure travelers looking to either return back to the central business district of the destination metropolis after conducting business all week, or looking to get away for the weekend. Note that both the origin and destination cities are huge tourist attractions, with very old and historic relics in addition to modern structures built to attract tourists.

Passengers traveling on Friday evening are mainly booked in 4 fare classes. The most expensive class, Class K, is only available for purchase at kiosks or through a clerk at the ticketing office in the station. Therefore, all bookings in Class K happen very close to the train departure. Class A is premier business class tickets, will the best amenities. Class B is half way between Class A and Class C. Class B is available to book up to 24 hours before departure. Class C is the cheapest fare, with the least amenities, and can only be booked 7 days before departure, or till available.

Figure 4.4 below shows the frequencies of bookings per hour for a period of time 30 days from train departure. It can be clearly seen in the figure that bookings peak about 2 to 4 hours prior to departure. During this time, Classes B and C are closed, and customers only have Classes A and K to book from. These bookings are
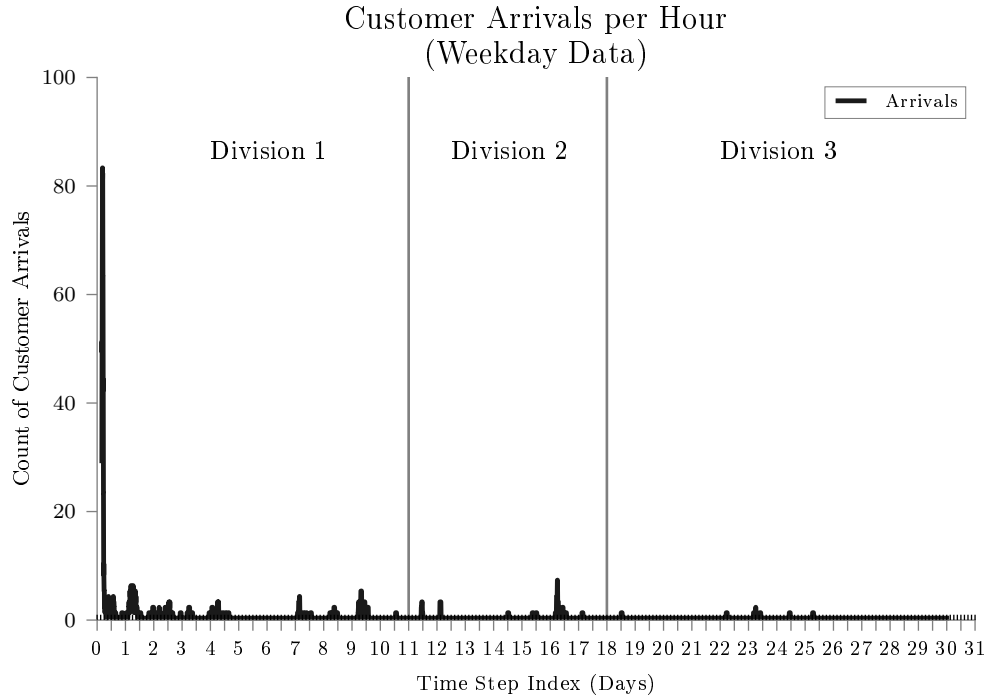
Figure 4.4: Frequency of Bookings per Hour (Weekday).

likely performed for corporate executive travel, and therefore are quite insensitive to higher fares. As an aside, note that there are only one or two bookings widely scattered beyond the 30 day window shown in figure 4.4.

In figure 4.5, the prices paid by the customers (converted to USD) are shown on the ordinate while the abscissa shows hours from departure. Zero on the X-axis denotes train departure. The length of the colored line shows the availability period of the respective fare class. Each of the colored markers on the lines represent a booking record for that fare class. The total number of booking records in each class are shown (in parenthesis) in the legend.

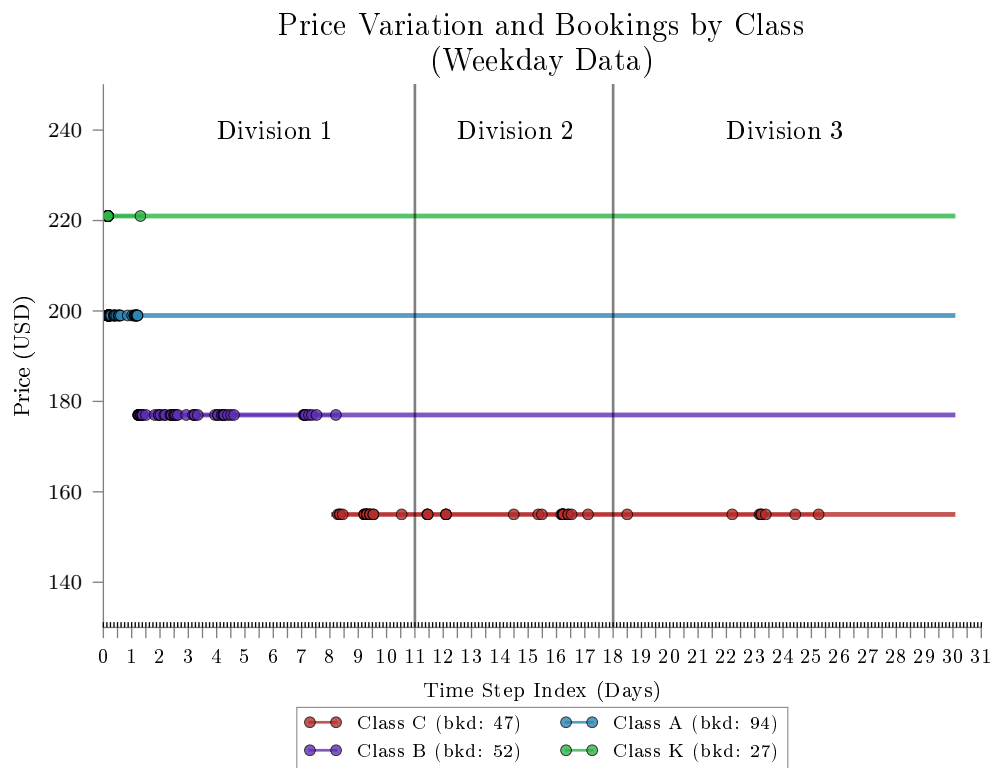Figure 4.5 also leads to an interesting finding: Class A does not have any

Figure 4.5: Booking Timestamps per Fare Class and Class Availability (Weekday).

bookings till Class B is almost closed, but parallel bookings on Class K can be observed. The reason for the former might be because passengers prefer the lower fare of Class B over the perks of Class A. This behavior is understandable on short journeys, where additional recline and leg-room are not as important. The few bookings of Class A that do overlap with Class B might be explained by high-paid executives traveling on business between the two metropolises.

The reason that Class K does not see bookings till very late into the booking period could be because people booking at the station with the assistance of a clerk are more likely to book the next available train, rather than a train in the future, especially when the fares on both trains are equal. Additionally, it is observed that Class A bookings cease about an hour before the last booking in Class K. This can be attributed to the access times to the train. The people further away from the station do not risk booking in Class A only to miss the train. But if they reach the station in time, they either use the kiosk or allow the clerk to book their tickets in Class K for a small extra amount. Further investigation reveals that the access times to the origin station on Friday evenings tend to be half an hour to an hour from the center of downtown serviced by that station. This finding bolsters the above surmise.

Bookings in Classes B and C are easily explained by the mindset of the normal passenger whose primary intent is to save money. There are no bookings in Class B till Class C is closed. Moreover, bookings in Class C occur at semi-random intervals (more bookings during daylight hours than night hours), indicating leisure travel booked well in advance to take advantage of the reduced fare. This leads

to the surmise that Class C is patronized by average income passengers, mostly on leisure, or semi-official trips (for example self-funded trips to conferences, or meetings arranged in advance).

Class B is more interesting than Class C, mainly because of the large gap in booking activity which lasts from about 5 days to 7 days from departure. This gap probably signifies that leisure booking ends about 7 days prior to departure, and business bookings begin about 5 days prior to departure. This conjecture is supported by the fact that leisure and cost-conscious travelers tend to avoid booking close to the departure date. Further, business schedules fall into place a few days before journey necessitate later bookings.

Since figures 4.5 and 4.4 provide complete information about the data, right from class availability periods, distribution of bookings and number of tickets sold, a data description table is not provided.

4.2.2 Weekend

To understand and process data from a weekend, where most travelers are traveling for leisure, the journey departing at 4 PM on Sunday, April 26 was selected. Bookings on the weekend service are concentrated mainly in 5 major fare classes, instead of 4 like the weekday service. Also, fares for each corresponding weekday class is lower on weekends. Classes A, B and C are similar to the classes found in the weekday data, with the addition of Classes D and E to weekend classes.

There are no bookings in class K, the kiosk class, and a smaller percentage of

the total bookings are in Class A. These can be easily attributed to lack of business travel on a Sunday evening. Most passengers are probably returning from the origin station after a weekend of leisure activities, such as sightseeing. The cheapest three classes have the combined maximum bookings, while classes B and A make up the rest of the total.

Bookings in Classes B does not start till Class C has closed, and bookings in Class C does not start till classes D and E are closed. The inexplicable thing about the D class is that it sees simultaneous bookings with Class E, probably by richer leisure traveler. It also prompts the idea that there must be some difference in the amenities of Classes D and E to justify paying more for Class D ticket. It is plausible that Class D offers reserved seats, or window seats, or other such advantages, or it may be a lower fare substitute for Class B.

It may well be that the fare of Class E is bumped up at 14 days prior to departure to the price of Class C, and Class D fare is increased to the level of Class B around the same time. Classes E and C will then share amenities with each other, just like Classes B and D (for example, refer to synthetic data and figure 4.1). However, there is no indication in the company policy or in the data to suggest this association, and therefore for this study they are treated as separate classes.

Figure 4.6 is similar to figure 4.4 above, only for weekend data. It shows that the bookings per hour for the weekend data are more evenly distributed than the weekday data. Further, it also shows that customers prefer to book early, and take advantage of lower fare classes, even though a peak before departure is observed.

Similarly, figure 4.7 shows the booking frequencies and availability periods for
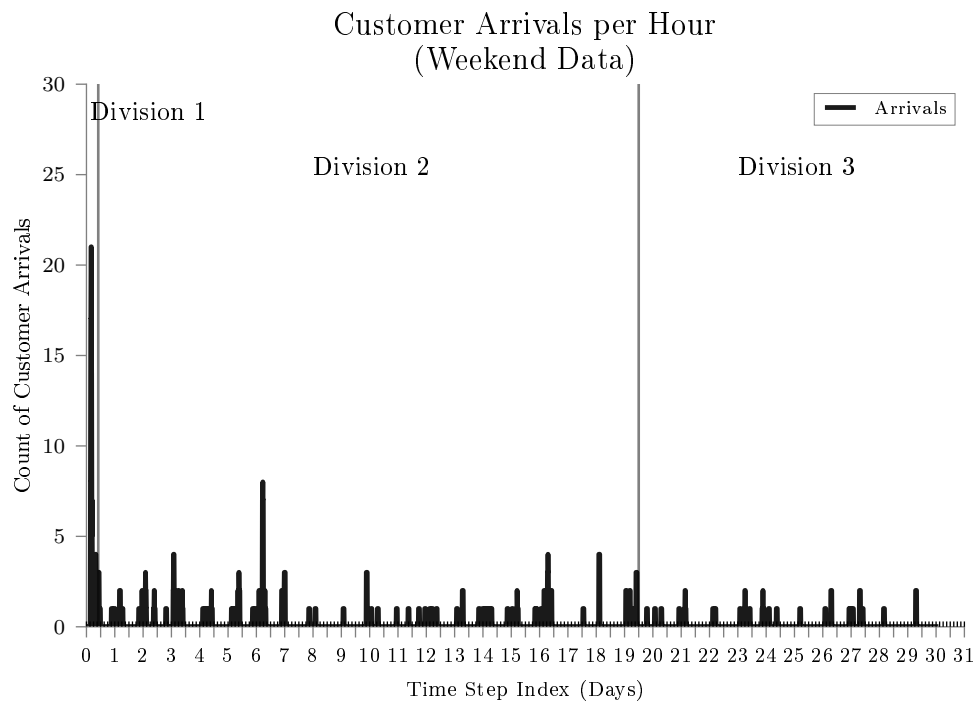
43

Figure 4.6: Frequency of Bookings per Hour (Weekend).
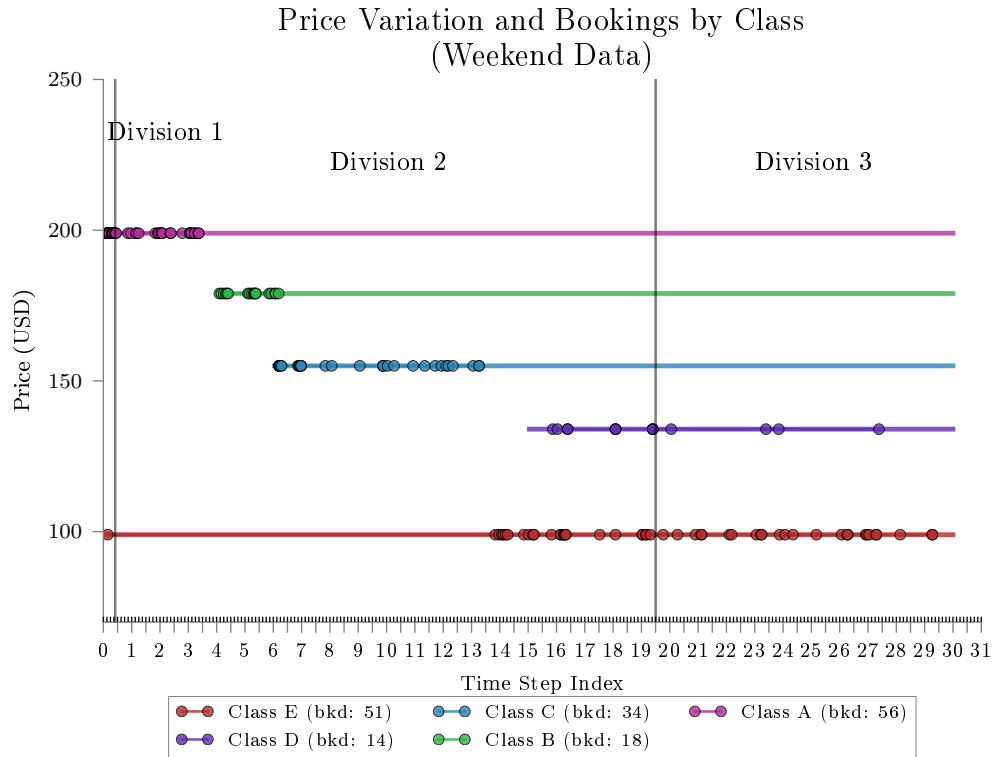
Figure 4.7: Booking Timestamps per Fare Class and Class Availability (Weekend).

all classes. The length of the lines represent availability periods, while the markers on them represent the time when a booking was done. Number of purchases in each fare class is shown in the legend against the fare class (in parenthesis).

It is clear from figure 4.7 that there is still some last minute bookings in class A. These are probably corporate bookings, by customers with early Monday morning meetings to attend.

4.3   Data Manipulation

The only real manipulation of the data is perturbing the timestamps so that the one arrival per time step assumption is honored. Other than that, the data is only partitioned and some cleaning is applied to ensure the data is uniform and unbiased. The latter two are explained in the following subsections.

4.3.1   Expanding

The records that represent the time stamps when a customer did not arrive, or when a customer arrived but did not book the ticket need to be imputed in the data set. Note that the weekday data has 220 observations while the weekend data has 173 observations. Essentially, expanding should ensure that each booking is placed in a unique time step. Referring to figures 4.4 and 4.6, it is directly inferred that weekday will need a time granularity of one second or less, while weekend can use a larger granularity of 30 seconds.

Based on the above observation, the weekday data was expanded using 1 second granularity to a total of 2,581,200 records (from three hours before train departure, when all bookings cease to 30 days prior to departure). Weekend data was expanded using 30 second intervals to 86,280 records.

4.3.2   Partitioned Expanding

The above subsection neatly illustrates that the expanded dataset is very huge. Moreover, it is an overkill for the periods when there are lean bookings, as it does

not justify the fine granularity of the time steps. In the light of these arguments, it was decided to vary the time step size. However since the time steps are assumed to be monotonic, and all mathematics based on that assumption, having variable time steps is not permissible. Therefore the data was partitioned such that a monotonic time step could be assumed within each partition, and because a separate model was fit to each partition, the time step sizes across the partitions can be varied.

The motivation for partitioning is obtained from the observation that very fine time granularity, required close to the departure, is not required earlier in the booking period. The data was therefore sliced into three parts, the first one (closest to departure) with time granularity of one second for weekday and 30 seconds for weekend, the middle partition with a time granularity of 5 minutes during weekdays and 10 minutes during weekends. The last partition, furthest from departure with a time granularity of 30 minutes.

Figures 4.4, 4.5, 4.6 and 4.7 already show the partitions. As can be seen from figures 4.4 and 4.5, the first weekday partition lasts from the departure to 11 days before departure, the second partition lasts from 11 to 18 days before departure, while the third partition covers the time periods from 18 to 30 days before departure. Similarly figures 4.6 and 4.7 show that the first division of weekend data lasts from train departure to 10 hours before departure, division two spans from 10 hours to 19 days and 12 hours before departure, while the rest of the booking space is taken up by division 3.

Partitioning and taking different sizes of time steps is justified because the data and human behavior does not indicate that people have visited the system

frequently, only to not purchase a ticket. Therefore, in periods with low arrival rates, larger windows of arrival times are permissible. It is loosely assumed that purchase rate is directly proportional to arrival rate. In addition, the way the model is set up, there is no clear indication of which time steps no customer arrived versus the time steps when someone arrived but did not purchase. Both of arrival probability and purchase choice can only be given individual probabilities, which applies uniformly to all intervals of time. Therefore having many more intervals will bias the model into thinking that the proportion of no-buy decisions are a lot higher than buy decisions.

4.3.3 Perturbing

Timestamps reported in the source dataset are capped at the nearest minute. Therefore, to satisfy the assumption of only one arrival per time stamp, the data had to be slightly modified. After expanding the data, either within partitions or without, the overlapping records were slightly shifted to occupy nearby time steps. It is like gently shaking a table with holes to ensure all marbles are properly seated. Since the time steps are much finer and more numerous than the number of bookings per minute, no booking record was shifted by so much that it presented in another minute.

Note that the above holds proportionately. That is, for records belonging in divisions 2 and 3 in both datasets, the shift is by the smallest number which will place overlapping records in adjacent time steps. Since arrivals and bookings are

sparse in these time steps, it does not cause any impact on the distribution of the data.

It is ensured that the distribution of the data does not change significantly. In the weekday data however, very close to the train departure when a horde of tickets are booked very close to each other, slight change in the statistics of the manipulated data can be observed. The difference is so small, however, that we do not expect there to be any impact on the final results.

4.3.4 Cleaning

Cleaning of the dataset merely involves manipulating certain special situations. For example, the system records multiple booked tickets under just one record. The price paid by the customer is a multiple of the number of tickets booked, and price per ticket. Such records are converted to price per ticket, to make the utilities similar to those customers who booked just one ticket. The reasoning is that customers decide based on the price of just one ticket.

Also, the company policy dictates that children under 12 years get a 50% discount on their tickets, up to two kids per booking family. Therefore, a person booking two tickets, one for an adult and another for a child will pay 75% of the ticket price per ticket; or a person booking three tickets, two children and one adult, will pay 66.67% of the price per ticket. With the 2 child tickets per booking limit, 66.67% of ticket price per ticket is the lowest price a person can pay for multiple tickets.

Applying this logic, for multiple ticket purchases a simple check will reveal if tickets for child passengers were booked. Child tickets are then reset to full fare, to keep the final price paid per ticket the same for all passengers in the model. The reasoning is again that children are mandatory companions of an adult, and therefore the purchase decision will be mainly based on the price adults pay for themselves.

At the end of cleaning, each record will have only one purchase transaction, and the fare amount paid will be reset to the base fare of the class the ticket is booked in.

4.3.5   Model Coefficients

The first experiments were conducted with a universal coefficient for all classes, similar to the synthetic data case. However, the real data model was not stable with a single coefficient. Therefore, a unique coefficient for each choice set was used. Not only did model stability improve, but also having a coefficient for each choice available with the customers offers better insight into the customer behavior. It explains where customers are willing to spend money, and if there is any customer bias. For example, if leisure travelers are few and far in between, it would make more sense to increase the prices much sooner into the booking period, rather than waiting till the last few hours, and vice-versa.

Chapter 5:   Method

The method used in this study is remarkably simple to understand. The first step is to fill in the data for the missing alternative: the no purchase choice. This is accomplished under an Expectation Maximization framework, where the coefficients of the logit model are iteratively fitted. These parameters are then used in the revenue management step where a Bellman equation is solved in order to select the appropriate price point. The following subsections in this section will further elaborate on the method. The first subsection introduces the variables and explains what they stand for. The next subsection details the Bayes's theorem that forms the heart of the EM algorithm. Third subsection shows the multinomial logit, while the fourth ties them up together in the EM algorithm

5.1   Variable Declaration

The notations used in this paper are as follows:

- β is a vector of coefficients for model parameters. $\hat{\beta}$ is its estimate, computed by the model.

- X are the model parameters, costs of ticket in each fare class.

- t is the time index from departure of train. $t = 0$ is considered as the train departure. t is monotonic and negatively indexed in the remaining time till departure. Only one customer arrival is permitted per t, at most.

- T is a set of all time periods t.

- λ is the mean of the arrival probability in t. $\hat{\lambda}$ is its estimate which is obtained as the total arrivals divided by the total number of periods (see Equation 5.1).

- $a(t)$ is the probability of arrival in time t. It is 1 when a confirmed booking has taken place in t, otherwise it is a distribution obtained by Bayes Theorem (see Equation 5.3). When confirmed booking has not taken place, the corresponding $a(t)$ is initialized to zero.

- D is a set of periods in T where a booking has been recorded. $\bar{D}$ is its inverse, the periods in T when a booking was not recorded.

- P_j is the probability of choosing j, which can be any of the offered fare classes, or the choice to not buy.

$\hat{\lambda}$ is given by the equation 5.1 below.

$$\hat{\lambda} = \frac{\sum a(t)_{\in D} + \sum a(t)_{\in \bar{D}}}{T} \tag{5.1}$$

This equation is nothing but the sum of all arrivals divided by the total number of time periods T. When confirmed bookings have been made, $a(t)$ is 1. When bookings are not confirmed, $a(t)$ takes a value from the probability distribution obtained using Bayes theorem (see Equation 5.3). At the start, when $a(t)_{\in \bar{D}}$ is

initialized to 0 (practically, very close to 0) this is simply the ratio of all confirmed purchases to the total periods. This forms our prior distribution for the Bayesian analysis further on.

5.2 Bayes' Theorem

The model relies on the Bayes theorem [25]. This theorem is useful when combining probability distributions to get a conditional distribution. In this case, the distributions that need to be combined are the arrival probability of customers to purchase tickets and the no-purchase probability given by the logit model. Equation 5.2 below provides the generic form of the Bayes theorem.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{5.2}$$

In this equation, $P(A|B)$ is the conditional probability of hypothesis A, conditioned on observations B. The probability $P(B|A)$ is the likelihood of the observations B given the hypothesis A. This can be understood as a guess about the observations needed to achieve the prior belief in the outcome of the trails. P(A) is the prior belief, and P(B) is to normalize the equation.

An example will make it easier to understand the equation. Consider the case of swatting a fly with a rolled up newspaper. The probability of successfully swatting the fly is conditioned on the observation of its flight or reaction as the newspaper approaches. This probability is given by $P(A|B)$ where A is the success probability of the swatting, and B is the observation about the reaction of the fly.

The Bayes theorem can be used to compute the probability of a successful swatting, conditioned on the flys reaction without actually swatting the fly.

Based on past successes in the act of swatting a fly, the necessary probability distribution of the flys reaction can be established. This is the conditional probability $P(B|A)$. It basically means that if the fly has a reaction close to the mean of B, then swatting will be successful. If the fly is jumpy, and so its reaction is further from the mean of B, the swatting will fail. This probability is purely obtained from past experiences.

The probability $P(A)$ is the independent probability that the fly will be swatted. This probability does not account for the flys reactions. It can be thought of as the aggregate probability obtained from repeated swatting of flies, by dividing the successful swatting with all attempted swatting. Since this probability is not known, usually a prior assumption is made, which is updated with each successive trial. $P(B)$ normalizes the above product, and is the byproduct of transforming a joint distribution to a conditional distribution. Readers interested in this further are requested to peruse literature on Bayes' identity.

In the present context we are attempting to estimate the probability of customer arrival to the ticketing system conditioned on the probability of no-purchase. The arrival probability of the customers is masked because no-purchase transactions are not observed. In this context $P(A|B)$ is the probability that a customer will arrive, conditioned on the observed no-purchase probability. The probability of not buying can be obtained from the arrival process of customers, and is given by $P(B|A)$. In this context, $P(A)$ is the probability of customer arrival; a prior

assumption of which is made using the observed purchase decisions, and the total number of periods ($\hat{\lambda}$). While $P(B)$ is the probability of not purchasing.

In the notations used in this paper, as described above, the Bayesian equation is given in equation 5.3 below.

$$\hat{a}(t)_{\in \bar{D}} = \frac{\hat{\lambda} \cdot P_0(\hat{\beta}, X_{\in \bar{D}})}{\hat{\lambda} \cdot P_0(\hat{\beta}, X_{\in \bar{D}}) + (1 - \hat{\lambda})} \qquad (5.3)$$

Here, is the final conditional probability $P(A|B)$, the probability that a customer arrived to the system, given the no-purchase decision probability we observe from the data. $\hat{\lambda}$ is the estimate of the arrival process of the customers ($P(A)$), and is the probability of no-buy ($P(B)$). The denominator normalizes it by appropriately accounting for the probability of not arriving $(1 - \hat{\lambda})$, and the probability of arriving and not buying $(\hat{\lambda} P_0(\hat{\beta}, X_{\in \bar{D}}))$.

5.3   Multinomial Logit

The multinomial logit (MNL) model used in this study is constructed to be very simple. For the synthetic data, there is only one model coefficient in the MNL, as the data was synthetic with only one coefficient. The real data, however, was stable only when coefficients for each choice were independently formulated. The model otherwise is a simple case of the MNL, with very straightforward formulation. Equation 5.4 gives the simple formula for computing the probability of choice using MNL.

$$P_i(\beta, X) = \frac{e^{U_i}}{\sum_{j=1}^{S} e^{U_j}} \tag{5.4}$$

Here U stands for the utility of choosing alternative i. The denominator is the sum of the utilities for each alternative in the alternative set S. In the context of this thesis, S comprises of all fare classes and the option to not buy (5 choices for the synthetic data and weekday data, 6 for weekend data). The utility of not buying is assumed, without loss of generality, as zero. The utilities of other choices are given by the generic Equation 5.5 below.

$$U_j = \beta_j \cdot X_j + \epsilon \tag{5.5}$$

Equation 5.4 gives the probability of a customer choosing alternative j. Equation 5.5 gives the utility obtained (or lost) by choosing alternative j. Utilities for each choice available with the customer is computed, and the choice with the maximum utility is chosen by the customer. Therefore, there is an equation 5.5 for each alternative j. However, there may not be a distinct β_j for each alternative (as in the case of the synthetic data, there is only one β for all alternatives). X_j denotes the cost of choosing alternative j. In all cases presented here, X is merely the cost of the ticket. For no purchase alternative, X is 0, and therefore equation 5.5 evaluates to 0.

The individual chooses the alternative that gives them the maximum utility. The alternative that maximizes the utility also maximizes the probability in equation 5.4, thus ensuring that choosing the alternative with maximum utility is the most

probable.

5.3.1   Multinomial Logistic Regression

The MNL model presented above needs to be worked backwards to obtain the values of the coefficients. The data contains the cost per ticket for each time step (X), and the choice made by the passenger. Therefore, to obtain the coefficients, a logistic regression is performed on the choices. Regression in this thesis is performed using an optimizer that finds the minimum value of a function. The solver used in the minimization is the Limited Memory Broyden–Fletcher–Goldfarb–Shanno with Box constraints (L-BFGS-B) algorithm. It belongs to the quasi-Newton family of methods.

L-BFGS-B is desirable because it uses a limited amount of computer memory and the solver requires bounds to be set on all estimation parameters. Both of these help limit the search space, making the convergence faster, and use less computer resources. Additionally, the MNL function may have other minima outside the search space, and therefore, if unconstrained, it might converge at a different but erroneous location.

The solver requires a single function value to be returned from the function being optimized. The traditional method of fitting a logistic regression involves a log-likelihood (LL). The LL makes it easier to solve the problem, because the likelihood of equation 5.4 needs the solution of a product. The logarithm converts all the product components into sum components, and then applies to the exponent in

each component, reducing them to the linear form of equation 5.5. The return value to the solver is the sum over all rows of this log of the probabilities, as demonstrated in equation 5.6

$$\sum_{t=0}^{T} \ln(P_{i(t)}) \tag{5.6}$$

Where t denotes the index of the time step, and i is the index of the choice made in the said time step.

5.4 EM Algorithm

The algorithm involves recursive optimization till the estimated coefficients attain a constant value. The equation of the objective function of the algorithm is given in equation 5.7 below.

$$\max_{\hat{\beta}} \left\{ \sum_{t \in D} a(t) \cdot \ln \left(P_{i(t)}(\hat{\beta}, X(t)) \right) + \sum_{t \in \bar{D}} \hat{a}(t) \cdot \ln \left(P_0(\hat{\beta}, X(t)) \right) \right\} \tag{5.7}$$

Equation 5.7 essentially consists of two parts. For those time steps where a booking is recorded ($t \in D$), the probability of MNL model is used to fit β as usual. That comprises the first part of the equation. The second part deals with those time steps when a booking has not been recorded. In these time steps($t \in \bar{D}$), the no-purchase probability is the product of the arrival probability $\hat{a}(t)$ and the natural logarithm of the purchase probability (logarithm because MNL equations are fit using their loglikelihood). The first part also has the arrival probability component $a(t)$ but that can be safely ignored because its value is always 1.

The final probability computed in the EM algorithm is the marginal probability of the arrival probability and the probability of choice. The $\hat{\beta}$ that maximizes this probability is the desired coefficient. Equation 5.7 is iteratively executed, and the results of each data is stored for the next iteration till convergence is achieved.

Solving equation 5.7 yields an estimate of β. This estimate is used in equations 5.5 and 5.4 to compute the no-purchase probability for the time periods when confirmed bookings have not occurred. This no-purchase probability is used in equation 5.3 to estimate the arrival probability during periods of no buy. The arrival probability estimate $\hat{a}(t)$ is used in equation 5.1 to better our guess about the overall arrival probability, and in equation 5.7 to get a better guess of $\hat{\beta}$ in the next iteration.

Note that for the first iteration, an initial value of $\hat{a}(t)$ must be supplied to get a preliminary estimate of β. This initial value, for the purposes of this thesis is set to 0.0001. The final algorithm has 3 major steps (1-3; 0 is the initialization step, and will be run once only). They are enumerated below:

0. Assume prior distributions for $\hat{a}(t)$: Set $a(t) = 1$ when a confirmed booking has been made, and 0 otherwise. Use this assumption to estimate the initial value of λ using equation 5.1.

1. **Solve for $\hat{\beta}$:** Use equation 5.7 to find the best estimate of β using the current values of $a(t)$ and $\hat{\lambda}$.

2. **Update the value of $\hat{a}(t)$:** Using the obtained value of $\hat{\beta}$ in equation 5.3, compute $\hat{a}(t)$ for time steps when confirmed booking has not been made. Also

update $\hat{\lambda}$ using equation 5.1.

3. **Convergence check:** Check for convergence in the values of $\hat{\beta}$ and $\hat{\lambda}$. If converged, stop, else repeat from step 1.

The final converged values from this algorithm can be used in forecasting customer arrival patterns and behavior.

5.5   Revenue Management

Using the coefficients obtained from above, and the MNL model fitted conse-quently, a Bellman equation can be used to take the optimal decision for the current time step. Solving the Bellman equation using a dynamic program will yield the most optimal control decisions for the give time step. In essence the equation looks at all the future time steps in the horizon and determines the total sale and revenue potential. Based on that potential, a price is offered at the current time step for the customer to choose.

However, the dynamic program was not implemented in this study, because the model coefficients from the real data are neither sensible nor reliable.

Chapter 6:   Results and Discussions

In this section, the results obtained by applying the EM algorithm to the data are presented. The problem statement being addressed has a degree of freedom and an identity problem both of which cannot be solved by just an EM algorithm. The algorithm needs either external information, or strong assumptions that are accurate.

If one keeps a track of and knows that a total of N people have arrived at the system, of which M purchased the tickets, then the purchase probability is given by $p = M/N$. However in the real world, neither N nor p is known, and M is the only known. It is evident that a range of values of N and p can yield the observed value of M. Therefore, this method has a degree of freedom that results in unstable solutions.

Identification problem arises because the time steps with no arrivals cannot be distinguished from time steps with no purchases. The EM framework presented is constructed to require a fixed number of time steps. It is also assumed that no more than one customer arrives in a given time step. These two requirements fix the total maximum population size of the system. Unless, the total number of time steps is accurately close to the real population size, the time steps when no arrivals occur

cannot be differentiated from the time steps when no purchases are made. In the case when the number of time steps is accurately close to the real population size, there are no periods when customers do not arrive, and the identifiability problem collapses.

This chapter first presents the results from the synthetic data, which forms a foundation for the results from the real data. Only results from the EM algorithm and model fitting are presented. The dynamic program has not been solved, and no results are presented that involve the final revenue management steps. Due to the two issues mentioned above, the results from the model fitting are not neither accurate nor reliable. Therefore, no correct results from the dynamic program can be expected.

All results below show the value of the coefficients (including the arrival probability, λ) and the function optimal value at each iteration. The iteration number is shown on the abscissa, while the values are shown on the ordinate. Legend contains the final value of the coefficients in parenthesis.

6.1 Synthetic Data

The results obtained by applying the method to the synthetic data are explained in the following sub-sections. Initially, the case where the dataset is expanded to the correct number of time steps is presented, and it is shown that the model works. Then the results from the case when a different number of time steps is used is shown, and the degree of freedom of the model is discussed. This discussion

forms a stepping stone to discuss the results from the real data.

6.1.1   Same Time Steps

After expanding the synthetic dataset to 30,000 records, the algorithm is applied. Since the time steps is kept the same as the original the results obtained are favorable and accurate. The plots in figure 6.1 show the value of the parameters ($\hat{\beta}$ and $\hat{\lambda}$) and the function on the ordinate against each run of the EM algorithm on the abscissa. The left subplot contains the coefficient of the MNL model ($\hat{\beta}$) and the estimated arrival rate $\hat{\lambda}$, while the function value is shown in the right hand side subplot. The final values of the parameters and the objective function are shown (in parenthesis) in the legend.

The figure beautifully shows convergence at about 80 iterations of the EM algorithm, and the initial parameters are fairly accurately recovered. The function value is also shown to be steadily decreasing, which is what one expects as the fit improves. This result leaves no doubt that the EM algorithm works as expected. However, this result demonstrates a very special case, and not a general rule. The special case being that the data passed to the EM algorithm has the exact same number of time steps as the original synthetic data, collapsing the identifiability problem.

Changing the number of time steps, however, produces inaccurate results, as demonstrated in the following subsection.
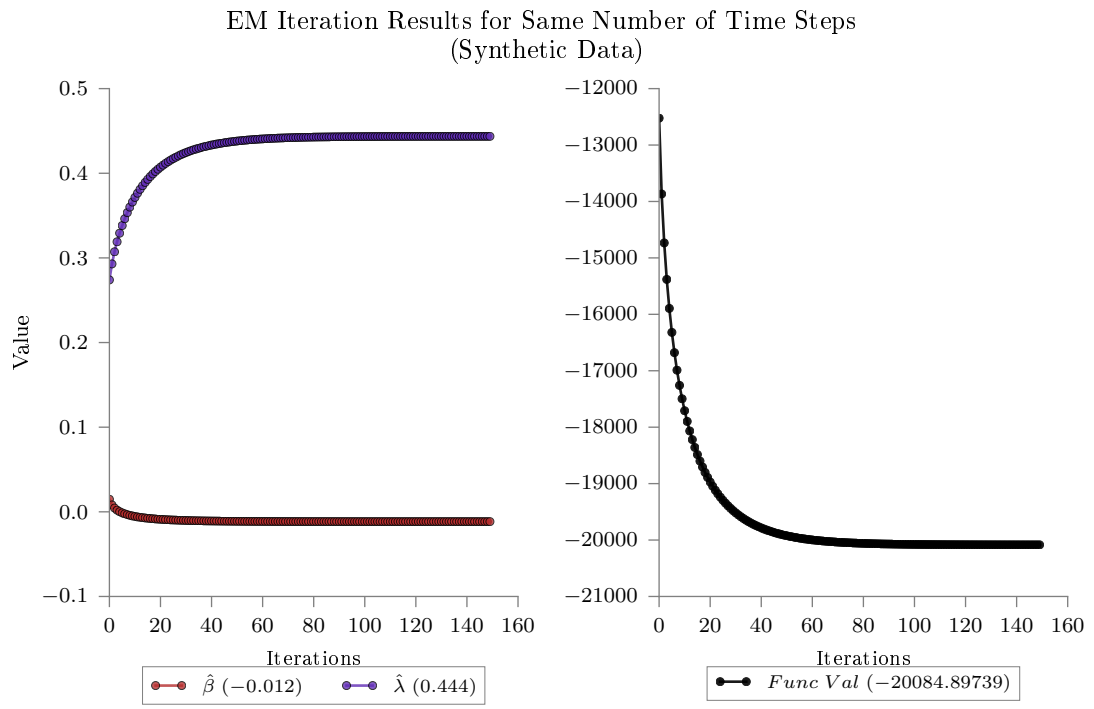
Figure 6.1: Synthetic Data Results with Same Time Steps

6.1.2 Different Time Steps

Using the same synthetic choice probabilities, but different number of time steps is similar to not knowing the conditions of the original synthetic data. The synthetic data was produced by assuming a total time period of 30,000 seconds. However, if an analyst does not know that, and has to pick a number of time steps, one may pick any arbitrary number. For illustration, let us assume 50% more time steps.

Instead of 30,000 time steps, we now have 45,000. The placement of the additional 15,000 time steps is uniformly distributed in the data. Obviously, no purchases are recorded during any of these additional time steps, and thus they are all either no-buy or no-arrive. Figure 6.2 below shows the resulting coefficients in a similar format as figure 6.1 above. It is no surprise that the results are different, and the coefficients have not been recovered as before. For all intents and purposes, these results are meaningless and should be discarded.

The reason behind the incorrect results, however is of academic importance and is the single most significant contribution of this thesis. The issue is caused by assuming one arrival per time step. This in essence fixes the total population size of people that can arrive at the system. If, in reality, A people arrive at the system in a T total periods, of which p purchase tickets, the purchase probability is given by $P(p) = p/A$, and arrival probability is given by $P(\lambda) = A/T$. In the first equation, only p is known to the analyst (it is obtained directly from the data, by counting the number of purchases in the dataset). In the second nothing is known to the
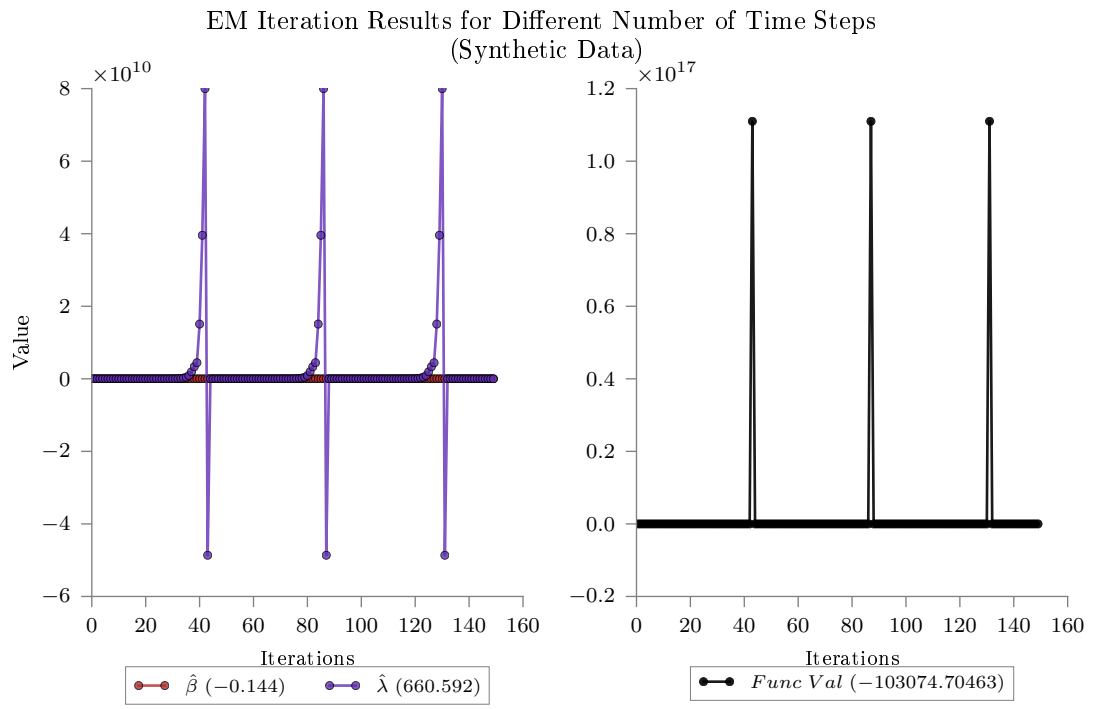
Figure 6.2: Synthetic Data Results with Same Time Steps

analyst.

It is also clear that the purchase probability $(P(p))$ is related to the arrival probability (λ) and the utility of purchase given by the choice model. The reason the EM algorithm fails is because it deals with two nested indeterminacies: it can neither identify no-purchases from no-arrivals nor control the degree of freedom. From just p, it attempts to compute A and λ. However, in this system of equations, knowledge of one other variable will fix the values of the other unknowns. In this case that variable is T. When the value of T is accurately set to the original value, everything works fine, and λ is what one would expect. Missing the value of T even by a little bit, however, produces incorrect values for λ (unless one is very lucky) and therefore, useless coefficients of the choice model.

A simple extrapolation of this discussion explains why the EM algorithm fails horribly at recovering the parameters from the real data. In the real data, it is next to impossible to guess the correct value of T, unless additional information is supplied.

6.2   Real Data

As discussed above, application of the EM algorithm to real data is not useful unless additional information informs one of the true value of T. However, for the sake of completeness, the results from real data are presented below.

6.2.1 Weekday

The results from applying the EM algorithm to the Weekday data is presented in this subsection. First the results from the complete data is presented, then the results from the partitioned data. It is seen that the EM algorithm does not converge, and coefficients quickly arrive at the bounds. Removing the bounds causes unstable model that fails to produce results. Also the degree of freedom issue is explored.

6.2.1.1 Complete Data

In this section the results from unpartitioned data are presented. The time granularity is on one second in this case, and a total of 2.58 million time steps are used in the model. The results are presented in figure 6.3 below. The figure is similar in format to the other figures above in this chapter.

Note that the coefficient of β_K is at bounds, and $\hat{\lambda}$ is 0. Also, the function value is increasing with each iteration instead of decreasing. This is due to the identifiability issue of the algorithm. Out of all 2.58 million time steps, it cannot identify which time steps were no-arrivals and which ones were no-purchases. Also, the oscillation of the coefficients does not affect the value of the objective function, which remains stationary. This is a classic case of freedom available to the model.

EM Iteration Results from Complete Data
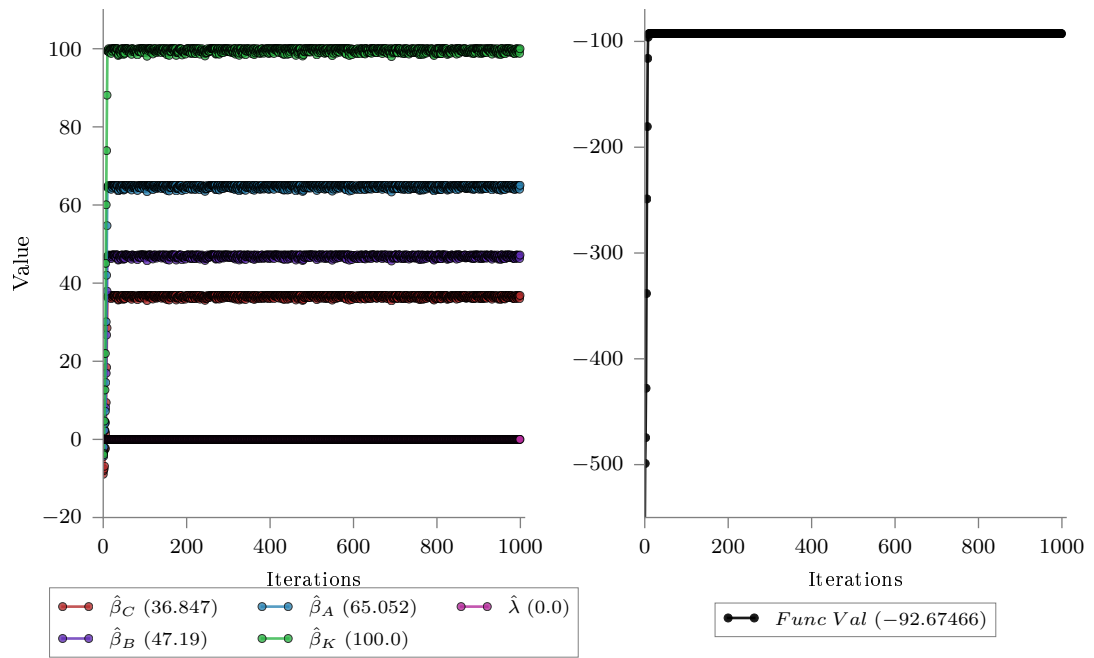(Weekday Data)

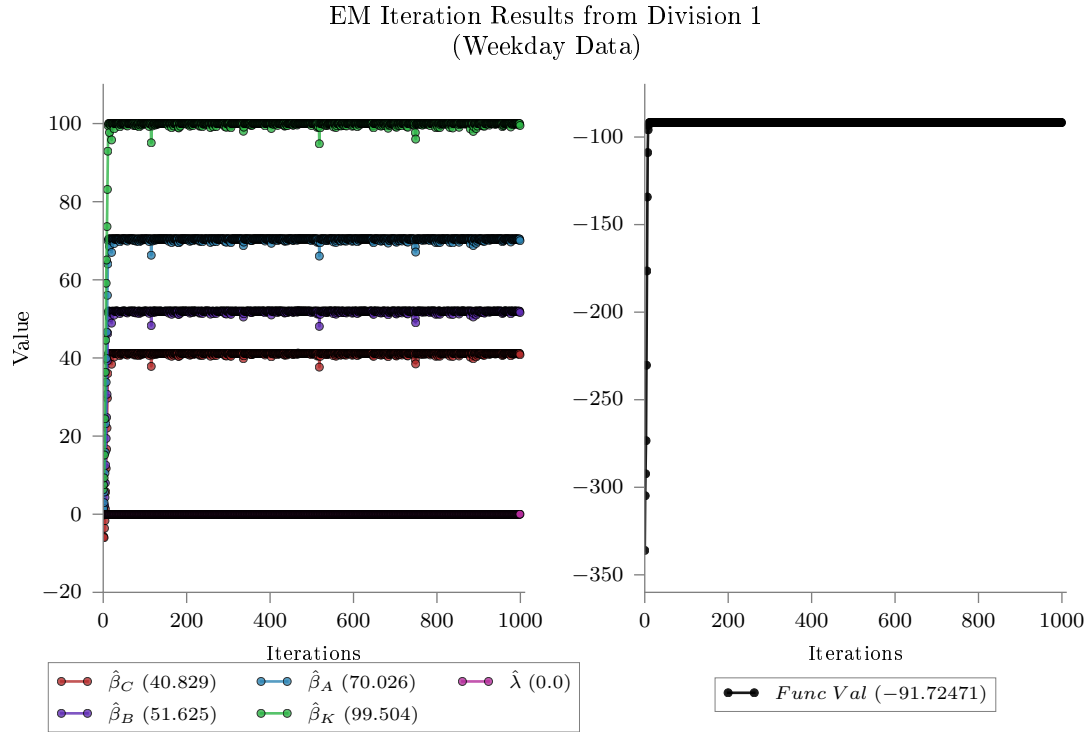Figure 6.3: Weekday Results Complete Data

EM Iteration Results from Division 1
(Weekday Data)

Figure 6.4: Weekday Results Division 1

6.2.1.2 Partitioned Data

The results from the partitioned data are presented below. Results from divisions 1, 2 and 3 are presented in figures 6.4, 6.5, and 6.6, respectively. Figure 6.4 is quite similar to figure 6.3, from above. This is because the data in division 1 has been expanded to 939,600 time steps, which is again far more than the number of people expected to arrive at the system. Therefore it suffers mainly from inability to identify the time steps.

Figures 6.5 and 6.6 on the other hand look much different from figures 6.4 or 6.3. The reason for this is that the data is not expanded to a very large number of

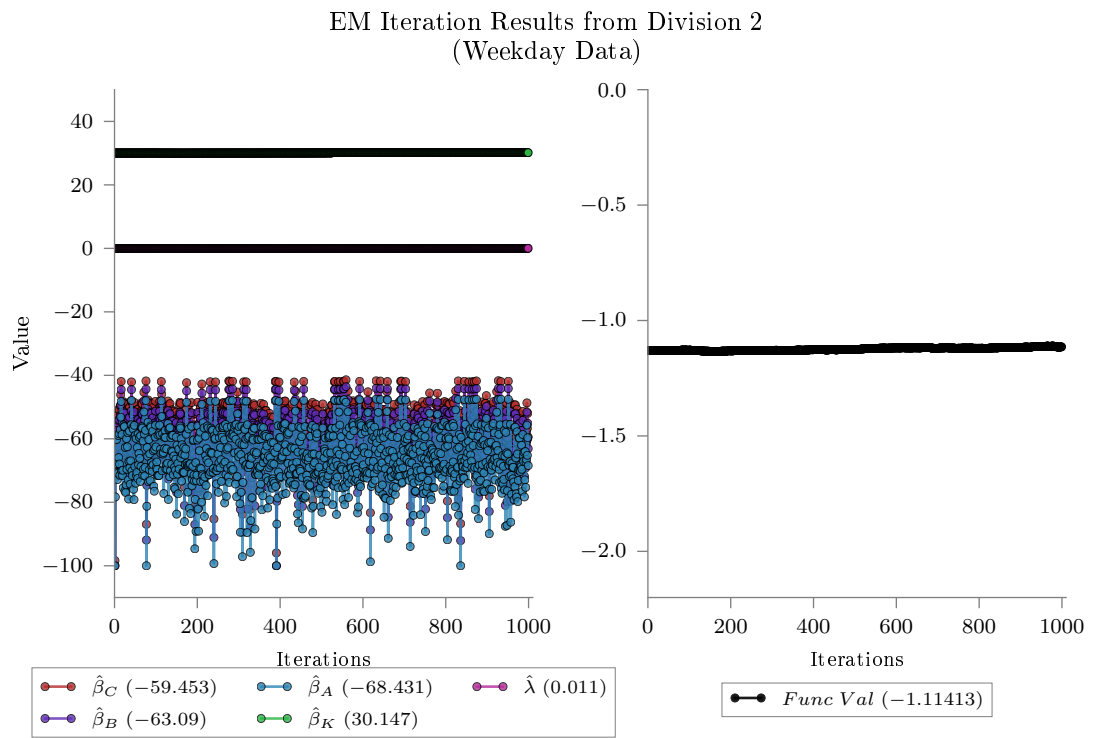EM Iteration Results from Division 2
(Weekday Data)

$\hat{\beta}_C$ (−59.453) $\hat{\beta}_A$ (−68.431) $\hat{\lambda}$ (0.011)

$\hat{\beta}_B$ (−63.09) $\hat{\beta}_K$ (30.147)

$Func\ Val$ (−1.11413)

Figure 6.5: Weekday Results Division 2

EM Iteration Results from Division 3
(Weekday Data)

$\hat{\beta}_C$ (−39.01) $\hat{\beta}_A$ (−44.304) $\hat{\lambda}$ (0.014)
$\hat{\beta}_B$ (−41.114) $\hat{\beta}_K$ (31.735)
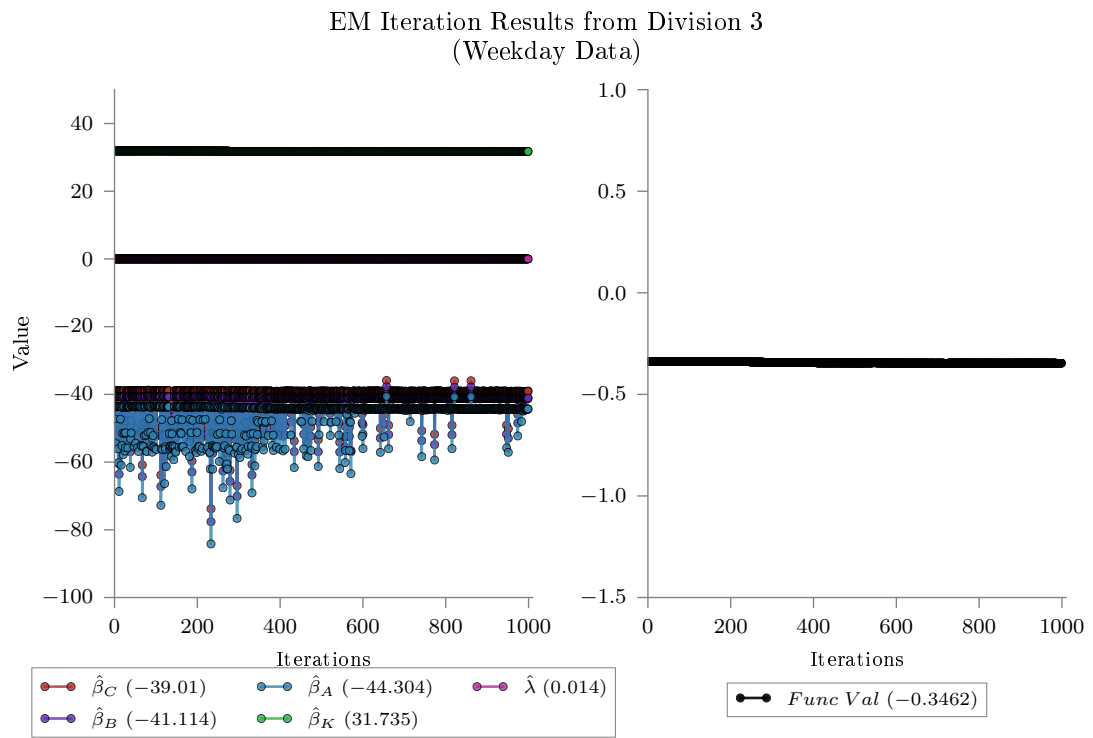
$Func\ Val$ (−0.3462)

Figure 6.6: Weekday Results Division 3

time steps (2,016 and 576 time steps respectively), and therefore the identity issue is minimized. However, the coefficients appear to be dancing all over the place, with wildly different numbers from one iteration to another. This jumping about is a direct consequence of the degree of freedom. Note that the objective function is constant, even though the coefficient values are not, because multiple values of the coefficient can result in the same objective function value. The extra degree of freedom consequently results in the EM never converging.

6.2.2   Weekend

Similar to the results from the weekday data above, the results from the weekend data are presented in this section. The results from the complete data are presented first, then the results from each of the partitioned datasets are shown. Weekend data results are also similar to the weekday data results, and hence the text is kept to a minimum to avoid repetition.

6.2.2.1   Complete Data

The results from complete data are shown in figure 6.7. Note the coefficient at the bounds and the general waviness caused by the identifiability and the degree of freedom respectively. It is interesting to note that at about 650 iterations, the coefficients have dipped to almost half their stationary values, while the objective function has barely dipped by 2 points, proving the adverse effects of the degree of freedom in the formulation.

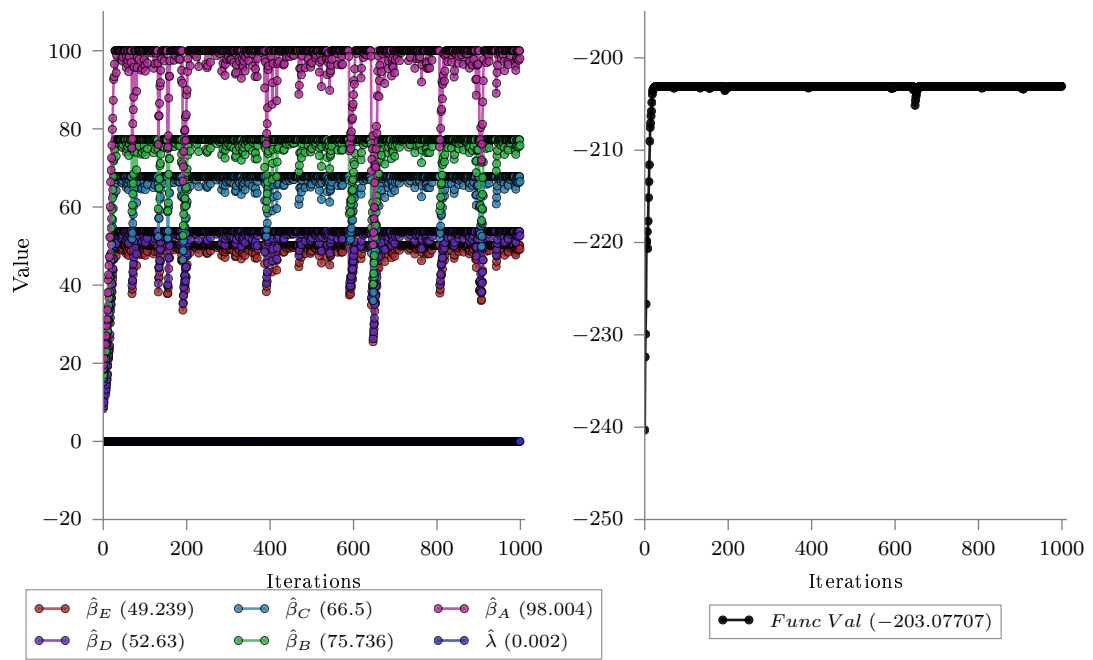EM Iteration Results from Complete Data
(Weekend Data)

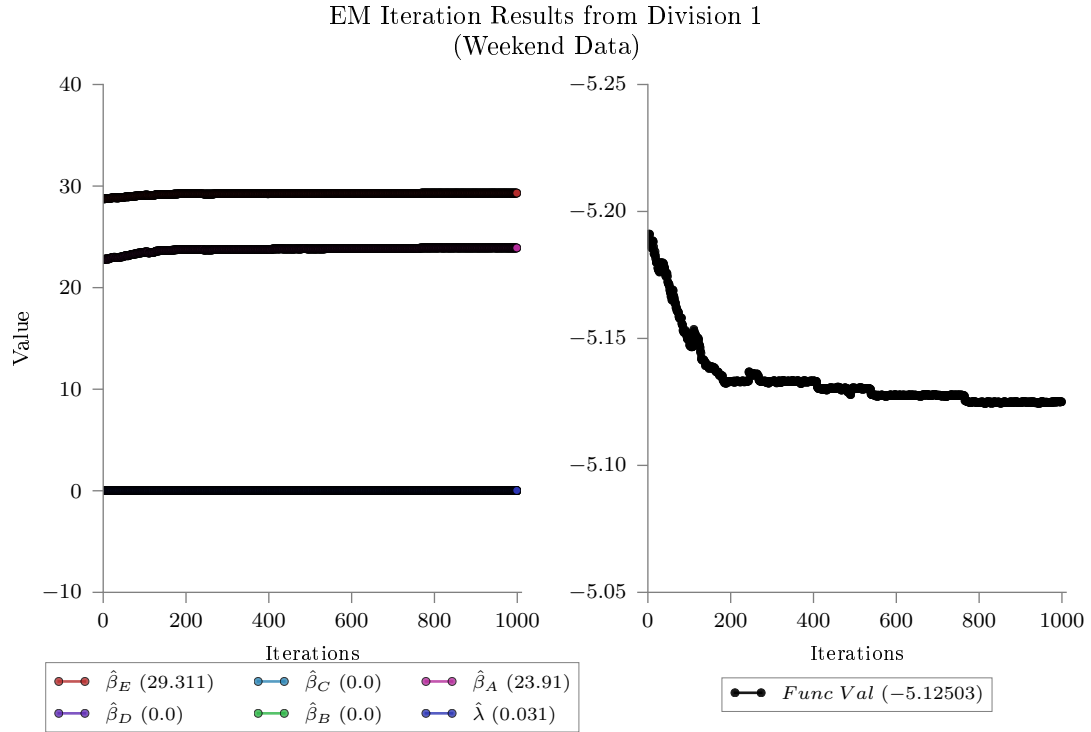Figure 6.7: Weekend Results Complete Data

EM Iteration Results from Division 1
(Weekend Data)

$\hat{\beta}_E$ (29.311) $\hat{\beta}_C$ (0.0) $\hat{\beta}_A$ (23.91)
$\hat{\beta}_D$ (0.0) $\hat{\beta}_B$ (0.0) $\hat{\lambda}$ (0.031)

$Func\ Val$ (-5.12503)

Figure 6.8: Weekend Results Division 1

6.2.2.2 Partitioned Data

Partitioned data results are presented in figures 6.8, 6.9, and 6.10 for divisions 1, 2 and 3 respectively. Note that none of figures 6.8, 6.9 and 6.10 resemble 6.7. The main reason is the diminished identification issue, because of the smaller number of time steps. Division 1 only has 1,080 records compared with the 86,280 records of the complete dataset, while divisions 2 and 3 have 2,748 and 504 respectively.

Despite the larger number of time steps (2,748) for division 2, note that it also encompasses a large period of the booking window: from 10 hours to 19 days and 12 hours prior to departure. This large period, and fewer number of time steps
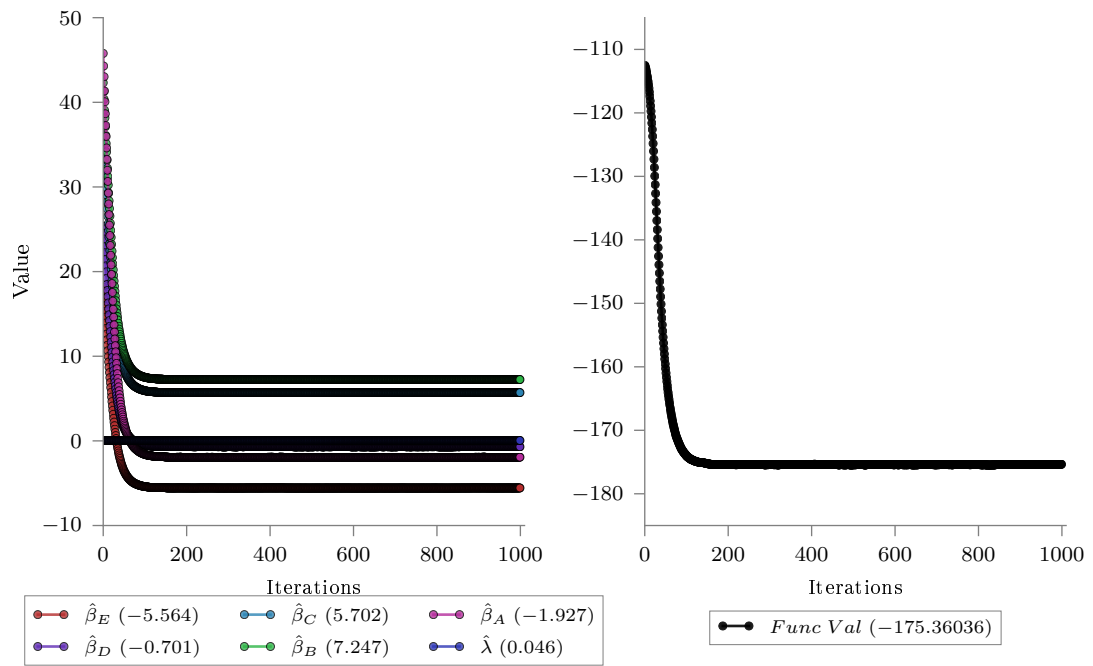
Figure 6.9: Weekend Results Division 2
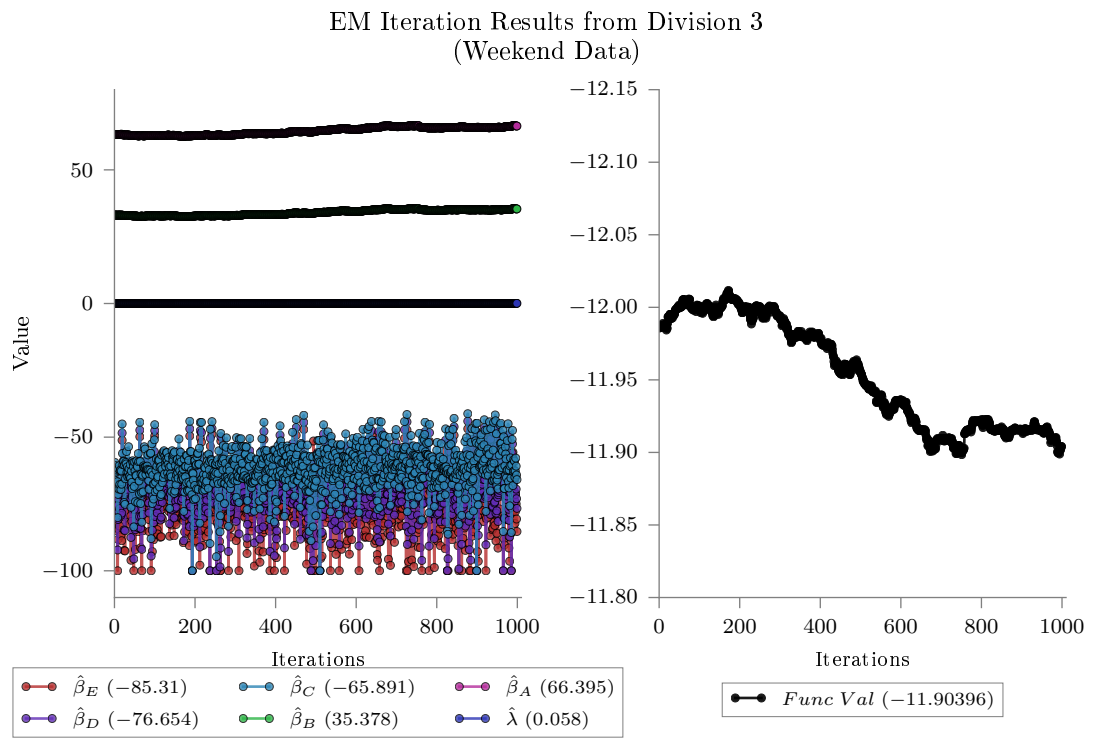
EM Iteration Results from Division 3
(Weekend Data)

Figure 6.10: Weekend Results Division 3

far reduce the problem of identification. Consequently, division 2 is the most stable result of all and converges with a decreasing objective function, as expected. The final values of some coefficients — shown in the legend (in parenthesis) — are also reasonable and negative. However, coefficients of $\hat{\beta}_C$ and $\hat{\beta}_B$ are positive, contrary to intuition.

Chapter 7: Conclusions and Contributions

This chapter presents the contributions made by this thesis. The conclusions drawn from this study are presented in the second section in this chapter.

7.1 Contributions

In this thesis, the framework developed and proposed in Talluri and van Ryzin (2004) was presented [49]. The framework is used to fit a logit model to a dataset when some data is missing or censored. The application to missing data demonstrated in this thesis involves data that is not recorded by the merchant providing a perishable service. In particular, the no-purchase choice made by customers and the customer arrival rate was deduced from synthetic data and real data obtained from a railway operator.

The railway operator, like other public transportation service operators, does not record the instances when customers arrived at their online or physical system to query prices, and decided to not buy the offered choices. These data is missing from the dataset, preventing the application of a choice model as discussed in this thesis. The goal of the EM framework developed in Talluri and van Ryzin (2004) is to handle such instances and impute the data with values to overcome the hurdle

of the missing data [49].

The whole procedure of the data gathering, cleaning and manipulation has been described in this thesis. The method to apply the algorithm to this data has also been discussed in detail. Results from applying the EM algorithm to the test datasets have been demonstrated. The primary takeaway from the results is the identification of the issues presented by the framework.

It was shown that the main issues with applying the method is that the missing data cannot be identified, and that there is a degree of freedom in the formulation that allows for a wide range of values to satisfy the objective function.

7.2   Conclusion

Consequently, the conclusion drawn from this work is that the EM algorithm by itself cannot solve the issues presented by missing data. The formulation is too lenient and the missing data is too vague to be solved by EM. With the help of some additional information, such as an accurate number of time steps, the EM algorithm can be used. However it is not possible to account for an accurate number of time steps. The number of arrivals to the system varies by market and by available alternatives.

A holistic approach to the problem is required to solve for missing data with any amount of accuracy. Firstly, the assumptions of one arrival per time step, and the monotonicity of the time step size should be relaxed. This will allow for a more natural modeling of the real world scenario. Then to solve for the identity

of the missing data, additional information from other sources should be used. For example, a trip planning model can be used to generate the commuters and get an idea about the size of the market.

Relaxing the assumptions about customer arrival will also help solve the degree of freedom in the model. By using a continuous function for time, for example, the model can be designed to fit the distribution of time such that the observed rate of purchases is maximally satisfied. Further, modeling arrivals as per a Poisson or similar process allows the flexibility of changing the arrival characteristics with time.

The ultimate takeaway from this thesis is that the EM algorithm cannot be applied to any other situation except the synthetic data case without a certainty in the results. If the model framework is sufficiently altered to address the issues raised and demonstrated here, there may be valid and dependable results produced.

Chapter 8:   Future Work

In this chapter, certain recent developments are demonstrated. It was illustrated in the results chapter that the EM algorithm will not work unless the number of time steps are assumed such that it is as close to the real number as possible. (See figure 6.2 for reminder.) In the real world scenario, it is almost impossible to accurately get an estimate of the real number of time steps. In fact, even within a day, the rate of bookings will vary widely, from the lull in the early pre-dawn hours to the frenzy in the evenings or mornings, close to office start and end times.

There is no natural way to take this uncertainty into account. The problem statement posed in this thesis has an identifiability problem that EM cannot solve by itself. In fact, EM might not even be a good tool to solve this problem. Over the period of history, since the seminal paper by Kalyan Talluri and Gareth van Ryzin, many attempts have been made to correct for this deficit, and somehow account for identifying the population size (see Section 2.3). Below, a method is presented which can potentially be used to resolve the problem.

8.1 Fitting Demand to Purchase Data

It is not a far fetched assumption to make that demand will be directly proportional to bookings. If the ratio of purchase transactions to total demand is assumed to be a constant, or nearly constant across all time periods, then a better estimate of the choice parameters can be achieved. This is not unlike Talluri (2009) and Newman, et. al. (2012) [46, 37], but instead of a global demand constancy, we argue that demand is equally proportional to booking transaction records even at smaller time intervals.

In keeping true to the assumption of only one arrival per time step, the monotonicity of the time step can be altered such that periods of lean bookings have longer time steps, while periods of heavy bookings have smaller time steps — similar to the partitioned data discussed above, but on much finer timescales. The process that created this data can be visualized as time walking from the start of the booking window to the end of the booking window. The strides taken by time on it's walk form the intervals for the data (also known as time steps).

Assume that time is taking strides of 30 minutes long from the tail end of the booking period (furthest from departure) towards the departure time. As long as a stride encapsulated one or fewer bookings, the stride length was unchanged. If more than one bookings were engulfed in a stride, the stride was broken by the number of bookings engulfed. For example, when taking strides of 30 minutes, if 3 bookings are engulfed, then the stride is broken down into 10 minute long strides, with each 10 minute period containing a booking point. The 10 minute long strides are then

continued towards departure, till more than one booking is covered in a single stride. The minimum stride length is one second long. Bookings falling in the same second (very rarely) were adjusted by moving them to the nearest unoccupied stride.

For weekend this resulted in much better results, with quick convergence and overall parameter stability as shown in figure 8.1. However, for weekday data, the method failed to produce credible results, as illustrated in figure 8.2. One explanation for this is that weekday bookings reduce the stride lengths more drastically than weekend data. That means that very soon, the time steps are quite small, and the frequency of empty strides increases rapidly, proportionately increasing the magnitude of the identification problem. A potential fix is to lengthen the strides when successive strides are empty. This raises the question of the length to elongate the strides back to.

8.2   Issues

As illustrated in figure 8.2, this method is not without problems. The problem is re-lengthening the strides, a question that is more difficult to answer than initially apparent. For example, if the average arrival rate for a given hour is a customer per five minutes, and two bookings occur 10 minutes apart, should the intervening stride be 10 minutes long, or 5 minutes long? Since the arrival rate is unknown before hand, it is not possible to implicitly answer the question.

Moreover, monotonicity of time steps is an assumption in Talluri and van Ryzin (2004), violating which changes the underlying mathematics of the framework [49].

84

EM Iteration Results from Future Method
(Weekend Data)

Legend (left plot): $\hat{\beta}_E$ (-5.467), $\hat{\beta}_C$ (-2.938), $\hat{\beta}_A$ (-9.025), $\hat{\beta}_D$ (-7.224), $\hat{\beta}_B$ (-4.193), $\hat{\lambda}$ (0.065)

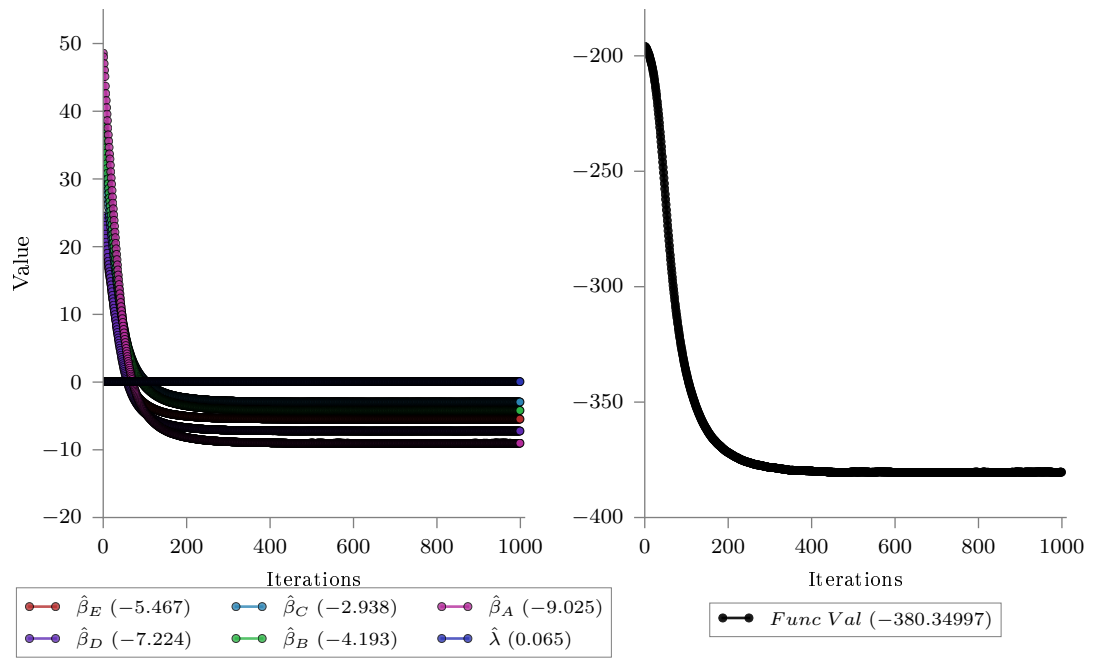Legend (right plot): $Func\,Val$ (-380.34997)

Figure 8.1: Results of Future Research Plans on Weekend Data

85

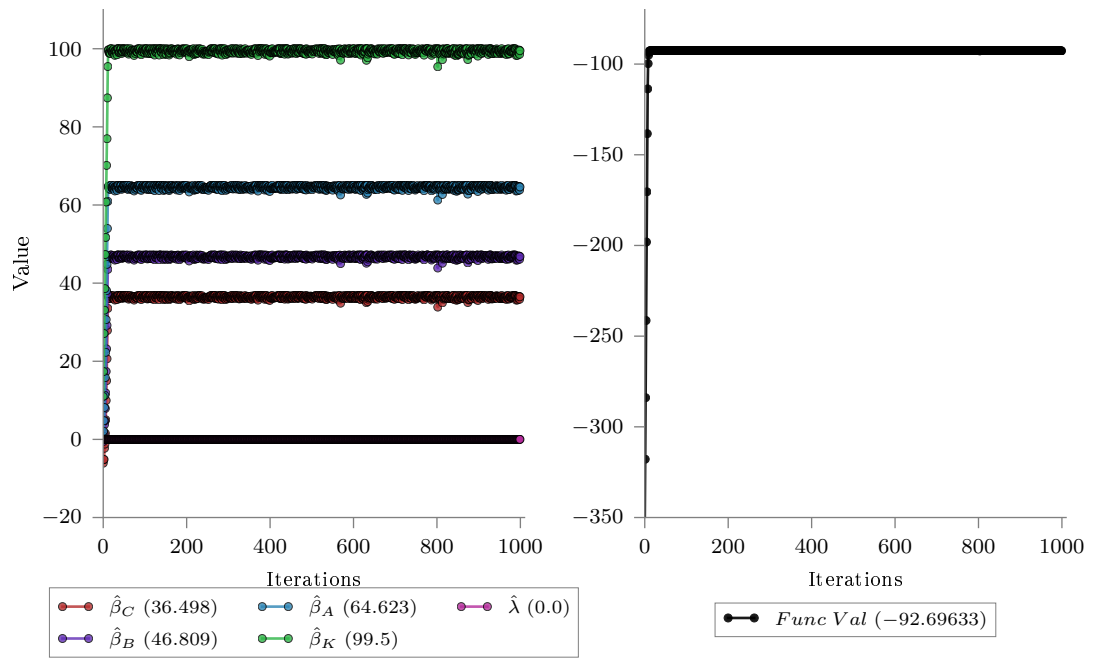EM Iteration Results from Future Method
(Weekday Data)

Figure 8.2: Results of Future Research Plans on Weekday Data

The arrival rate (λ) of the customers, and the arrival probability per time step ($a(t)$) is dependent on the time steps. Changing the time step size as a function of demand changes the arrival rate also to a function of demand through time. Change in the arrival rate affects the Bayesian equation (equation 5.3), which will no longer have the form derived in Talluri and van Ryzin (2004) (and reproduced in equation 5.3). Therefore, this alteration to the model blows up the mathematics, and the results (presented in 8.1 and 8.2)are consequently incorrect.

The solution to these issues lies in changing the model, such that time is properly modeled as a function of demand, and the following mathematics is corrected. A cursory glance at this fix indicates that the framework presented will need a complete makeover. The best way forward seems to use a hierarchical Bayes model that uses Bayes inference to solve for parameters. The required parameters are treated as prior distributions with an assumed shape. These parameters are used in the formulation of a probabilistic model which usually comprises of multiple levels, wherein the prior distributions of the parameters are updated in the light of the given data.

Exploring the hierarchical Bayes models and improving the framework are part of research planned for the future.

Bibliography

[1]   Staffan Algers and Muriel Beser. "Modelling choice of flight and booking class-a study using stated preference and revealed preference data". In: *International Journal of Services Technology and Management* 2.1-2 (2001), pp. 28–45.

[2]   S-E Andersson. "Passenger choice analysis for seat capacity control: A pilot project in Scandinavian Airlines". In: *International Transactions in Operational Research* 5.6 (1998), pp. 471–486.

[3]   Cynthia Barnhart and K Talluri. "Airline operations research". In: *Design and Operations of Civil and Environmental Engineering Systems* (1997), pp. 435–469.

[4]   Richard E Bellman and Stuart E Dreyfus. *Applied dynamic programming.* Princeton University Press, 2015.

[5]   Peter Belobaba. "Air travel demand and airline seat inventory management". PhD. Dissertation. Flight Transportation Laboratory, Massachusetts Institute of Technology, 1987.

[6]   Peter P Belobaba. "OR practiceapplication of a probabilistic decision model to airline seat inventory control". In: *Operations Research* 37.2 (1989), pp. 183–197.

[7]   Peter P Belobaba. "Survey paper-airline yield management an overview of seat inventory control". In: *Transportation science* 21.2 (1987), pp. 63–73.

[8]   Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand.* Vol. 9. MIT press, 1985.

[9]   Dimitris Bertsimas and S. de Boer. "Joint network pricing and resource allocation". In: (2002).

[10]   Shelby L Brumelle and JEFFREY I McGill. "Airline seat allocation with multiple nested fare classes". In: *Operations Research* 41.1 (1993), pp. 127–137.

[11]   Ek Peng Chew, Chulung Lee, and Rujing Liu. "Joint inventory allocation and pricing decisions for perishable products". In: *International Journal of Production Economics* 120.1 (2009), pp. 139–150.

[12]   A Ciancimino et al. "A mathematical programming approach for the solution of the railway yield management problem". In: *Transportation science* 33.2 (1999), pp. 168–181.

[13]   Cinzia Cirillo and Kay W Axhausen. "Dynamic model of activity-type choice and scheduling". In: *Transportation* 37.1 (2010), pp. 15–38.

[14] Cinzia Cirillo, Michael Maness, and Nayel Urena Serulle. "Measuring value of travel time and travel time variability in the presence of managed lanes: results from a pilot stated preference survey on the Capital Beltway". In: *Transportation Letters* 6.1 (2014), pp. 23–35.

[15] Claire Claire Jia Ling Cizaire. "Optimization models for joint airline pricing and seat inventory control: Multiple products, multiple periods". PhD. Dissertation. Massachusetts Institute of Technology, 2011.

[16] W Cooper. "Asymptotic behavior of some revenue management policies". In: *Univ. of Minnesota Working Paper, Minneapolis, Minn* (2000).

[17] Jean-Philippe Côté, Patrice Marcotte, and Gilles Savard. "A bilevel modelling approach to pricing and fare optimisation in the airline industry". In: *Journal of Revenue and Pricing Management* 2.1 (2003), pp. 23–36.

[18] Robert G Cross. *Revenue management*. Crown Business, 2011.

[19] Renwick E Curry. "Optimal airline seat allocation with fare classes nested by origins and destinations". In: *transportation science* 24.3 (1990), pp. 193–204.

[20] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[21] Moshe Dror, Pierre Trudeau, and Shaul P Ladany. "Network models for seat allocation on flights". In: *Transportation Research Part B: Methodological* 22.4 (1988), pp. 239–250.

[22] Fred Glover et al. "The passenger-mix problem in the scheduled airlines". In: *Interfaces* 12.3 (1982), pp. 73–80.

[23] Pratt Hetrakul. "Discrete Choice Models for Revenue Management". PhD. Dissertation. University of Maryland, 2012.

[24] Ian SA Hood. "MERLIN: Model to evaluate revenue and loadings for intercity". In: *Yield management: Strategies for the service industries, 2nd edn. Thomson, London* (2000), pp. 98–107.

[25] Harold Jeffreys. *Scientific inference*. Cambridge University Press, 1973, p. 31. ISBN: 978-0-521-18078-8.

[26] A Kuyumcu and A Garcia-Diaz. "A polyhedral graph theory approach to revenue management in the airline industry". In: *Computers & industrial engineering* 38.3 (2000), pp. 375–395.

[27] Ting Li et al. "Passenger travel behavior model in railway network simulation". In: *Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference. 2006, pp. 1380–1387.

[28] K Littlewood. "Forecasting and Control of Passengers". In: *12th AGIFORS Symposium Proceedings*. Vol. 95. 1972, p. 128.

[29] Yangwen Liu and Cinzia Cirillo. "Measuring transit service impacts on vehicle ownership and use". In: *Public Transport* 7.2 (2015), pp. 203–222.

[30] Yangwen Liu, Jean-Michel Tremblay, and Cinzia Cirillo. "An integrated model for discrete and continuous decisions with application to vehicle ownership, type and usage choices". In: *Transportation Research Part A: Policy and Practice* 69 (2014), pp. 315–328.

[31] R. Ducan Luce. *Individual Choice Behavior a Theoretical Analysis*. John Wiley and sons, 1959.

[32] Charles F Manski and Steven R Lerman. "The estimation of choice probabilities from choice based samples". In: *Econometrica: Journal of the Econometric Society* (1977), pp. 1977–1988.

[33] Charles F Manski and Daniel McFadden. "Alternative estimators and sample designs for discrete choice analysis". In: *Structural analysis of discrete data with econometric applications* (1981), pp. 2–50.

[34] Daniel McFadden. "Conditional logit analysis of qualitative choice behavior". In: (1973).

[35] Jeffrey I McGill and Garrett J Van Ryzin. "Revenue management: Research overview and prospects". In: *Transportation science* 33.2 (1999), pp. 233–256.

[36] Jeffrey P Newman, Mark E Ferguson, and Laurie A Garrow. "Estimating GEV models with censored data". In: *Transportation Research Part B: Methodological* 58 (2013), pp. 170–184.

[37] Jeffrey Newman, Mark Ferguson, and Laurie Garrow. "Estimating discrete choice models with incomplete data". In: *Transportation Research Record: Journal of the Transportation Research Board* 2302 (2012), pp. 130–137.

[38] Jeffrey Newman, Mark Ferguson, and Laurie Garrow. "Estimating nested logit models with censored data". In: *Transportation Research Record: Journal of the Transportation Research Board* 2343 (2013a), pp. 62–67.

[39] Saratchai Ongprasert. "Passenger Behavior on Revenue Management Systems of Inter-city Transportation". PhD. Dissertation. Citeseer, 2006.

[40] R Phillips. "State-contingent airline yield management". In: *Presentation in Session TC33* 4 (1994).

[41] Lawrence W Robinson. "Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes". In: *Operations Research* 43.2 (1995), pp. 252–263.

[42] Garrett van Ryzin and Gustavo Vulcano. "An expectation-maximization method to estimate a rank-based choice model of demand". In: *Operations Research (Forthcoming)* (2016).

[43] Soheil Sibdari, Kyle Y Lin, and Sriram Chellappan. "Multiproduct revenue management: An empirical study of auto train at Amtrak". In: *Journal of Revenue & Pricing Management* 7.2 (2008), pp. 172–184.

[44] Robert Warren Simpson. *Using network flow techniques to find shadow prices for market demands and seat inventory control*. MIT, Department of Aeronautics and Astronautics, Flight Transportation Laboratory, 1989.

[45] Arthur Sullivan. "Economics: Principles in action". In: (2003).

[46] Kalyan Talluri. "A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters". In: *Available at SSRN 1374853* (2009).

[47] Kalyan Talluri and Garrett Van Ryzin. "A randomized linear programming method for computing network bid prices". In: *Transportation Science* 33.2 (1999), pp. 207–216.

[48] Kalyan Talluri and Garrett Van Ryzin. "An analysis of bid-price controls for network revenue management". In: *Management Science* 44.11-part-1 (1998), pp. 1577–1593.

[49] Kalyan Talluri and Garrett Van Ryzin. "Revenue management under a general discrete choice model of consumer behavior". In: *Management Science* 50.1 (2004), pp. 15–33.

[50] KT Talluri. "Revenue management on a line network". In: *Work-ing paper, Universitat Pompeu Fabra, Barcelona, Spain* (1996).

[51] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[52] Gustavo Vulcano, Garrett Van Ryzin, and Richard Ratliff. "Estimating primary demand for substitutable products from sales transaction data". In: *Operations Research* 60.2 (2012), pp. 313–334.

[53] Lawrence R Weatherford. "Using prices more realistically as decision variables in perishable-asset revenue management problems". In: *Journal of Combinatorial Optimization* 1.3 (1997), pp. 277–304.

[54] Gerard Whelan and Daniel Johnson. "Modelling the impact of alternative fare structures on train overcrowding". In: *International journal of transport management* 2.1 (2004), pp. 51–58.

[55] Gerard Whelan et al. "Optimal fares regulation for Britains railways". In: *Transportation Research Part E: Logistics and Transportation Review* 44.5 (2008), pp. 807–819.

[56] Elizabeth Louise Williamson. "Airline network seat inventory control: Methodologies and revenue impacts". PhD. Dissertation. Flight Transportation Laboratory, Massachusetts Institute of Technology, 1992.

[57] Elizabeth Louise Williamson. "Comparison of optimization techniques for origin-destination seat inventory control". Master's Thesis. Flight Transportation Laboratory, Massachusetts Institute of Technology, 1988.

[58] Richard D Wollmer. "An airline seat management model for a single leg route when lower fare classes book first". In: *Operations Research* 40.1 (1992), pp. 26–37.

[59] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.