# Summary of Bayesian Data Combination Approach Integrating Micro and Macro Information: Application to Duration Models for Incomplete Marketing Data

by

**Ryosuke Igari**

Graduate School of Economics
KEIO UNIVERSITY

# Chapter 1. Introduction

## Recent Changes in the Marketing Environment

The marketing environment has changed drastically in recent years. Customers purchase products or services in various places: supermarkets, corner stores, department stores, drug stores, catalog or telephone shopping channels, online electric commerce (EC) sites, and free-market applications on PCs or smartphones. As the marketing environment has changed, activities such as marketing communications, distribution channels, or consumers' preferences have also changed. Researchers have begun to study the issues associated with the new marketing environment. For example, the use of the internet is considered one of the most important changes in marketing practices and research. Consumers now purchase products on EC sites such as Amazon or Rakuten. Additionally, companies, that provide products or services, use the internet as a distribution channel in addition to traditional channels such as stores or catalogs.

However, in these complex marketing environments, the complete data sets for analyzing consumers' behaviors are difficult to obtain. For example, to estimate the effects of internet advertising or SNS pages on consumers' purchasing behavior directly, single-source data, which captures purchase history from all stores and EC sites and browsing behavior from PCs, smartphones, and tablet devices for the same consumer, are required. However, complete data sets of consumer behavior are difficult to obtain because of privacy issues or data collection costs. In statistical analysis (including marketing models) or econometric models, complete data, which is obtained from single-source data, is required to capture the relationships between variables such as correlations, covariance or regression coefficients. However, as we mentioned, consumers purchase products from various channels including own stores or competing stores, EC sites, and catalogs or telephone shopping, which means that it is virtually impossible for researchers to capture their behavior completely, that is, it is difficult to obtain single-source data. If researchers could obtain the complete data in the form of panel or longitudinal data, these data would be biased, because obtaining these complete data forces target samples to respond to extensive questions, which leads to biased responses. If we analyze the biased data, the estimated results would be biased, which leads to incorrect interpretations of the effects of marketing variables such as price (one of the most important variables in marketing and economics).

In this setup, some studies use macro information to strengthen the accuracy of micro-level data modeling. Imbens and Lancaster (1994) proposed a method of incorporating the macro-level information into micro-level models using generalized method of moments (GMM). Similarly, Chaudhuri *et al.* (2008) proposed empirical likelihood approaches that include auxiliary information in individual-level modeling. We consider these methods a type of data combination used in economics (Ridder and Moffit, 2007). Combining micro and macro data is considered one type of data combination; we call this data combination in this study, and propose a Bayesian approach for data combination.

## Aims of This Paper

In this paper, we discuss a microeconometric model to capture consumers purchase behavior. Especially, we focus on a duration analysis (e.g., Klein and Moeschberger, 2003; Ibrahim *et al.*, 2005) and its application to an interpurchase timing model (e.g., Jain and Vilcassim, 1991; Helsen and Schmittlein, 1993) using incomplete marketing data. Duration analysis captures the time when the next event will occur, and particularly in marketing, it is used to capture consumer's purchase timing. The role of interpurchase timing models is to estimate the effect of marketing promotions such as price coupons and predict the time when consumers are highly likely to purchase products.

However, duration analysis has serious problems with respect to competing events, dropouts, missing responses, missing covariates, missing censoring indicators, and intermittent missingness. These problems are generalized as a missing data analysis (e.g., Little and Rubin, 2002). In a missing data analysis, the missing indicator, which shows whether each observation is missing or not, is fully obtained. Then, specification of the missing mechanisms is required. However, it is sometimes difficult to obtain the missing indicators and specify the missing mechanisms for missing observations such as purchase behavior in competing stores or EC sites. That is, as we mentioned, in a complex marketing environment, the complete data sets for analysis are difficult to obtain, and these incomplete or missing data problems are generated from purchase behavior in competing stores or on EC sites. Under these circumstances, the missing data analysis is not simply applicable, because fully observed missing indicators and the correct specification of missing mechanisms are difficult. Thus, we focus on the situation where the censoring indicator cannot be obtained, the missing mechanism cannot be specified, and the missing indicator cannot be observed even when we consider the interpurchase timing model using duration analysis.

Then, for these incomplete marketing data problems, we propose solutions that integrate macro-level information into micro-level data modeling using a Bayesian data combination approach. Moreover, we apply our Bayesian data combination approach to duration analysis for incomplete data without specification of missing mechanisms and without fully observed missing indicators by integrating micro and macro data. By integrating the macro information, we can estimate the parameters from micro-level data composing the restriction. Additionally, we propose the Bayesian estimation method using the Markov chain Monte Carlo (MCMC) method (e.g., Gelman *et al.*, 2013), which allows us to estimate the parameters flexibly without requiring numerical optimization.

# Chapter 2. Literature Review

In this chapter, we focus on a literature survey of subjects that are relevant to this paper. First, we introduce duration analysis and its application to a marketing interpurchase timing model. Second, we introduce the incomplete data problems in duration analysis. When we calculate the likelihood function of duration models, we must consider the censoring indicator. If we ignore the censoring indicator, the estimator will have bias.

However, censoring indicators are sometimes missing. Therefore, we introduce the missing censoring indicator problems. Additionally, in duration analysis, we must consider the competing events in the likelihood function if there are any competing events in the duration analysis. Competing events are considered the same as censoring problems. Additionally, we consider the intermmittent missingness in repeated duration analysis and treat the problem as an unobserved missing indicator. Third, we introduce data combination approaches that combine micro and macro data using GMM and empirical likelihood approaches.

# Chapter 3. Duration Models Considering Competing Events: Application to Interpurchase Timing Model in Multi-channel

In recent years, multi-channel strategy has become more significant in the marketing field as the EC market grows. Many companies, that manufacture products, have real stores and online channels such as EC sites for the distribution of own products. In this environment, marketing managers should consider an online to offline (OtoO) strategy and the interrelationships of consumer's purchase behavior in multi-channel distribution. Some consumers purchase the products in the real store or EC sites, but the same consumers' purchase histories in multi-channel cannot be usually obtained jointly, because these data such as scan-panel data in-store and click-stream data in EC sites are given from different data source. In this way, analysis of consumers' purchase behaviors in multi-channel distribution is difficult because of data-accessibility, therefore companies usually analyze these data separately. However, the private brand (PB) companies, which owns both the manufacturing and the retail stores, are likely to record both online and real store purchase histories for the same consumers by member registration through smartphone apps or frequent shopper programs. Therefore, we can determine what type of customer will purchase the products in a real store or EC site and who will purchase the products from both channels.

In this chapter, we focus on consumer's purchase behavior in online and offline channel using single-source data that captures the purchase behavior in each channel simultaneously. Especially, we deal with interpurchase timing and purchase amount in online and offline channel. When consumers purchase products, they choose one channel, online or real store. In this context, online and real stores are competing. If there are competing channels, we must consider modeling the interpurchase timing for each single event and the competing interpurchase timing. For example, when a consumer purchases a product on an EC site, the purchase record of the competing event, such as a store purchase, will not be observed. This situation is called competing event, and single-event analysis will lead to mistaken results. In this situation, competing risk models, considering competing events, have been proposed (e.g., Kalbfleisch and Prentice, 2002). Then, we propose a joint model for consumer's interpurchase timing and amount in online and offline channels using single-source data that records the purchase behavior of a PB company in both channels. The proposed model consists of three components: (1) a competing risk model that captures interpurchase timing in multi-channel distribution, (2) a linear regression model that captures purchase amounts, and (3) a latent class model to explain consumer heterogeneities. We extend our model to a hierarchical Bayes model and estimate the parameters using the MCMC method.

For empirical analysis, we use single-source data that captures purchase records in both EC and real stores for a PB. We analyze two product categories, *beauty-health* category and *food* category. We confirm that the proposed model that considers competing events performs better than a single-event model that does not consider the competing events. The number of classes, 3 performed the best in terms of MSE in two categories, and the two store-centric classes and one EC-centric class are extracted. For the competing risk model, mileage affects interpurchase timing positively in almost all classes and categories, which implies that point programs can hasten consumers' purchase timing in almost all channels, classes, and product categories. On the other hand, for the purchase amounts model, mileage programs do not positively affect purchase amounts unlike interpurchase timing. Additionally, we show that we can predict the probability of purchase incidence in each channel and expected purchase amounts by changing the value of the marketing variables.

In this case, we assume that purchase behaviors in online and offline channels can be fully obtained as single-source data. If the single-source data is available, the proposed model works appropriately. However, in the marketing environment, single-source data capturing multi-channel or competing stores, such as own and competing stores, are not available in practice. That is, the censoring indicator caused by competing events cannot realistically be obtained by researchers. Therefore, the censoring problems by competing events remains. To solve this problem, in the next chapter, we focus on the missing censoring indicator problem and the Bayesian data combination approach by integrating macro information.

# Chapter 4. Bayesian Data Combination Approach Integrating Micro and Macro Data using Quasi-Bayes Method: Application to Duration Model with Missing Censoring Indicator

Missing data problems, in which complete data cannot always be obtained, are widely known to researchers in many fields such as social science or nature science (e.g., Little and Rubin, 2002). Missing or incomplete data problems are caused by non-response, censoring, truncation, or dropout. In this chapter, we focus on duration analysis for incomplete data where there is some nonignorable missingness, and censoring indicators cannot be fully observed. We address incomplete data problems caused by nonignorable missingness in which specifying the missing mechanism is difficult, and we also address incomplete data problems caused by *unknown censoring* or *missing censoring indicators* using auxiliary information obtained from other data sources.

In this chapter, we focus on the nonignorable missingness and missing censoring indicator problem, and propose solutions using a Bayesian data combination approach. As we mentioned, methods that combine micro and macro data by GMM or empirical likelihood have been extensively studied (e.g., Imbens and Lancaster, 1994; Chaudhuri *et al.*, 2008) . However, the Bayesian approach for combining micro and macro data has never been developed. Particularly, moment restrictions such as Imbens and Lancaster (1994) are not used in Bayesian methods. Then, we propose a Bayesian data combination approach using quasi-Bayesian inference (Chernozhukov and Hong, 2003) or Bayesian GMM (Yin, 2009) approach, that utilizes population-level information to identify true duration time and estimates parameters by MCMC method. Even when the optimization of the objective function is difficult, the quasi-Bayesian inference or Bayesian GMM using MCMC may still be available.

We show that analysis from incomplete data may cause biased estimates, and the proposed model can improve the accuracy of estimates by two simulation studies: Simulation 1. nonignorable missing data, and Simulation 2. a missing censoring indicator problem. From MSE and coverage, it is shown that the proposed model with auxiliary information can estimate the parameters appropriately, but the estimated results from existing models without auxiliary information would be biased and may lead to an incorrect interpretation under nonignorable missingness or missing censoring indicators. For real data analysis, we apply our model to purchase duration analysis by using purchase panel data in the Japanese marketing field. We use the *Syndicated Consumer Index* (SCI) data provided by Intage Inc. In marketing practice, researchers sometimes assume that when a customer does not purchase products for over 30 days from the last purchase, the data are deleted in the analysis, which is considered to be the same as missingness or missing censoring indicators. In the analysis, we use *tissue/toilet paper* as the product category. We estimate two models: (1) normal Bayes model not considering missingness, and (2) proposed model using Bayesian data combination approach. The results show that the coefficients of the two models are different, especially for price and gender. Especially, the normal Bayes model underestimates the effect of price. The results from the normal Bayes model contradict the previous studies and empirical knowledge for marketing practice, but the results from the proposed model have adapted to that knowledge and practice. Thus, the duration analysis from the incomplete data leads to biased estimates, and researchers should consider the incomplete observations. Then, the proposed method can improve the results effectively.

Finally, in this chapter, we discuss the missing censoring indicator problem. However, it is important for marketing managers to distinguish between customers who are censoring and those who are not (which is missing), that is, those who purchases the products in competing stores. If we consider the purchase behaviors in competing stores within the framework of duration analysis, we must address the intermittent missingness problems. Although our Bayesian combination approach integrating macro information is strong for missing data analysis, if we deal with intermittent missingness that has unobserved missing indicators, we must consider latent variable modeling. However, the usual quasi-Bayesian method or Bayesian GMM have not been expanded to latent variable modeling generally. In the next chapter, we propose a Bayesian data combination approach using a new quasi-Bayesian inference, in which latent variables are dealt with, and apply it to the interpurchase timing model with intermittent missingness.

# Chapter 5. Bayesian Data Combination Approach using a New Quasi-Bayes Method: Application to Duration Model with Intermittent Missingness

In this chapter, we focus on duration analysis for repeated events (e.g., Bijwaard *et al.*, 2006). Especially in repeated measurement data, missing data often become a problem. In duration analysis studies, this problem has been extensively considered. In this chapter, we focus on the intermittent missingness in duration analysis with repeated measurements. Under repeated duration analysis, if any missing events exist between two observed events, the observed duration is not the true duration. That is to say, we observe only the cumulated duration for two or more events. Under intermittent missingness in repeated duration analysis, neither the covariates nor the incidences of events can be observed. Moreover, the missing indicator cannot be obtained by researchers and the usual missing data analysis cannot be applied simply. Therefore, ignoring intermittent missingness will lead to biased estimates and incorrect interpretations about the effects of some important covariates. For illustrative purposes, we deal with the interpurchase timing model in marketing. Consumers purchase products from various companies or stores such as supermarkets, drugstores, and corner stores. By analyzing the purchase histories including interpurchase timing, marketers plan various marketing interventions such as coupons or direct mail. However, the purchase data consist of only customers' behaviors from the company's own stores. Therefore, data on consumers' behaviors in competing stores are unavailable. Output from analyzing this incomplete data may lead to biased estimates and incorrect decision-making. In these conditions, the Bayesian combination approach combining micro and macro data proposed in Chapter 4, is useful.

However, our method introduced in Chapter 4 cannot be applied to the duration analysis with intermittent missingness, because if we deal with intermittent missingness that has unobserved missing indicators, we must consider latent variable modeling. Therefore, we propose a Bayesian data combination approach using a new quasi-Bayes estimation method and MCMC algorithm that utilize population-level information to identify unobserved intermittent missingness. This allows us to estimate the repeated duration model with unobserved missing indicators. The proposed model consists of the following: (1) latent variable model, (2) latent missing indicator model which separates true and composite duration, (3) mixtures of duration models and (4) moment restriction from population-level information to deal with nonignorable intermittent missingness. We use a new estimation procedure that combines objective functions of likelihood and GMM simultaneously with latent variables.

From the simulation study, we show that ignoring the intermittent missingness in repeated measurement data may result in severely biased estimates. Here we estimate seven models including the proposed model by Bayesian data combination approach to show the performance of the proposed model. From MSE, the proposed model performs the best totally in all models, and the results show that the existing models missspecify the model structure. For real data analysis, we applied the proposed model to interpurchase timing in marketing, in which we can trace the whole purchase histories observed both in own stores and competing stores. We use the SCI data provided by Intage Inc in Japan as in Chapter 4. Then, we select and use only the purchase histories observed in own stores to mimic the situation in database marketing. That is, though the scanner panel data are recorded for purchase histories in competing store chains, we regard it to be a database from a particular store, which is incomplete and lacks information on competing stores. In the analysis, we use the purchase data of the haircare category that consists of shampoo, hair rinse, and hair treatment. We confirm that the proposed model can estimate the coefficients of the duration model appropriately compared to the results from the complete data.

Our model can be applied to other issues in marketing. For example, we apply our model to internet marketing using web access data. In internet circumstances, the complete data on consumer's website browsing behaviors cannot be collected because consumers visit websites of competing companies and purchase products. On the other hand, the proposed model can be applied to other research fields such as social and natural sciences. For example, in medical statistics, researchers often use longitudinal data about clinical trial for patients, but such data often record histories within the limited medical institution and patients may go to another clinic or take over-the-counter drugs. In this situation, researchers may underestimate the effects of therapy programs, since there exists unobserved events between observed events. Additionally, in economics, researchers use panel data on factors such as job employment, marriage, and wages. Here, incomplete data problems can occur in the same way. We can strengthen incomplete observed data using population-level information from government statistics or other research institutes.

# Chapter 6. Conclusion

In this paper, we focused on duration analysis with incomplete data problems generated from purchase behavior in competing stores or on EC sites. In Chapter 3, we dealt with the duration analysis considering competing events and applied the model to interpurchase timing models in online and offline channels. We showed that if there are any competing events in duration analysis, the results that do not consider the competing events will be biased. In Chapter 4, we focused on the missing censoring indicator problems in duration analysis and proposed its solution by integrating macro information. Then, we incorporated macro-level information into micro-level data modeling by using a data combination approach. By integrating macro information, we estimated parameters composing the restriction. Additionally, we proposed the Bayesian estimation using MCMC, which allowed us to estimate parameters flexibly without numerical optimization. In Chapter 5, we proposed a duration model with intermittent missingness in repeated events. We dealt with unobserved intermittent missingness using a new quasi-Bayesian inference with hybrid posterior incorporating population-level information. We applied the proposed model to interpurchase timing in marketing. We confirmed that the proposed model can estimate the coefficients of the duration model appropriately compared to the results from the complete data.

Government statistics, such as the census and questionnaire surveys with finite sample numbers, can be used as auxiliary information. If these data have a finite sample size, the stochastic information can be used for the auxiliary information in the same manner (Imbens and Lancaster, 1994). Recently, government statistics such as the *Family Income and Expenditure Survey* or the *Consumer Confidence Survey* have been made available to the public, and custom-made aggregation services are available in Japan. In the future, auxiliary information will be open and available to researchers and to private companies. We consider that our approach will be helpful and powerful. Additionally, company databases, that provide consumer products, only record the purchase histories of the companies' brands, not of competing brands. In this situation, our model assists marketing managers in the use of auxiliary information, such as the market share or large-scale syndicated surveys, to develop subsequent marketing strategies.

# Bibliography

Bijwaard, G. E., Franses, P. H., and Paap, R. (2006). Modeling purchases as repeated events. *Journal of Business & Economic Statistics*, **24**, 487-502.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B*, **70**, 311-328.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, **115**, 293-346.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian data analysis 3rd.* Chapman & Hall/CRC.

Helsen, K. and Schmittlein, D. C. (1993). Analyzing duration times in marketing: Evidence for the effectiveness of hazard rate models. *Marketing Science*, **12**(4), 395-414.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2005). *Bayesian survival analysis.* John Wiley & Sons, Ltd.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies.* **61**, 655-680.

Jain, D, C., and N, J. Vilcassim. (1991). Investigating household purchase timing decisions: A conditional hazard function approach. *Marketing Science*, **10**, 1-23.

Kalbfleisch, J. D., and Prentice, R. L. (2002). Competing risk and multistate models, in *The statistical analysis of failure time data 2nd* . John Wiley & Sons.

Klein, J. P., and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data 2nd.* Springer Science & Business Media.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data.* John Wiley & Sons.

Ridder, G., and Moffitt, R. (2007). The econometrics of data combination. in *Handbook of Econometrics*, **6**, 5469-5547.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, **4**, 191-207.