TM&ARG

Discussion Paper No. 127

Social Media and the Diffusion of an Information Technology Product

Yinxing Li and Nobuhiko Terui

August 2016

TOHOKU MANAGEMENT & ACCOUNTING RESEARCH GROUP

GRADUATE SCHOOL OF ECONOMICS AND
MANAGEMENT TOHOKU UNIVERSITY
KAWAUCHI, AOBA-KU, SENDAI
980-8576 JAPAN

Social Media and the Diffusion of an Information Technology Product

Yinxing Li and Nobuhiko Terui[1]
Tohoku University

August 2016

Abstract

The expansion of the Internet has led to a huge amount of information posted by consumers online through social media platforms such as forums, blogs, and product reviews. These text data are useful especially when numeric sales data are not enough, as is typically the case with new product diffusion. This study proposes a diffusion model that accommodates pre-launch social media information and combines it with post-launch sales information in the Bass model to improve the accuracy of sales forecasts. The model is characterized as the extended Bass model, with time varying parameters whose evolutions are affected by the consumer's communications in social media.

Specifically, we first extract information from social media to build variables, such as the number of positive and negative comments, and also latent topics. These data are fed as key parameters in the diffusion model's evolution process for the purpose of plugging the gap between the time-invariant key parameter model and that of observed sales.

We examine several models using text analysis techniques, e.g., sentiment analysis for counting numbers of positive and negative comments and topic analysis by topic model to extract relevant topics. These results are then compared with the conventional Bass model using only post-launch sales data.

An empirical study of the first-generation iPhone during 2006 and 2007 shows that the model using additional variables extracted from sentiment and topic analysis on BBS performs best based on several criteria, including DIC (Deviance Information Criteria), marginal likelihood, and forecasting errors of holdout samples. We discuss the role of social media information in the diffusion process for this study.

Keywords: Bass Model, Diffusion, Hierarchical Bayes Model, Predictive Density, Social Media Data, Text Analysis, Sentiment Analysis, Time Varying Parameter, Topic Model

[1]Terui acknowledges the grant by JSPS KAKENHI Grant Number (A)25245054.

1. Introduction

The expansion of the Internet has led to massive information posted by consumers online through social media such as forums, blogs, and product reviews. This provides an opportunity for firms to know consumers' product expectations and evaluations without the need for a direct survey. Using text mining, Grimes (2008) found that 80% of business-relevant information originates primarily as unstructured text.

A growing number of studies have examined the influence of user-generated content in marketing. Lee and Bradlow (2011) have proved that customer reviews can complement existing methods for generating attributes used in marketing analysis by comparing expert guides and consumer surveys. Netzer et al. (2012) have utilized large-scale, consumer-generated data on the Web to understand consumers' top-of-mind associative network of products and the implied market structure insights. Moe and Trusov (2011) showed that when studying the effect of consumer's ratings, the potentially endogenous relation between sales and ratings must be considered. Tirunillai and Tellis (2012) used a naïve Bayes classifier and support vector machine to classify user-generated online content to positive news and negative news and incorporated this information into a financial econometric model to forecast stock returns.

In recent years, online product reviews have taken on a larger role in the consumer decision process. Not only do consumers prefer buying products online but they also rely increasingly on others' online comments. Considering that only a limited number of samples are available for conventional new product diffusion models, online conversations, such as SNS, blogs, and BBS, are becoming very popular and could have complementary roles. Combined with word-of-mouth (WOM) data, these could improve forecasting performance through a deeper understanding of the market structure.

In this study, we use not only sales data but also user-generated online content (or online comments) to describe and forecast the diffusion process of a new product, where online WOM data is plugged into the model as covariates for affecting the change of key parameters over time. From the modeling perspective, our model is characterized as a diffusion model with a time-varying parameter. This parameter variation of the diffusion model has been discussed for several reasons; for example, as a competitive activity, changes in marketing

practice, different segments adopting products at different times (Eliashberg and Chatterjee, 1986), specification and measurement errors (Putsis, 1998), and aggregation and omitted variables (Sarris, 1973; Judge et al., 1985). By capturing changes in consumer expectation and evaluation before and after launch, our study incorporates an additional reason for the diffusion model of an IT product where the WOM effect by BBS would be significantly present by reflecting the change of consumer expectation and evaluation before and after launch. In particular, we show the information extracted from BBS text data leads to sales, thus motivating the development of a systematic variation model with covariates constituted from BBS text. This is distinguished from stochastic process models using a Kalman filter (Bretschneider and Mahajan, 1980; Judge et al., 1985; Putsis, 1998; Xie et al., 1997), where the sources of parameter variation are not always obvious.

Our proposed models belong to the class of systematic variation models (Mahajan et al., 2000, Ch. 11) and they share the advantages with other time-varying parameter models in producing fewer forecasting errors, as was shown by Putsis (1998) and Xie et al. (1997). In addition, our models provide insights into the time variation of parameters to guide the transition.

We evaluate parameter estimates using a Bayesian approach, and our inference is exact in the sense of not relying on asymptotic theory. The predictive density is numerically evaluated to reflect the uncertainty of point forecasts in decisions, as recently discussed by Terui and Ban (2014) and Takada et al. (2015). This characteristic of inference is intrinsic to the new product diffusion process as it uses a limited number of data points.

In the next section, we briefly introduce the text analysis used in our study, i.e., sentiment analysis using a naïve Bayes classifier and topic analysis by LDA. In Section 3, we propose the models and explain the estimation procedure. The empirical application is reported in Section 4. We apply our model to the diffusion process of the first-generation iPhone by augmenting the information set with user-generated content from the BBS of this product. We show that temporal variables extracted from social media contain useful information not only for improving forecasts than the original Bass model but also for understanding changes in the diffusion process due to interactions among potential and actual purchasers through WOM in the BBS environment. The sentiment analysis gathers

the subjective emotional responses of consumers, and the topic analysis gathers rather subjective information, including the effects of marketing, reviews, and discussions, named by the characterization of extracted topics. We conclude our study in Section 5.

2. Text Analysis of Social Media

2.1 Sentiment Analysis

Our model utilizes user-generated information from social media on a new product. We use two methods to analyze this text data: sentiment analysis and topic analysis. We first generate the numeric information from text data by classifying users' comments into one of three comment categories: positive, negative, or neutral (no relation). In particular, the number of positive comments before launch reflects the expectations of potential customers that can lead to after-launch sales.

We use the naïve Bayes classifier for text analysis, which has been effectively applied to the marketing problem (Tirunillai and Tellis, 2012). It is a simple probabilistic model based on the Bayes theorem, with independence assumptions between words, and it is well-recognized to show good performance in text analysis.

When the vector of words **x** in a comment is given, the posterior probability $p(C_k \mid \mathbf{x})$ of classifying it to C_k (category *k*, i.e., positive, negative, or neutral) is calculated by Bayes's theorem as $p(C_k \mid \mathbf{x}) \propto p(C_k)\, p(\mathbf{x} \mid C_k)$. $p(C_k)$ is the prior probability and can be defined by calculating the share of positive comments among all comments in the training data. $p(\mathbf{x} \mid C_k)$ is the likelihood, implying the probability that this comment with the vector of words **x** happens when it belongs to C_k under the assumption of independence of word, i.e., $p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p(x_i \mid C_k)$. Then, we classify the comments using the value of $\hat{y}$ from the function below:

$$\hat{y} = \arg\max_k \; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k). \tag{1}$$

2.2 Topic Analysis

Next, we extract the "topics" from a collection of documents in social media using the latent Dirichlet allocation (LDA) model (Blei et al., 2003), which is well-established in natural language processing and applied in a variety of disciplines. The LDA model is based on the assumption that each document can be viewed as a mixture of various latent topics, where topics follow a multinomial distribution over words. Contrary to the fact that the naïve Bayes classifier assumes that one document only has one topic, LDA assumes that each document is a mixture of various topics.

More specifically, LDA is a generative model allowing sets of observations to be explained by unobserved groups, explaining why some parts of the data are similar. Denote $w_{d,i}$ as the *i-th* word in document *d* and $z_{d,i}$ the (latent) topic of the *i-th* word in document *d*. The model assumes that $w_{d,i}$ has a vocabulary (*v*) distribution in topic *k* that follows a multinomial distribution ($w_{d,i} \sim Multi(\phi_{z_{d,i}})$) and $\theta_{k,d}$ follows topic distribution ($z_{d,i} \sim Multi(\theta_d)$) in document *d*. Then, the model describes the probability that vocabulary *v* appears in document *d* and is represented as the sum of the products of topic distribution and vocabulary distribution over possible *K* ways:

$$p(v \mid d) = \sum_{k=1}^{K} p(v \mid k) p(k \mid d) = \sum_{k=1}^{K} \phi_{v,k} \theta_{k,d} \quad . \tag{2}$$

In the LDA model, the most common method to estimate latent parameter **z** is to use Gibbs sampling. However, when there is a large volume of text data like in our study, Gibbs sampling requires a lot of time to sample the parameters. Then we employ a popular way known as "collapsed Gibbs sampling," which analytically uses the natural conjugate of prior distribution to integrate out $\theta_{k|d}$ and $\phi_{w|k}$. Details of the MCMC procedure are given in the Appendix.

3. Models

3.1 Diffusion Model with Social Media Information

We use the new product diffusion model by Bass (1969) as the base model and extend it in the way of incorporating social media information. Then, we assume that the potential market size (*m*), the innovator ratio (*p*), and imitator ratio (*q*) are changing over time and their

dynamics are partially driven by temporal communications among potential users.

We expect different roles for sentiment analysis and topic models. The sentiment analysis extracts emotional and rather subjective feelings of "like" (positive) or "dislike" (negative) from consumers' BBS communications. On the other hand, topic analysis involves objective factors based on consumers' expectations and evaluations before and after the launch of a new product and their responses to marketing activity.

We employ the empirical model of Srinivasan and Mason (1986), which uses a continuous form of expression for the difference of cumulative sales ($x_t - x_{t-1}$) to define the model as

$$y_t = m_t\left[F\left(t \mid p_t, q_t\right) - F\left(t-1 \mid p_{t-1}, q_{t-1}\right)\right] + \varepsilon_t \tag{3}$$

where the cumulative density is written by

$$\begin{aligned} F\left(t \mid p_t, q_t\right) &= \frac{1-\exp\left\{-\left(p_t+q_t\right)t\right\}}{1+\frac{q_t}{p_t}\exp\left\{-\left(p_t+q_t\right)t\right\}} \\ &= \frac{1-\exp\left\{-\left(\frac{1}{1+\exp\left\{-p_t^*\right\}}+\frac{1}{1+\exp\left\{-q_t^*\right\}}\right)t\right\}}{1+\left(\frac{1+\exp\left\{-p_t^*\right\}}{1+\exp\left\{-q_t^*\right\}}\right)\exp\left\{-\left(\frac{1}{1+\exp\left\{-p_t^*\right\}}+\frac{1}{1+\exp\left\{-q_t^*\right\}}\right)t\right\}} \end{aligned} \tag{4}$$

and ε_t is assumed to follow a normal distribution $\varepsilon_t \sim N\left(0,1/\tau\right)$.

We assume that the dynamics of parameters are partly explained by extracted variables from social media on the grounds that they contain changes in consumers' emotions, expectations, and evaluations. We describe this mechanism using a hierarchical model for the parameters in addition to diffusion model (4). More specifically, for the appropriately transformed parameter vector $\theta_t = \left(m_t^*, p_t^*, q_t^*\right)'$, where $m_t^* = \log m_t$, $p_t^* = \log\left(\frac{p_t}{1-p_t}\right)$, and $q_t^* = \log\left(\frac{q_t}{1-q_t}\right)$, and covariate vector z_t (including constants and variables) by analyzing social media data. We define the hierarchical model as

$$\theta_t = Gz_{t-1} + \eta_t \tag{5}$$

where z_{t-1} is a covariate vector constituted from social media data, η_t is the three-dimensional vector of error terms and assumed to follow a normal distribution $\eta_t \sim N_3(0,\Sigma)$, where $\Sigma = diag\left(\tau_1^{-1},\tau_2^{-1},\tau_3^{-1}\right)$. That is, the models are canonically represented by hierarchical nonlinear regression models.

We denote the static Bass model as Model 1, where we set the covariate as $z_{t-1}=1$. Then, the first model (Model 2) uses three quantities to describe the comments: total number of comments and numbers of positive and negative comments. We define the covariate as $z_{t-1}=(1,s_{t-1},p_{t-1},n_{t-1})'$, where s_{t-1} means the number of comments in t−1. p_{t-1} and n_{t-1} are, respectively, the numbers of positive and negative comments in *t*−1.

The second model (Model 3) is defined when the covariate comprises constant terms and extracted topics as $z_{t-1}=(1,T_{1t-1},T_{2t-1},T_{3t-1})'$, where T_{it-1} is the number of *i*-th topics at *t*−1. Although there are some approaches about how to select the number of topics, we assume three topics for simplicity. The third proposed model (Model 4) combines Models 2 and 3 by setting $z_{t-1}=(1,s_{t-1},p_{t-1},,n_{t-1},T_{1t-1},T_{2t-1},T_{3t-1})'$.

The proposed models are characterized as the hierarchical regression model whose parameters evolve over time, synchronizing with temporal changes of variables constructed from social media communications at a previous time. Since the first column of coefficient matrix G is the vector of parameters of the static Bass model (Model 1), these models are nested and include the original Bass model as a special case when additional text information has no information on parameter evolutions in (5).

3.2 Posterior Density for Model Parameters

In terms of (4) and (5), the model is canonically described as a hierarchical nonlinear regression model with time-varying parameters. We use a Bayesian MCMC method to estimate parameters since the procedure of hierarchical regression models has been well-established and the necessary conditional posterior densities are available in closed form, except the time-varying parameter $\{\theta_t\}$. Then we can proceed with relatively efficient

computational steps by combining Metropolis–Hasting sampling for three key parameters, with Gibbs sampling for the other parameters.

In fact, the joint posterior density of model parameters is formulated by

$$\begin{aligned} p\left(\{\theta_t\},\tau,G,\Sigma\,|\,\{y_t,t\},\{z_t\}\right) \propto\; & p\left(\{\theta_t\}\,|\,\{y_t,t\},\tau\right) p\left(\tau\,|\,\{y_t,t\},\{\theta_t\}\right) \\ & \times p\left(G\,|\,\{\theta_t\},\{z_t\},\Sigma\right) p\left(\Sigma\,|\,\{\theta_t\},\{z_t\},G\right) \end{aligned} \tag{6}$$

where the right-hand side of first line of (6) means the product of conditional posterior density for parameters in the diffusion model (4) and the second line means those for hierarchical model (5).

The sampling scheme of MCMC for this model is as follows. Starting from the initial parameter values, once $\{\theta_t\}$ is generated, the posterior density of hierarchical models $p\left(G\,|\,\{\theta_t\},\{z_t\},\Sigma\right)$ and $p\left(\Sigma\,|\,\{\theta_t\},\{z_t\},G\right)$ are available in closed forms, i.e., normal and inverted gamma distributions with given hyper parameters. On the other hand, the likelihood function $p\left(\{y_t,t\}\,|\,\{\theta_t\},\tau\right)$ in (4) is combined with prior density $p\left(\{\theta_t\}\,|\,G,\{z_t\},\Sigma\right)$ from hierarchical model (5) to evaluate the conditional posterior density as

$$p\left(\{\theta_t\}\,|\,\{y_t,t\},\tau,G,\{z_t\},\Sigma\right) \propto p\left(\{y_t,t\}\,|\,\{\theta_t\},\tau\right) p\left(\{\theta_t\}\,|\,G,\{z_t\},\Sigma\right). \tag{7}$$

We employ Metropolis–Hasting sampling for this posterior density. When $\{\theta_t\}$ is given, the conditional posterior density $p\left(\tau\,|\,\{y_t,t\},\{\theta_t\}\right)$ of the right-hand side of (6) is known as an inverted gamma distribution.

Finally, the posterior density of key parameters of the Bass model at the original scale is obtained by inverse transformation of $\theta_t=\left(m^*,p^*,q^*\right)$ to $\left(m_t,p_t,q_t\right)$, i.e., $m_t=\exp\left(m_t^*\right)$, $p_t=\dfrac{1}{1+\exp\left(p_t^*\right)}$ and $q_t=\dfrac{1}{1+\exp\left(q_t^*\right)}$; then, we can evaluate the joint posterior density as

$$p\left(\{m_t,p_t,q_t\},G\,|\,\{y_t,t\},\{z_t\}\right). \tag{8}$$

The details of this algorithm, including the setting of a prior distribution, are given in the Appendix.

4. Empirical Application

4.1 Data

We use the numbers for quarterly global sales of first-generation iPhones from June 2007 to September, 2008. These data were obtained from www.statista.com and are displayed in Figure 4.1.

Figure 4.1: iPhone Sales (June 2007–September 2008)

As for social media information, corresponding to global sales data, we use "*gsmarena*" (http://www.gsmarena.com/), a well-known BBS for mobile phones, where users from all over the world put their comments regarding mobile phones in which they are interested. Users of this BBS can access information on topics for all phones and provide their own comments or discuss topics with other users. We extract social media text data on the first-generation iPhone and collect its sales data until the next-generation iPhone (iPhone 3G) is released. In the BBS of *gsmarena*, a new topic for a mobile phone is usually created when this phone is first announced to the public by the company. On January 9, 2007, Steve Jobs gave a presentation on the iPhone and a thread was created the following day. Each comment has three elements: user, date, and comment text. We extract date and comment text only because user information is not used in this study. A total of 8,121 comments uploaded between January 10, 2007 and November 24, 2007 are divided into two groups: 1,500 comments for training data and 500 comments for test data. The daily text data are converted to quarterly data and we use the first four quarters for estimating models while the last two quarters are kept for holdout samples.

4.2 Sentiment Analysis

A conventional sentiment analysis uses two categories—positive and negative comments—to classify comments. However, this BBS usually has many unrelated comments such as questions and discussions. Then we classify training data into three groups: positive, negative, and no relation. We confirmed that the no relation group improves the accuracy of classification.

We classify all 8,121 comments of training data and then test the accuracy using test data comprising 500 comments. The prior distribution and accuracy are given in Table 4.1. The prior distributions $p(C_k)$ for three categories are calculated by counting the number of positive comments in training data by interpreting each comment manually, i.e., by making a dictionary: 39% for positive, 33.1% for negative, and 27.8% for neutral. The accuracy is defined as the ratio of the number of hits over the number of comments in the test data of 500 comments. We found 94.2% of positively predicted comments in test data to be truly positive, with hit rates of 90% and 84.7%, respectively, for negative and neutral predicted comments. This shows the high precision of our dictionary for sentiment analysis.

Table 4.1: Summary of Naïve Bayes Classifier

Figure 4.2 shows time-series plots for the numbers of positive and negative comments used in our study. The movements of these numbers are synchronized with those of sales with the lag of one period; thus, these can be leading indicators for sales.

Figure 4.2 Time-series Plots of Positive and Negative Comments

---

4.3 Topic Analysis

In the topic model, we set the number of topics as three and the Bayesian Collapse Gibbs sampling algorithm is used to estimate the model (see the Appendix for details). The number of M = 4,000 samples is used to evaluate posterior probability after discarding the previous 1,000 samples as a burn-in period. This computation needed a tremendously long time of about one week. Table 4.2 shows the top words for each topic.

Table 4.2: Top Words for each Topic

First, the topic number in the table means the estimate of probability that topic *k* is in all

D documents, i.e., the posterior mean $E\left(\theta_{k,\cdot} = \sum_d \theta_{k,d} / D\right)$. Topic 1 has the largest probability of 0.554 and is dominant compared with other topics with almost half the probability. Next, the top twelve words for each topic are given in their order of frequency. The number next to each word refers to its frequency in the document. According to these classifications, we can easily characterize each topic. Topic 1 contains "phone," "n95" (Nokia cellphone), "nokia," "good," etc., which are used regarding reviews. Topic 2 includes the words "ur," "me," "install," "tell," "help," "thanks," "plz," and other words used in the context of discussions. Topic 3 contains "apple," "will," "market," "Europe," "us," "released," "uk," and other words in the context of marketing. Thus, we call Topic 1 "Reviews," Topic 2 "Discussion," and Topic 3 "Marketing."

Figure 4.3 shows time-series plots for the number of words in each topic and global sales data.

Figure 4.3: Number of Words in each Topic

The figure shows that the number of topics, especially Reviews and Marketing (Topics 1 and 3) leads to sales with a one-period lag and suggests that they could be leading indicators for accurate sales forecasting. In contrast, Discussion (Topic 2) is synchronized with sales.

4.4 Model Comparison

The models were estimated by generated sample of $\theta_t^{(k)}, k = 1, ..., M$, and we used *M* = 5,000 samples for constructing the posterior density after discarding the previous 1,000 samples as the burn-in period. This required many iterations and almost 20 hours for the MCMC sequence to converge for Model 4. Other models did not need such a high number of iterations. We confirmed their convergence using the Geweke's test (Geweke, 1992), with a significance level of 95%. In the above, the non-informative diffuse prior was set for parameters. The specification of prior distribution and necessary conditional posterior densities are provided in the Appendix.

Five models defined in previous section are compared based on three measures: log of marginal likelihood (LMD), deviance information criteria (DIC), and root mean squared errors of forecasts for holdout samples (RMSE). The results are provided in Table 4.3.

Table 4.3: Model Comparison

First, the models with time-varying parameters perform significantly better than the static model (Model 1) in terms of respective criterion. This means that BBS contains useful information to describe the new product diffusion process. Among the dynamic models, the model with sentiment analysis (Model 2) is supported slightly better than the model with topics (Model 3). However, their combined model (Model 4) performs best.

4.5 Parameter Estimates

Table 4.4 shows the estimates of coefficient matrix G in (5) for the respective models. The first column of each table show the estimates of constant term of time evolution model for transformed parameters. This means the estimation of transformed parameters for the original Bass model. Other columns show the time-varying factors of transformed parameters induced by several variables constructed by sentiment and topic analysis and using BBS information.

Table 4.4: Parameter Estimates

The estimates are defined as posterior mean and 95% CI (credible interval) with the boundary created by upper and lower 2.5 percentiles of posterior density given in parentheses below the estimate. First, the estimates of the intercept term are shown in the column denoted "0" in the tables, and the time-invariant part of key parameters $\theta_t = \left(m_t^*, p_t^*, q_t^*\right)'$ is significantly estimated in the sense that 95% CI does not include zero. This means that the original Bass model by itself (Model 1) and the parts of original Bass model for other models are well estimated if we interpret them when they are inversely transformed. They drive a

smooth orbit of sales and the mechanism of the original Bass model works for all models as an intrinsic part of the diffusion process. Second, three topics extracted by topic model affect all parameters not only when used solely, i.e., in the case of Model 3, but also together with sentiment variables, i.e., Model 4. Third, the numbers of total, positive and negative, comments from the sentiment analysis almost explain the changes of m_t^*, p_t^*, but it does not hold for q_t^* from the results of Models 2 and 4, and we could explain that imitators rely not on emotional but subjective factors such as review, discussion, and marketing by other people.

Next, we consider the result of the most supported model (Model 4) in more detail. The time transition equation of m_t^* has significant positive coefficients on all covariates and they induce a positive increase of m_t^* when they are increased. According to the magnitude of estimates due to the fact that the measurement scale is common in each category, the order of effectiveness is as follows: (i) Topic 2 (discussion) > Topic 3 (marketing) > Topic 1 (review) for topic models and (ii) positive comments > negative comments > number of comments.

As for p_t^*, Topic 3 (marketing) has a positive effect, implying that the recognition of a new product through the firm's marketing activity would be creating new innovators at each period. On the other hand, Topic 1 (review) and Topic 2 (discussion) have negative coefficients, and we interpret that active review and discussion are reflecting the circumstance where there is too much product information, which would discourage innovative offerings on the part of consumers.

Finally, the q_t^* equation has positive significant estimates of coefficient on three topic variables; however, there is no effective variable in sentiment analysis. This means that the change of imitator would be induced not by the subjective emotional factors in sentiment analysis, but rather by objective product evaluation through review and discussion in topic analysis.

4.6 Temporal Change of Key Parameters

Figure 4.4 shows the posterior density of key parameter estimates for Model 4, where the

estimates of $\theta_t = \left(m_t^*, p_t^*, q_t^*\right)'$, t = 1…, 4 (estimates), and 5, 6 (forecasting) are inversely transformed to their original scales $\left(m_t, p_t, q_t\right)$ for the model interpretations. Most posterior densities are skewed by the form of log and logistic transformations. We share this skewness throughout the models and then define the estimates of the original key parameters by the median, which provides a more reasonable point estimate in the case of a skewed distribution.

Figure 4.4: Posterior Density of Key Parameters

The temporal change of key parameter estimates is depicted in Figure 4.5.

Figure 4.5: Temporal Change of Key Parameters

First, as for the market potential parameter *m*, the dynamic models have larger values than the original Bass model (Model 1) and show the same pattern, growing with the peak at the third period and declining after that, although the levels are different with the highest potential numbers for Model 4. The variables constructed by sentiment analysis and topic models have similar effects on determining the orbit of m_t.

Second, the innovator *p*'s estimates take similar low values and Model 3 produces rather fluctuating innovator estimates with the highest at the third period. The estimates of imitator q_t are heterogeneous among models. In particular, Models 2 and 3 have relatively lower values, around 0.6–0.8 and, thus, share higher values with Model 4.

4.7 Forecasting

Bayesian inference in this model constitutes unconditional predictive density. The predictive density for *s*-step ahead forecast y_{T+s} can be written by the model structure as

$$p\left(y_{T+s} \mid \text{Data}\right) = \int p\left(y_{T+s} \mid \theta_{T+s}, G, \Sigma\right) p\left(\theta_{T+s} \mid G, \text{Data}\right) p\left(G \mid \Sigma, \text{Data}\right) p\left(\Sigma \mid \text{Data}\right) d\theta dG\Sigma \quad (9)$$

where y_{T+s} is the *s*-step ahead forecast and θ_{T+s} is the corresponding time-varying parameter vector. The integration in (9) can be numerically evaluated by efficient Monte Carlo methods, i.e., by sequentially generating samples in addition to MCMC iterations for posterior density. That is, starting from some initial values of $\left(G^{(0)}, \Sigma^{(0)}\right)$, we take the steps: (i) $\Sigma^{(k)}$ is generated from $p\left(\Sigma \mid \text{Data}\right)$; (ii) $G^{(k)}$ is generated from ; $p\left(G \mid \Sigma^{(k)}, \text{Data}\right)$, (iii) ${\theta_{T+s}}^{(k)}$ is generated from $p\left({\theta_{T+s}}^{(k)} \mid G^{(k)}, \text{Data}\right)$ using equation (6); and (iv) ${y_{T+s}}^{(k)}$ is generated from $p\left({y_{T+s}}^{(k)} \mid {\theta_{T+s}}^{(k)}, G^{(k)}, \Sigma^{(k)}\right)$ using equation (5). We note that when the diffusion model contains an explanatory variable of "time," the structural equation is easily updated by shifting *T* to *T*+*s*, without assuming scenarios for future explanatory variables, as is done by Takada et al. (2014).

Figure 4.6: In-Sample and Out of Sample Fit

Figure 4.7: Predictive Density for Model 4

Figure 4.6 shows the generated forecasts of respective models from the fifth and sixth periods, where in-sample fits from the first to fourth periods are also depicted and where the forecasts are defined as the mean of predictive density. The predictive densities for Model 4 are shown with observation by the x-mark in Figure 4.7. They are well-defined and accommodate holdout observations in the center of density, implying that the forecast using predictive density has high precision. In addition, we can evaluate the predictive interval easily by evaluating percentiles of predictive density.

5. Concluding Remarks

The fusion of numeric structured data and unstructured text data is a challenging issue in big data analysis and it is also demanded in marketing research.

In this article, we proposed time-varying diffusion models to accommodate social media information. These models belong to the class of systematic variation models and provide useful insights on parameter variations, where we enlarge the information set regarding the diffusion process using product-related BBS text data from before and after the launch of a new product. We use this information based on the recognition that communications in BBS reflect changes in consumer expectations before launch as well as changes in product evaluations of not only the product itself but also the marketing activity and its competitive products. In particular, the communications among potential customers waiting to launch innovative IT products used in our study contain a sort of a proxy variable for consumers' expectations before launch and changes in perception and evaluation after launch.

Our proposed models contain additional variables constituted from BBS text data by applying two approaches for analyzing text data, i.e., sentiment analysis and topic analysis. These variables are used as covariates to explain parameter temporal transitions. These analytical techniques are expected to extract subjective emotional variables and evaluation-based objective variables in BBS, respectively. The empirical study showed that these additional variables lead to an improvement in the model fit and precision of forecasting by filling a gap between smooth transitions of sales generated by a static diffusion model and realized sales, and they provide the roles of constructed variables in text analysis for the change in model parameters. For example, both of the emotional sentiment variables, rather than objective topic variables, have positive effects on market potential; on the other hand, topic variables affect the innovator and imitator with reasonable interpretation while sentiment variables affect the change of innovator transition, but not for that of the imitator. We also showed that the proposed model with the augmented information set produces a great improvement in the precision of forecasting.

IT products, such as the iPhone, continue to evolve and, together with growing social media networks, we can consider the extension of our model to successive product generations, including second- and third- generation products, by using the models of Norton and Bass (1987), Mahajan and Muller (1996), Kim et al. (2000) and others. Future research can investigate into this problem.

Reference

Bass, F. M.(1969), "A New Product Growth for Model Consumer Durables," *Management Science*, 15, 215-227.

Bass, F. M.(2004), "Comments on "A New Product Growth for Model Consumer Durables"," *Management Science*, 50, 1833-1840.

Blei, D.M., Ng A.Y. and Jordan, M.I.(2003) "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D. M. (2012), "Introduction to Probabilistic Topic models," *Communications of the ACM* ,55, 77-84.

Bretschneider, S.H., and Mahajan, V. (1980), “Adaptive Technological Substitution Models,” *Technological Forecasting and Social Change*, 18, 129-139.

Eliashberg, J. and Chatterjee, R. (1986), “Stochastic Issues in Innovation Diffusion Models with Stochastic Parameters,” Mahajan. et al. ed. *Innovation Diffusion Models of New Product Acceptance*, 151-203, Cambridge, MA.

Che, H. Sudhir, K. and Seetharaman, P.B. (2007), "Bounded Rationality in Pricing Under State-Dependent Demand: Do Firms Look Ahead, and if So, How Far?," *Journal of Marketing Research*, 434-449.

Duan, W. and Gu, B, and Whinston, A.B. (2008), " The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry," *Journal of Retailing*, 84, 233-242.

Geweke, J. (1992). Evaluating the accuracy of sampling‐based approaches to the calculation of posterior moments. In Bayesian Statistics 4, Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), 169‐193. Oxford: Oxford University.

Press. Ikonomakis, M. and Koisiantis, S. (2005), "Text Classification Using Machine Learning Techniques," *WSEAS Transactions on Computers*, 4, 966-974.

Iyengar, R. and Van de Burde, C.(2011) "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30, 195-212.

Iwata,T. and Sawada H. (2013), "Topic model for analyzing purchase data with price information," *Data Mining and Knowledge Discovery*, 26, 559-573.

Judge, George, G. et al. (1985), *The Theory and Practice of Econometrics* (2nd ed.), Wiley, New York.

Kim, N. Chang, D. and Shoker, A.(2000), “Modeling Inter-Category Dynamics for a Growing

Infromation Technology Industry: The Case of the Woreless Telecommunications Industry," *Management Science*, 46, 496-512.

Lee, T.Y. and BradLow, E.(2011), "automated marketing research Using Online Consumer Reviews," *JournaL of Marketing Research*, 48, 881-893.

Mahajan, V. and Muller, E.(1996), "Timing, Diffusion, and Substitution of Successive Generations of Technological Innovations: The IBM Mainframe Case," *Technological Forecasting and Social Change*, 51, 213-224.

Mahajan, V., Muller, E. and Wind, Y. eds(2000), *New Product Diffusion Models*, Kluwer Academic Press, Boston.

Moe, W. and Trusov,M.(2011), "The Value of Social Dynamics in Online Products Ratings Forums," *Journal of Marketing Research*, 49, 444-456.

Netzer, O., Feldman, R. and Goldberg, J. and Freskko, M. (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, vol. 31, pp. 522-543, 6 2012.

Norton, J. A. and Bass, F. M. (1987), "A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products," *Management Science*, 33, 1069-1086.

Putsis, W.P. (1998), "Parameter Variation and New Product Diffusion," *Journal of Forecasting*, 17, 231-257.

Sarris, A.H. (1973), "A Bayesian Approach to Estimation of Time Varying Parameter Regression Coefficients," *Annals of Economic and Social Measurement*, 6, 31-311.

Seth Grimes, C.(2008), "Unstructured Data and the 80 Percent Rule," *Clarabridge Bridgepoints*, Q3, 11-111.

Takada, H. Saito, K., Terui, N. and Yamada, M. (2015), "The Ubiquitous Model for Dynamic Diffusion of Information Technology," Discussion Paper of DSSR No.35, Graduate School of Economics and Management, Tohoku University.

Terui and Ban (2014), "Multivariate Structural Time Series Models with Hierarchical Structure for Over-dispersed Discrete Outcome," *Journal of Forecasting*, 33, 376-390.

Tirunillai, S and Tellis, G.(2011)"Does Online Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance," *Marketing Science*, 31, 198-215.

Xie, J. Song, M., Sirbu, M. and Wang, Q. (1997), "Kalman Filter Estimation of New Product Diffusion Models," *Journal of Marketing Research*, 34, 378-393.

Appendix: MCMC Algorithm

I. LDA Topic Model

Under the prior distributions of Dirichllet distributions on hyperparameter α, β $\theta_d \sim Dir(\boldsymbol{\alpha})$ $(d = 1, \ldots, M)$, and $\phi_k \sim Dir(\boldsymbol{\beta})$ $(k = 1, \ldots, K)$, where M is the number of documents, and K means the number of topics. Following Griffiths and Steyvers (2004), we set the vectors with element of 50/K for $\boldsymbol{\alpha}$ and 0.1 for $\boldsymbol{\beta}$ respectively. Then, the collapsed Gibbs Sampler provides the posterior density in the closed form,

$$p\left(z_{d,i} = k \mid w_{d,i} = v, \mathbf{w}^{\backslash d,i}, \mathbf{z}^{\backslash d,i}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{n_{k,v}^{\backslash d,i} + \beta_v}{n_{k,\cdot}^{\backslash d,i} + \sum_{v'} \beta_{v'}} \frac{n_{d,k}^{\backslash d,i} + \alpha_k}{n_{d}^{\backslash d,i} + \sum_{k'} \alpha_{k'}},$$

where $w_{d,i}$ means word i in document d, $z_{d,i}$ is latent topic of word i in document d, θ_d means topic distribution of document d, and ϕ_k is vocabulary distribution of topic k. $\mathbf{w}^{\backslash d,i}$ is all words from the text data except word $w_{d,i}$, $\mathbf{z}^{\backslash d,i}$ is all topics except $z_{d,i}$, n_d is the number of words in document d, $n_{k,v}^{\backslash d,i}$ means frequency of word v in topic k except word i in document d, and $n_{k,\cdot}^{\backslash d,i} = \sum_{v} n_{k,v}^{\backslash d,i}$.

II. Proposed Model

1. Prior Distributions

Parameter	Setting
$m_k \sim N(\mu_{m0}, {\tau_{m0}}^{-1})$	$\mu_{m0} = 0, \tau_{m0} = 0.3$
$p_k \sim N\left(\mu_{p0}, {\tau_{p0}}^{-1}\right)$	$\mu_{p0} = 0, \tau_{p0} = 0.1$
$q_k \sim N\left(\mu_{q0}, {\tau_{q0}}^{-1}\right)$	$\mu_{q0} = 0, \tau_{q0} = 0.1$
$\tau \sim IG(\alpha, \beta)$	$\alpha = 3, \beta = 10$
$\tau_j \sim IG(\alpha, \beta)$	$\alpha = 3, \beta = 10$
$\gamma \sim IG(\alpha, \beta)$	$\alpha = 3, \beta = 10$

2. Conditional Posterior Distributions

(1) $G_{jk}|\{\theta_t\},\{z_t\},\Sigma$

$$N\left(\left(n\tau_j+\ \tau_{j0}\right)^{-1}\left(\tau_j\sum_{i=1}^{n}\left(\theta_{jt}^{*}\ -\sum_{z=1}^{K\neq k}\left(G_{jz}z_i\right)\right)G_{jk}^{-1}+\mu_{j0}\tau_{j0}\right),\left(n\tau_j+\tau_{j0}\right)^{-1}\right)$$

(2) $\tau_j|\ \{\theta_t\},\{z_t\},G$

$$IG\left(\alpha+\mathrm{n}/2,\beta+\sum_{i=1}^{n}\left(\theta_{jt}^{*}-\sum_{z=1}^{K}\left(G_{jz}z_i\right)\right)^{2}/2\right)$$

(3) $\left\{\theta_t\right\}|\left\{y_t,t\right\},\tau,G,\left\{z_t\right\},\Sigma$

The conditional posterior density of θ_t is generated by Metropolis–Hastings sampling by the proposed density on the right hand side of

$$p\left(\left\{\theta_t\right\}|\left\{y_t,t\right\},\tau,G,\left\{z_t\right\},\Sigma\right)\propto p\left(\left\{y_t,t\right\}|\left\{\theta_t\right\},\tau\right)p\left(\left\{\theta_t\right\}|G,\left\{z_t\right\},\Sigma\right).$$

For *iter*(=1,…,R) of MCMC iterations, we use Metropolis–Hastings with a random walk algorithm,

$\theta_{jt}^{(iter)}=\ \theta_{jt}^{(iter-1)}+\ \lambda_\theta;\ \lambda_\theta\sim N(0,0.05)$, where the acceptance probability is $\alpha=\min\left(1,\frac{p\left(\theta_{jt}^{(iter)}|\{y_t,t\},\tau,G,\{z_t\},\tau_j\right)}{p\left(\theta_{jt}^{(iter-1)}|\{y_t,t\},\tau,G,\{z_t\},\tau_j\right)}\right)$, where t = 1, …, N and j = 1,2,3..

(4) $\tau|\{y_t,t\},\{\theta_t\}$

$$IG\left(\alpha+\mathrm{n}/2,\beta+\sum_{i=1}^{n}\left(\mathrm{y}_t-m_t\left(F(t|p_t,q_t)-\ F(t-1|p_{t-1},q_{t-1})\right)\right)^{2}/2\right)$$

Table 4.1: Summary of Naïve Bayes Classifier

Class	Prior(Trainning data)	Accuracy(Test data)
Positive	0.390	0.942
Negative	0.331	0.900
No relation	0.278	0.847

Table 4.2: Top Words for each Topic

Topic1 (0.554)		Topic2 (0.224)		Topic3 (0.222)	
phone	0.0093	ur	0.0117	apple	0.0063
n95	0.0072	me	0.0081	mobile	0.0059
nokia	0.0071	can	0.0080	will	0.0058
good	0.0070	fone	0.0080	market	0.0058
but	0.0069	install	0.0080	contract	0.0054
better	0.0069	tell	0.0061	network	0.0051
people	0.0054	help	0.0061	europe	0.0049
camera	0.0053	thanks	0.0060	sim	0.0048
think	0.0041	bluetooth	0.0049	us	0.0047
like	0.0041	plz	0.0048	released	0.0041
really	0.0039	installer	0.0041	uk	0.0032
because	0.0038	files	0.0040	june	0.0032

Table 4.3: Model Comparison

	RMSE	Log(ML)	DIC
Model 1	0.376	1.619	2.522
Model 2	0.101	3.059	2.365
Model 3	0.125	2.472	2.403
Model 4	0.016	5.017	2.217

Table 4.4: Parameter Estimates

Model 1

	0	num.comment	pos.comment	neg.comment	Topic1	Topic2	Topic3
m*	2.035 *	-	-	-	-	-	-
	[2.026,2.044]	-	-	-	-	-	-
p*	-3.019 *	-	-	-	-	-	-
	[-3.036,-3.002]	-	-	-	-	-	-
q*	24.473 *	-	-	-	-	-	-
	[24.511,24.438]	-	-	-	-	-	-

Model 2

	0	num.comment	pos.comment	neg.comment	Topic1	Topic2	Topic3
m*	0.145	1.173 *	0.338	0.902 *	-	-	-
	[-1.177,0.468]	[0.834,1.512]	[-0.145,0.823]	[0.336,1.469]	-	-	-
p*	-3.899 *	4.067 *	-4.539 *	-3.839 *	-	-	-
	[-4.627,-3.171]	[3.364,4.771]	[-6.016,-3.063]	[-5.277,-2.401]	-	-	-
q*	0.526	0.445	0.667	1.160 *	-	-	-
	[-0.128,1.181]	[-0.245,1.137]	[-0.218,1.553]	[0.274,2.045]	-	-	-

Model 3

	0	num.comment	pos.comment	neg.comment	Topic1	Topic2	Topic3
m*	0.624 *	-	-	-	0.320 *	0.886 *	-0.431 *
	[0.530,0.691]	-	-	-	[0.233,0.417]	[0.552,1.120]	[-0.531,-0.367]
p*	-1.139 *	-	-	-	-3.144 *	1.305	4.513 *
	[-1.215,-1.097]	-	-	-	[-3.699,-2.589]	[-0.104,2.716]	[4.218,4.807]
q*	-0.289 *	-	-	-	0.931 *	1.674 *	1.907 *
	[-0.368,-0.210]	-	-	-	[0.548,1.314]	[0.930,2.418]	[1.627,2.187]

Model 4

	0	num.comment	pos.comment	neg.comment	Topic1	Topic2	Topic3
m*	-0.473 *	0.090 *	0.805 *	0.198 *	0.330 *	0.831*	0.442 *
	[-0.498,-0.449]	[0.063,0.117]	[0.037,0.123]	[0.151,0.245]	[0.308,0.352]	[0.796,0.866]	[0.402,0.482]
p*	-1.826 *	0.609 *	-0.398 *	0.019	-0.612 *	-0.209 *	0.131 *
	[-1.862,-1.774]	[0.551,0.666]	[-0.299,-0.497]	[-0.087,0.126]	[-0.676,-0.548]	[-0.284,-0.134]	[0.017,0.244]
q*	-0.260 *	0.011	-0.023	-0.013	0.248 *	0.503*	0.964 *
	[-0.302,-0.218]	[-0.036,0.059]	[-0.099,0.052]	[-0.094,0.068]	[0.208,0.288]	[0.429,0.577]	[0.906,1.023]

Figure 4.1: iPhone Sales (June 2007–September 2008)

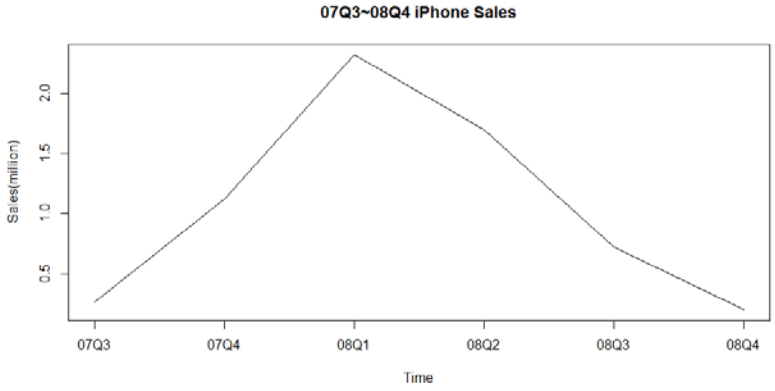

Figure 4.2 Time Series Plots of Positive and Negative Comments

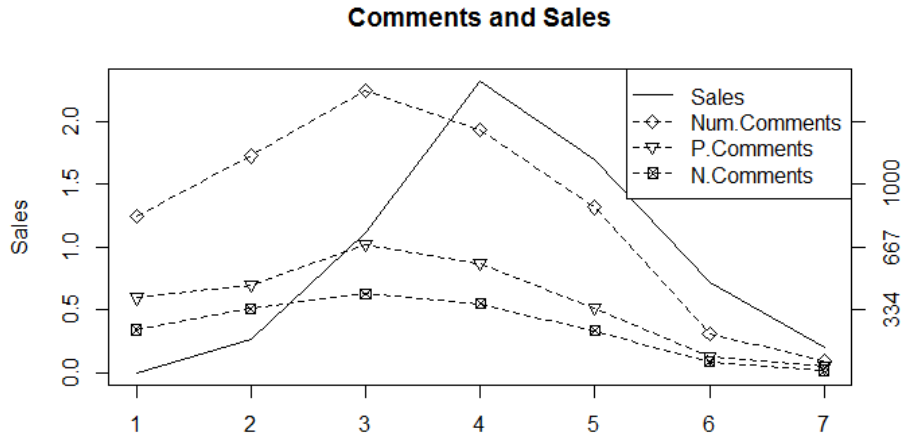

Figure 4.3: Number of Topics

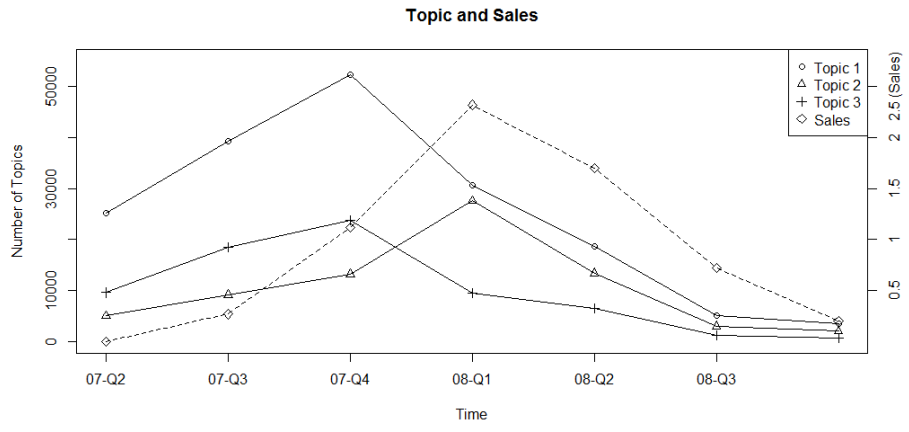

Figure 4.4: Posterior Density of Key Parameters: Model 4

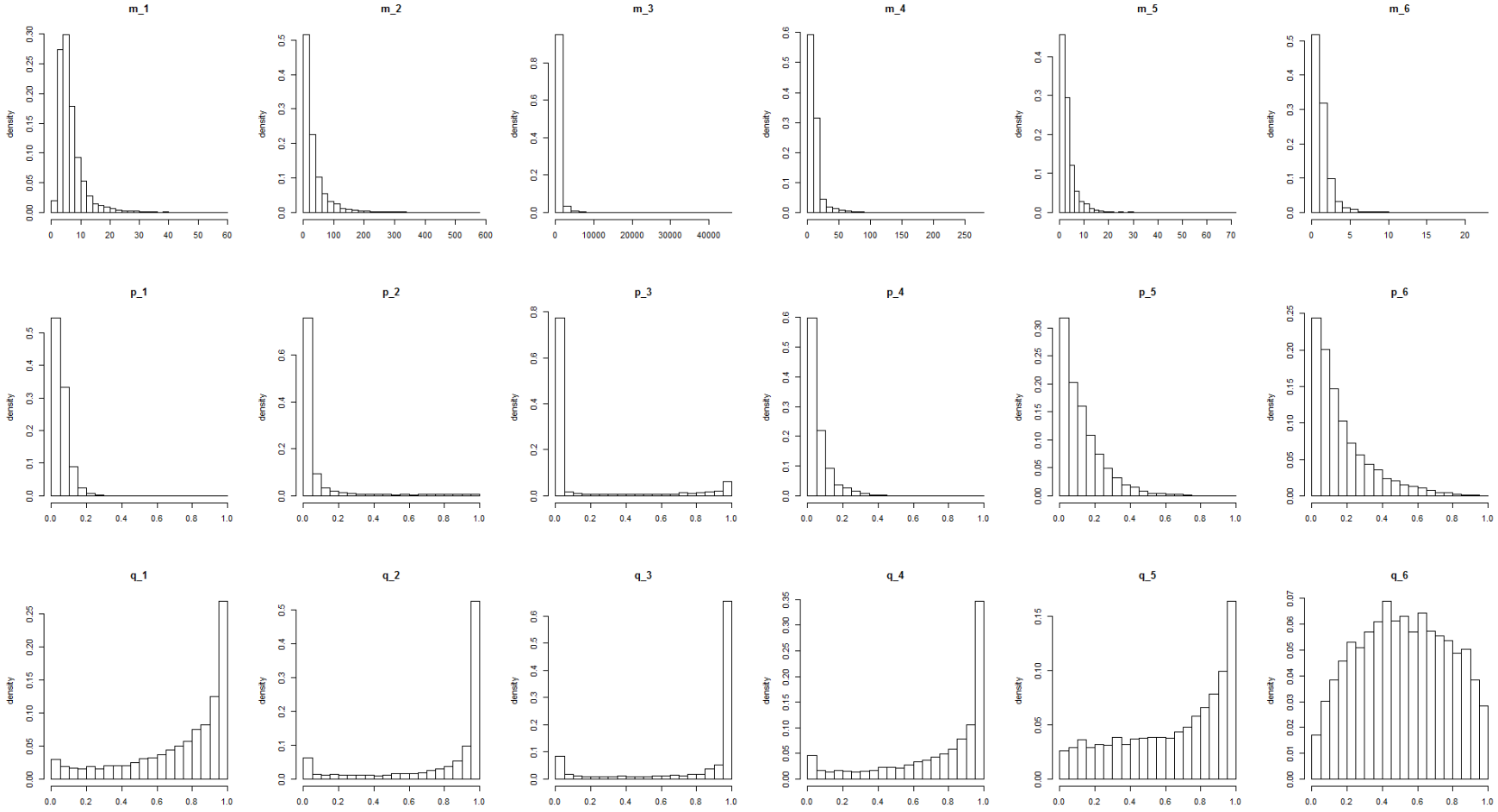

Figure 4.5: Key Parameter Estimates

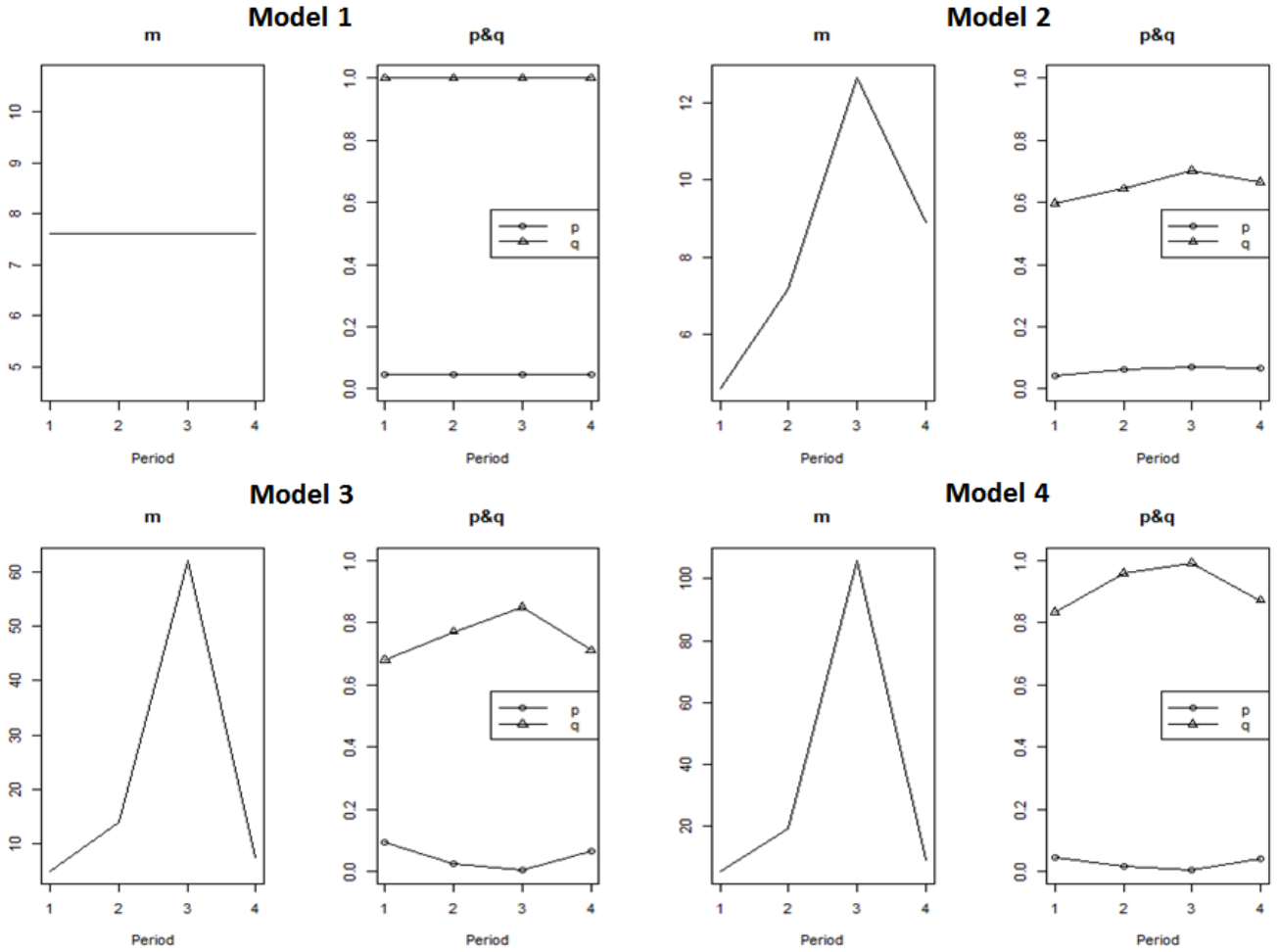

Figure 4.6: In-sample and Out of Sample Fit

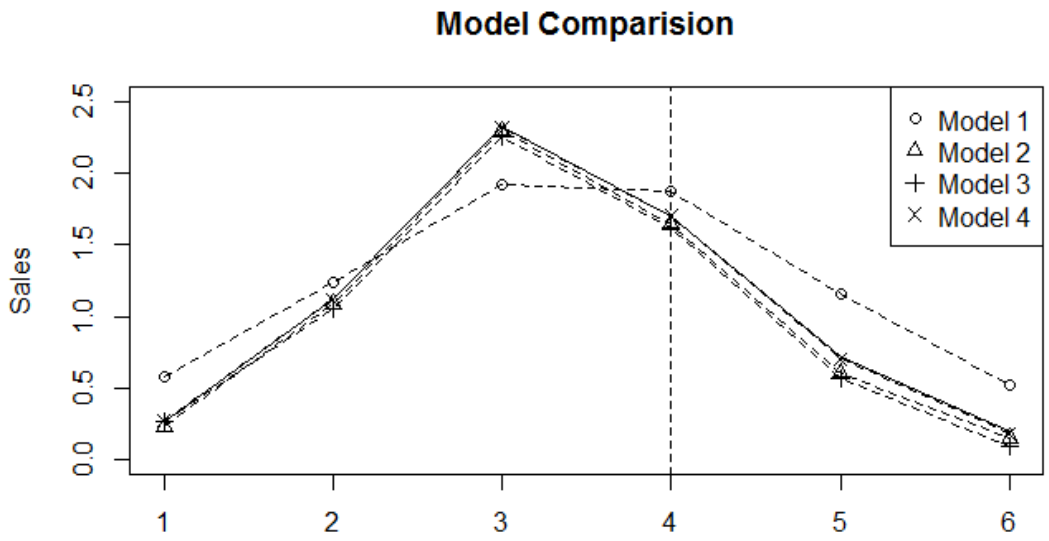

Figure 4.7: Predictive Density for Model 4