Nonparametric Estimation of Demand Structures in Airline Revenue Management

Johannes Ferdinand Jörg, Catherine Cleophas

Lehr- und Forschungsbereich Advanced Analytics (ADA)
RWTH Aachen University, Germany
Email: johannes.ferdinand.joerg@ada.rwth-aachen.de
Email: catherine.cleophas@ada.rwth-aachen.de

***Abstract*—Airline revenue management employs forecasting and optimization techniques to offer the right price at the right time to the right customer. With the ability to store large amounts of data comes the challenge to incorporate the information contained in those data sets. This contribution considers the estimation of demand segments present in a specific market using nonparametrical methods on panel data. We employ finite mixtures to model booking events in different time frames and to obtain an estimator for the number of demand segments. Via an airline revenue management simulation tool, we perform an experimental study of the derived estimator. We discuss the results with respect to the underlying demand structure of the simulation and identified demand segments. The findings are discussed with respect to theoretical and practical use. Finally, we discuss the real world applicability and possible further research intentions.**

Keywords–nonparametric; demand segments; estimation; revenue management

I. INTRODUCTION

A central theme of revenue management is analyzing historical sales data to draw conclusions on the underlying demand structure. A multitude of sources influence booking data in airline revenue management. In addition, availability control and product restrictions censor sales. This leads to a discrepancy of historical sales data and actual demand. Studying this issue helps to understand the market and to improve revenue management optimization parameters.

To apply revenue management techniques, we have to be able to segment the market. The goal is to optimize the availability of products over time such that the customers' willingness to pay is exploited. Talluri and van Ryzin [1] give a detailed overview of revenue management techniques. Naturally, this motivates identifying demand segments and their behavior to forecast the number of customers arriving in a certain time frame. Airline revenue management usually segments demand by characteristica including customers' willingness to pay, utility costs for different product properties (e.g., weekend stay, minimum stay, economy or business class, number of transfers) and date of the request or cancellation. One example would be to simply segment the market into business travellers with a high willingness to pay and late booking requests and private travellers with lower willingness to pay and earlier booking requests. In practice, this relatively simple distinction does not suffice to obtain a satisfying forecast or optimization. Also, most of the related works assume a fixed number of demand segments as input into estimation or optimization procedures, e.g., [2], [3]. Thus, we are motivated to find ways to obtain a suitable amount of demand segments.

Revenue management forecasting usually relies on historical data to extrapolate demand. It is important to accurately forecast demand, as for example a study by Pölt [4] suggests that a 20% increase in forecasting accuracy may lead to an 1% increase in revenue. As the amount of data stored by companies steadily increases, methods to analyze these data sets are needed. Most current, practice-oriented approaches explain demand structures with parametric statistics or heuristics. For an overview of forecasting methods see, e.g., the taxonomy of Azadeh [5]. Parametric estimation needs specified underlying distributions. We focus on nonparametric estimation of demand structures here, which uses large data sets to remove the assumption of a specific underlying distribution.

One example of the application of nonparametric statistics to revenue management is presented by Van Ryzin and Vulcano [6]. They propose an expectation-maximization approach for analysing market structures. They identify demand segments with their preferences over the set of alternatives and use an iterative algorithm to create new sets of demand segments, such that the probability for the observed booking distribution is maximized.

To model customer decisions, we employ a discrete choice model: Customers buy one of a set of products at each specified time period. In the case of quantity-based airline revenue management, we identify products with discrete booking classes on a specific flight itinerary and time periods correspond to flight departures. Our proposed method to estimate the number of demand segments works as follows: In a first step, we formulate our model for panel data of two time periods using a finite mixture model to represent the probability of a booking event. An introduction to finite mixture models can be found in McLachlan and Peel [7]. In the next step, we decompose the model such that the estimation of the number of demand segments becomes a rank estimation problem. This idea is based on an approach of Kasahara and Shimotsu [8], in which the authors derive sufficient conditions for the nonparametric identifiability for various finite mixture models in the framework of discrete choices.

This short paper is structured as follows: In Section 2, we define a mathematical model of the booking process and derive a lower bound for the number of demand segments in a market. Section 3 discusses a study, which employs an airline revenue management simulation tool in order to test the methods in a theoretical environment. Finally, Section 4 reviews our findings, provides an outlook of the intended extensions to the model, and addresses practical applicability.

 ISBN: 978-1-61208-416-9

II. METHODS

The notion of finite mixture models is often applied in economics, chemistry and health care [9]. In principle, this modeling technique can represent an observation set as a composition of several subpopulations. In the context of airline revenue management, this can explain seemingly homogeneous booking data by presuming the presence of a number of demand segments.

As input for our estimation method, we use panel data. Panel data is the generic term for data which is collected at different times for the same population and the same indicators. Here, we will observe customers for a set amount of flight departures and store their booking decisions.

Suppose we have panel data of individuals over a number of T time periods. In each time period $t \in T$, an individual chooses to buy one alternative x_i out of the set of available alternatives X. Let M be the number of mixture components contributing to the observed data. The probability of buying a product is denoted by $p^*_m[x_1]$ for the first time period and by $p_{m,t}[x_t]$ for $t > 1$. Each mixture component m contributes a factor $\omega_m \in [0, 1]$ to the observation. The baseline model of our investigation then looks as follows:

$$P[\{x_t\}_{t=1}^T] = \sum_{m=1}^{M} \omega_m p^*_m[x_1] \prod_{t=2}^{T} p_{m,t}[x_t], \tag{1}$$

where $\sum_{m=1}^{M} \omega_m = 1$ ensures that the probability (1) consists of the contribution of demand segments. This means that the probability for an event $\{x_t\}_{t=1}^T$ is completely explained by the M mixture components.

Now, we can identify the nature of demand segments: Each mixture component m represents a demand segment with its own probability to book a specific class x_t at different times t. Thus, our objective to estimate the number of demand segments corresponds to estimating parameter M of our model. We assume that panel data of customers for $T = 2$ time periods is available.

Since the temporal spacing of time periods is not yet defined and we consider an airline revenue management context, let the time periods correspond to flight departures. On all flights, the same set of booking classes X is offered such that a booking event for two time periods is a tuple (x_1, x_2). Therefore, the model (1) simplifies to

$$P[(x_1, x_2)] = \sum_{m=1}^{M} \omega_m p^*_m[x_1] p_{m,2}[x_2]. \tag{2}$$

Let N be the number of individual observations in the panel data and denote by $x^i = (x_1, x_2)$ the observation tuple of the i^{th} individual. Given this data, we can estimate the probabilities $P[(x_1, x_2)]$. In a first approach, we use plug-in estimates for this probability, i.e.,

$$p_{i,j} = \hat{P}[(x_1 = i, x_2 = j)] = \frac{\sum_{i=1}^{N} 1_{\{(x_1=i, x_2=j)\}}(x^i)}{N}. \tag{3}$$

With these estimates, we can calculate an observable $|X| \times |X|$ matrix $P = (p_{i,j})$

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,|X|} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,|X|} \\ \vdots & \vdots & \ddots & \vdots \\ p_{|X|,1} & p_{|X|,2} & \cdots & p_{|X|,|X|} \end{pmatrix} \tag{4}$$

Note here, that in order to identify M, it is required to have $|X| \geq M$.

Defining $V = \text{diag}(\omega_1, \ldots, \omega_M)$, this admits the decomposition $P = P_1 V P_2$, where P_1 and P_2 are $|X| \times M$ and $M \times |X|$ matrices consisting of the entries p^*_m and $p_{m,2}$, respectively. We can now apply a simple rank argument: Since $\text{rank}(V) = M$ and $\text{rank}(P_i) \leq M$, we have that

$$\text{rank}(P) \leq \min\left\{(\text{rank}(P_1), \text{rank}(V), \text{rank}(P_2)\right\} \leq M \tag{5}$$

Therefore, the rank of matrix P yields a lower bound for the number of demand segments M, i.e.,

$$M \geq \text{rank}(P). \tag{6}$$

III. RESULTS

Our results are based on an airline revenue management simulation. It models the complete revenue management process from forecasting demand to optimizing available booking classes, to taking reservations of arriving customer requests. It also includes an extensive stochastic demand model to create artificial demand. Through simulation experiments, we obtain data sets, which provide input for the estimation process. Customers are generated according to several parameters (e.g., willingness to pay, preferred flight departure time, utility costs for specific product properties). The algorithm draws these values with a given error term from a normal distribution, to create realistic variation. Demand segments in the context of our simulation tool are then a specific set of these underlying parameters. The set of available alternatives X is defined by a set of 12 booking classes and the no-purchase alternative. We distinguish booking classes by price, compartment and additional product properties, such as being refundable or rebookable or requiring a minimum stay. These product properties impose, depending on the demand segment, either a penalty on the utility of the customer or an acceptance probability, such that the customer has to decide if he accepts this restriction or not without having further implications on the utility. As expected, the more limitations are imposed on a booking class, the cheaper its price. It is also possible for booking classes to only differ in price. Three of the 12 booking classes are located in the business compartment, the remaining 9 are located in the economy compartment.

Customers are identified by a unique customer identification number. Thus, we can track their purchases over several flight departures. We also assume that the simulation parameters do not change over the course of both time periods, i.e., the willingness to pay, product restriction costs, and request and cancellation date are constant for both flights for which they requested tickets. In order to obtain different bookings given constant demand, we varied the forecasts for both flights such that the revenue management system does not offer the same product availabilities in both cases. This was achieved by introducing a forecast error as follows: For each booking class on one flight itinerary, we draw a realization of a random variable with uniform distribution on $[0.8, 1.2]$. The forecast

 ISBN: 978-1-61208-416-9

for this booking class is then multiplied by the value of the realization. This ensures that customers do not always book the same class and that the panel data obtained is sufficiently diverse.

To compute the rank of matrix P, we used a QR decomposition with varying tolerance for eigenvalues which are considered to be zero. This will be denoted by QRx, where x is the number of decimal places which are considered non-zero and the values of QRx are the lower bound for the number of demand segments computed. We also modified the number of customers we tracked over several simulation runs, to assess the amount of data needed for convergence.

We performed the estimation procedure with three different scenarios. The first data set is based on a simulation with only two demand segments. These customers who book mainly prefer the economy compartment, one demand segment that has higher willingness to pay and the other has a lower willingness to pay. The second data set describes a scenario with six demand segments. Each of the demand segments has its own set of parameters and therefore its own subset of the available booking classes, which customers consider buying. While this may be the most realistic scenario considered here, we plan to perform the estimation for additional scenarios. The third data set is similar to the first, but one demand segment is mainly business oriented while the other one is mainly economy oriented, such that there is an inherent separation within the data set.

TABLE I. SIMULATION WITH TWO DEMAND SEGMENTS

Number of customers	*QR7*	*QR4*	*QR3*	*QR2*	*QR1*
100	7	7	7	7	7
1000	7	7	7	7	7
10000	7	7	7	7	5
50000	8	8	7	6	5
100000	8	8	7	6	4

At first glance, the results exhibit some peculiarities. Table I shows that the lower bound of demand segments actually increases when the number of observations increases, i.e., from seven to eight demand segments when observing 10000 and 50000 customers. We would expect the number of demand segments to decrease as demand observations increase. The reason that this is not always the case lies in the simulation design. Since customers are stochastically generated, outliers buy products that were not bought in simulation runs before. The algorithm recognises these customers as a new demand segment.

TABLE II. SIMULATION WITH SIX DEMAND SEGMENTS

Number of customers	*QR7*	*QR4*	*QR3*	*QR2*	*QR1*
100	12	12	12	12	12
1000	13	13	13	13	13
10000	13	13	13	13	12
25000	10	10	10	10	10
50000	8	8	8	8	8
100000	7	7	7	7	7

Table II shows the results for the simulation with six demand segments. A similar effect to the one in Table I takes place for 100 and 1000 observations: Again, the lower bound for the number of demand segments increases due to the increase in generated customers and thereby customers who booked booking classes that were not observed before. Here, we also note that the sensitivity to the non-zero eigenvalues is lowered compared to Table I. The only occurence of a difference is at 10000 customers observed from 13 to 12 demand segments.

TABLE III. SIMULATION WITH TWO DEMAND SEGMENTS WITH DISTINCT BOOKING BEHAVIOUR

Number of customers	*QR7*	*QR4*	*QR3*	*QR2*	*QR1*
100	10	10	10	10	10
1000	10	10	10	10	8
10000	8	8	8	8	6
50000	7	7	7	7	4
100000	5	5	5	5	4

Table III shows the results of the estimation for the third scenario setup. It includes two demand segments, which largely differ in their booking behaviour. This means that customers from one demand segment usually book cheaper economy classes while the others book the more expensive economy and business classes. A low amount of observations leads to a high amount of estimated demand segments, since there is a broad spectrum of booking classes booked. As the number of observations increases, we obtain an even lower amount of demand segments than in the first scenario (compare Table I). Since it should be easier to separate two demand segments, which are clearly distinct, than two which overlap in booking behaviour, this meets our expectations.

These preliminary results indicate that the approach does better for a higher amount of demand segments. The identification of demand segments in this model still remains open, as the presented methods do not explain the relationship between the simulation-generated demand segments and those identified by the estimation procedure presented here.

IV. CONCLUSION

We have developed a procedure to estimate a lower bound of the number of demand segments from panel data. This approach does not rely on the specification of an underlying demand structure and thus can be used without the knowledge of specific distribution functions. First preliminary results, which are based on a simulation study of an airline revenue management tool, show promising results.

There are several gaps to be filled in order to use the presented methods in real world applications. Some simplifications limit the generality of the model, e.g., the assumption that the probabilities are not dependent on previous time periods. This is clearly not the case in real world scenarios. In the introductory example, we can easily deduce that for business travellers, the probability to book a business class again should be higher than buying a lower economy class. We expect that the probabilities are not too eratic. Still, assuming that we obtained a number of demand segments in a market, we may now use that knowledge to derive the probability distribution of each segment. This requires panel data for at least three time frames as is described in [8]. With these probability distributions and the information of the booking data, we may calculate a forecast for future flights.

Another point of interest is the distribution of requests for the demand segments. As the estimation procedure uses final booking data, i.e., booked classes after departure of the flight,

 ISBN: 978-1-61208-416-9

the information about when the customer booked his flight is lost. However, the request date is a crucial part of dynamic revenue management optimization techniques. We may argue that if we can obtain panel data in the first place, we should be able to know when the booking was made. This would present us with the possibility to make inference of when the demand segments likely request over the booking horizon.

It is also difficult to assess the quality of the preliminary results, as we do not know yet which initially created segments are identified. The ultimate test should be to compare the forecast, which resulted from this estimation procedure, with forecasts of state-of-the-art methods currently used in practice. One may also substitute different parts of already established forecasting methods with nonparametric estimation procedures in order to alleviate the need for a specific underlying distribution which has to be specified beforehand. For example, we may be able to represent price elasticities with the probabilities of the demand segments.

In current practice, creating observation sets in the form of panel data is difficult, since we usually only have access to the information of how many bookings were observed in each offered class after the flight departed. Neither the time of the request nor the set of available classes at that time is stored. The need to track each single customer for a set amount of time frames is imperative for the presented estimation method. In airline revenue management, we may consider the use of bonus cards or improved tracking of online purchases of flights to create such data sets. Other areas, where revenue management is commonly practised, e.g., hotels and car rentals, may also use the fact that customers usually have to leave personal information when renting rooms or cars. The access to this kind of information should be available and give more detailed data sets.

The next steps to extend this model include to incorporate more observable characteristica into the data set in order to lower the amount of observations needed to obtain reasonable results. We may additionally consider the availability of booking classes at the request time as an explaining characteristic. Also, other panel data sets have to be constructed from other simulations to assess the behaviour of the lower bound. We plan to create scenarios with up to 12 demand segments present. This would amount to a scenario where each demand segment has exactly one booking class which he is willing to buy. In revenue management, these kinds of customers are called product customers as opposed to priceable customers, e.g. see [10], since their choice set only consists of the specified booking class and the no-purchase alternative. Another approach would be to allow for incomplete observations of customers, i.e., assess the viability of this method for customer data sets which do not track every single customer for two or more time periods. Incomplete or inaccurate observations may also be created when the technique of collecting this panel data is not able to exactly assign recurring customers. We plan to use different rank estimation methods for our matrix rather than an explicit calculation, e.g., the rank stastistic of Robin and Smith [11] or Kleibergen and Paap [12].

REFERENCES

[1] K. T. Talluri and G. J. Van Ryzin, The theory and practice of revenue management. Springer Science & Business Media, 2004, vol. 68, ISBN: 978-1-4020-7701-2.

[2] P. Hall and X.-H. Zhou, "Nonparametric estimation of component distributions in a multivariate mixture," Annals of Statistics, 2003, pp. 201–224, ISSN: 0090-5364.

[3] E. S. Allman, C. Matias, and J. A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," The Annals of Statistics, 2009, pp. 3099–3132, ISSN: 0090-5364.

[4] S. Pölt, "Forecasting is difficult–especially if it refers to the future," in AGIFORS-Reservations and Yield Management Study Group Meeting Proceedings, 1998, pp. 61–91.

[5] S. Sharif Azadeh, "Demand forecasting in revenue management systems," Ph.D. dissertation, École Polytechnique de Montréal, 2013.

[6] G. van Ryzin and G. Vulcano, "A market discovery algorithm to estimate a general class of nonparametric choice models," Management Science, vol. 61, no. 2, 2015, pp. 281–300, ISSN: 1526-5501.

[7] G. McLachlan and D. Peel, Finite mixture models. John Wiley & Sons, 2000, ISBN: 978-0-471-00626-8.

[8] H. Kasahara and K. Shimotsu, "Nonparametric identification of finite mixture models of dynamic discrete choices," Econometrica, vol. 77, no. 1, 2009, pp. 135–175, ISSN: 1468-0262.

[9] P. Schlattmann, Medical applications of finite mixture models. Springer Science & Business Media, 2009, ISBN: 978-3-540-68650-7.

[10] T. Fiig, K. Isler, C. Hopperstad, and P. Belobaba, "Optimization of mixed fare structures: Theory and applications," Journal of Revenue & Pricing Management, vol. 9, no. 1, 2010, pp. 152–170, ISSN: 1476-6930.

[11] J.-M. Robin and R. J. Smith, "Tests of rank," Econometric Theory, vol. 16, no. 02, 2000, pp. 151–175, ISSN: 0266-4666.

[12] F. Kleibergen and R. Paap, "Generalized reduced rank tests using the singular value decomposition," Journal of econometrics, vol. 133, no. 1, 2006, pp. 97–126, ISSN: 0304-4076.