

# A HIDDEN MARKOV MODEL FOR CONVERSION RATE DYNAMICS IN ONLINE RETAIL

*Completed Research Paper*

**Ming Fan**

Foster School of Business  
University of Washington  
Seattle, WA 98195-3226, USA  
mfan@uw.edu

**Lin Hao**

Mendoza College of Business  
University of Notre Dame  
Notre Dame, IN 46556-5646, USA  
lhao@nd.edu

**Zhenzhong Sheng**

Taobao  
Hangzhou, China 310099  
hetai.szz@taobao.com

**Yong Tan**

Foster School of Business  
University of Washington  
Seattle, WA 98195-3226, USA  
ytan@uw.edu

## **Abstract**

*In this study, we use a proprietary data set from an online marketplace to study the conversion rate dynamics in online retail. We examine how seller-level covariates, such as online sellers' pricing and product strategies, marketing efforts, service responsiveness, reputation scores, product quality ratings, and other attributes, affect conversion rates. Specifically, we address the following research questions: (i) How do sellers' covariates affect their conversion rates? (ii) Is the relationship between the conversion rate and sellers' covariates state dependent? (iii) If the relationship is state dependent, what are the factors that determine the states and the state transitions? A hidden Markov model is adopted in the effort to answer these questions. The estimation results indicate that there are two states that affect conversion rate dynamics. The relationship between the conversion rate and the sellers' covariates is state dependent, that is, given the different states of sellers, the effects of the sellers' covariates on their conversion rates are different. We also estimate the thresholds between states. The results provide important guidance for sellers regarding what attributes they need to improve, to increase their conversion rates.*

**Keywords:** Conversion rate, hidden Markov model, online retail

## Introduction

In the online environment, the purchase conversion rate is an important measure of the effectiveness and success of online stores. There has been increasing interest in studying online conversion rates using click-stream data (e.g., Sismeiro and Bucklin 2004, Moe and Fader 2004). While click-stream data are useful in studying consumer-level behavior for a specific online company, seller-level analysis of the conversion rate is important in understanding how online sellers' strategies and tactics affect the performance of online stores.

In this study, we used a proprietary data set from an online marketplace to conduct seller-level analysis of conversion rates in online retail. The panel data set had a large sample of 2,766 sellers, who we studied over a seven-month period. This unique data set allowed us to examine how sellers' characteristics, such as pricing and product strategies, marketing efforts, service responsiveness, reputation scores, and product quality ratings, affect sellers' conversion rates. Specifically, we addressed the following research questions:

- What are the seller-level factors that affect the conversion rate?
- Is there any hidden state that potentially governs the dynamics of the conversion rate? Is the relationship between the explanatory variables and the conversion rate state dependent?
- If the relationship is state dependent, what are the factors that determine the states?

The organization of the paper is as follows: In the next section, we review prior literature, followed by theory development. We then discuss the empirical model, data, and variables. Further, we present our estimation results, discussions, and concluding remarks.

## Literature Review

Prior studies have examined website characteristics and consumer-behavior-related conversion rates. Mandel and Johnson (2002) show that page design can affect purchase decisions. Moe and Fader (2004) develop an individual probability model for visit-to-purchase conversion. Based on previous-visit patterns and purchases, the model predicts the subsequent visits that are likely to convert to purchases. Sismeiro and Bucklin (2004) develop and estimate a conversion rate model using click-stream data. The model predicts conversion rate by linking the purchase decision to what visitors do and what they are exposed to while visiting the seller's site.

At the seller level, Perdikaki et al. (2012) examine the effect of traffic on sales and conversion rates in brick-and-mortar stores. They break down sales volume into conversion rate and basket value and analyze the impact of traffic. They find that stores' sales volume exhibits diminishing returns to scale with respect to traffic.

Previous literatures mainly focus on the relationship between the user's clickstream pattern and the conversion probability. Our study uses a panel data set to examine the dynamics of conversion rates across a large number of startup sellers. We believe the study of conversion rates at the seller level provides unique insights into how sellers' product and pricing strategies, marketing efforts, and service responsiveness, together with reputation scores and product quality ratings, affect conversion rates. To the best of our knowledge, there has not been any literature on this specific issue. Our results provide important guidance regarding what attributes the seller needs to improve and how much the seller needs to improve them to become more likely to jump to a more favorable status.

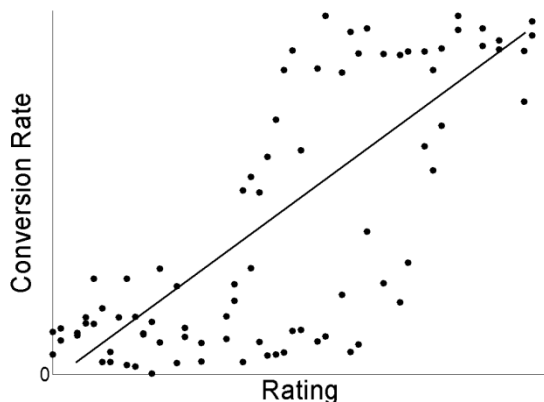
## Theory Development

In this section, we develop the theoretical foundation of the seller's conversion rate dynamics. Online sellers regularly monitor the status of their stores via a key set of variables, such as reputation scores. These variables are of great concern to sellers, because these variables highly influence consumers' ex ante perception of expected purchasing utility, which subsequently affects consumers' purchase decisions. In addition, sellers' characteristics and strategies, such as price, marketing efforts, and service responsiveness, also affect conversion rates. We call these variables—that could potentially affect the conversion rates—explanatory variables.

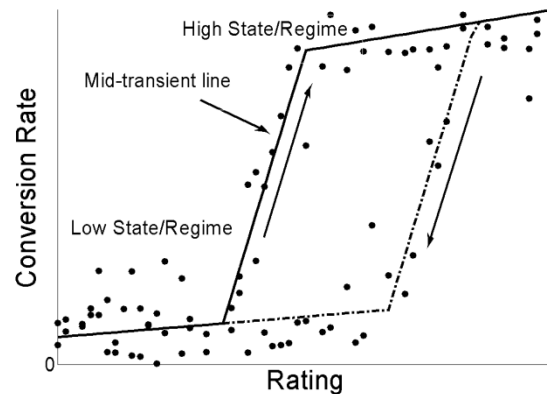
For sellers, it is important to understand how changes in the values of explanatory variables can affect conversion rates. These answers could provide important guidance for sellers to increase their conversion rates. In order to examine the relationship between conversion rates and the explanatory variables, a common method is to approximate it using one linear function, in which the conversion rate is the dependent variable and the explanatory variables are the independent variables. We take a commonly believed explanatory variable, the seller's product quality rating, for our example. Figure 1 illustrates the approximated linear function.

While a simple linear function can reveal the direction of the relationship, it neglects a significant amount of information if the consumer's purchase decision-making process follows a "regime-switching" framework. In a regime-switching model, consumers conceive of single or multiple thresholds when assessing the value of some of the explanatory variables. We call these explanatory variables "state variables."

The thresholds divide the value range of state variables into multiple regimes. In contrast to the single-regime case, where a single linear function governs the relationship between conversion rate and the explanatory variables over the explanatory variables' entire value range, the relationship between the conversion rate and the explanatory variables is different across different regimes in the regime-switching model, that is, the unit change in explanatory variables can have different effects on consumers' purchase decisions across different regimes. We take the product quality rating as an example of a state variable. When assessing the product quality rating, consumers usually conceive a minimum acceptable rating as the threshold for considering a purchase, such as 3 stars in a 5-star rating scale system. Thus, an increase in product quality rating from 1 star to 2 stars might not affect a consumer's purchase decision. However, once the rating crosses the threshold, perhaps 3 stars, the rating will start to have influence on the consumer's purchase decision. The conversion rate may have a sudden, positive jump from around 3 stars to the 3.5 stars area, because consumers now think the rating is "good enough" for them to make purchases. Further, when the rating is above 4.5 stars, the marginal benefit of a rating increase diminishes because the rating is already in the consumers' comfort range. This stepwise phenomenon is illustrated in Figure 2, in which the observations are identical to Figure 1. Instead of a straight line, however, a stepwise line is employed for the estimation, which can help us capture a more accurate picture of the purchase decision process in reality. The single-regime model can be considered a special case of the regime-switching model.



**Figure 1** Single-regime model



**Figure 2** Regime-switching model

The focus of this study is to estimate the underlying regime switching. This is a nontrivial task, especially in the panel context. It is related to the following questions. First, what are the thresholds conceived by consumers? Second, in a given time period, which regime does a seller belong to? Is the seller in the high regime or the lower one? Third, what is the relationship between the conversion rate and the explanatory variables in each regime? We develop a hidden Markov model (HMM) to investigate these issues and estimate the regime-switching model. An HMM can be characterized by a combination of the following

three components: (i) the state-dependent outcome probability distribution  $F$ , (ii) the state-transition probability matrix  $G$ , and (iii) the initial distribution  $\pi$ .

To apply the HMM, we first need to define what hidden states exist in our research context. In our study, threshold values are not known. They are random variables. Therefore, given a set of specific values for state variables, which regime a seller belongs to is an unobserved random variable. This unobserved random variable, which represents the seller's unobserved regime location, is the seller's hidden state. In Figure 2, for instance, there are potentially two regimes: one is depicted by the upper right arm of the curve and the other is depicted by the lower left arm. Without knowing where the mid-transient line is located, we cannot conclude with certainty which arm a specific dot (observation) is associated with, that is, we do not know for sure the seller's hidden state. Once the transient line is given, we can consider the seller on the upper right side to be in the high business state and the seller on the lower left side to be in the low business state. The terms "low business state" and "high business state" come intuitively from the fact that the high state has a higher baseline conversion rate than the low state.

We want to mention that we do not model the mid-transient region as a state. Although our model can be flexible in generating more than two states, we hypothesize a two-state setting for simplicity. We assume the slope of the transient line is so steep that the mid-transient region is too narrow to have significant practical implications. We believe the most significant implication of the transient line is its position, not the inside-transient-region dynamics.

As the values of state variables change over time, a seller's regime location can change as well. Therefore, the seller's hidden state can also change over time. A seller may switch from the low state to the high state or vice versa, once the values of the state variables cross the corresponding threshold values. Due to the probabilistic nature of the threshold values, the state switching also operates in a probabilistic manner. The state-transition probabilities are loaded in the state-transition matrix  $G$  of the HMM. Since the switching is apparently caused by variation in the values of the state variables, the state-transition probabilities depend on the values of the state variables.

It is intriguing to postulate that threshold values themselves may be associated with the path taken by state variables. To return to the example of the product quality rating, if a rating of 2 starts in the low state, reaches 4.5 stars in the high state, and then starts to fall, the gain in conversion rate during the rating ascending process may already have been completely "consumed," that is, reverted back to the low state, when the rating drops from 4.5 stars to 3 stars. The economic underpinning of such a phenomenon comes from Prospect theory, which suggests that the individual perception of loss and gain depends not only on the absolute magnitude of the change in utility, but also the reference point of where she started from (Kahneman et al. 1979). This is illustrated in Figure 2 by the dashed line at the lower level of the curve. The higher-level solid line represents the path taken by the rating as it is ascending. The lower-level dashed line represents the path taken by the rating as it is descending. While an increase from 2 stars to 4.5 stars would be considered a fair improvement, falling from 4.5 stars to 3 stars may have already been counted as an "unacceptable" bad signal that damages the conversion rate greatly, that is, puts the seller back in the low state. In the HMM, by modeling the transition probabilities as state dependent, we allow our model to finely capture this potential effect.

After defining the hidden states and demonstrating how they can switch back and forth, we turn our attention to the state-dependent outcome probability distribution  $F$ , which characterizes the relationship between the conversion rate and the explanatory variables given a specific state. It is important to emphasize that in our study, state variables are a subset of explanatory variables. The criterion for an explanatory variable to be incorporated into a state variable set is that it has structural influence—that is, it can effect state switching—on the conversion rate. However, other than its influence on state switching, given a specific state, a change in the value of a state variable may still influence the conversion rate, as it is the nature of an explanatory variable. Product rating is a good example because it can serve as both a state variable and an explanatory variable. Price can serve as an explanatory variable, but not a state variable. Price can influence the conversion rate in a specific given state, but it has no structural impact on the conversion rate.

The estimation result of state-dependent outcome distribution  $F$  is able to reveal the answer for a long-time puzzle in the seller's mind: for some seller, tweaking the explanatory variables, for example reducing the price, effectively boosts her conversion rate, while the same operation done by another seller in the

similar context doesn't have the similar effect, or even quite far from. Some sellers may feel regardless the operations they take they can hardly effectively increase the conversion rate, while looking at other sellers enjoy the high conversion rate seemingly by doing "nothing". The explanation given by our theory is that those sellers are probably in different hidden states. Furthermore, by knowing which state she is most likely to be seated in at a given time period, the seller can learn from the transition probability matrix  $G$  that what are the state variables she should improve to make herself switch to a more desirable state.

## The Model

In this section, we explain the HMM setup. The data sampling time window starts from  $t = 1$  and ends at  $t = \bar{T}$ , which count as total  $\bar{T}$  periods. Sellers open their online retail stores all at time  $t = 1$ . A seller either terminates its service at  $t = T_i < \bar{T}$ , or remains open till  $t = \bar{T}$  in which case  $T_i = \bar{T}$ . Corresponding to our theory, we define two sets of covariates for seller  $i$ : (1) state variables, which determine seller  $i$ 's state transition at the end of  $t$ , and (2) outcome variables  $O_{it}$  which affect seller  $i$ 's current period conversion rate.  $R_{it}$  and  $O_{it}$  can be potentially overlapped, for example, the product rating. Seller  $i$ 's unobserved business state at time  $t$  is denoted by  $s_{it}$ . In our model,  $s_{it} = \{1, 2\}$  where 1 represents the "low business state" and 2 represents the "high business state." Figure 3 is an illustration of the HMM model in our study.

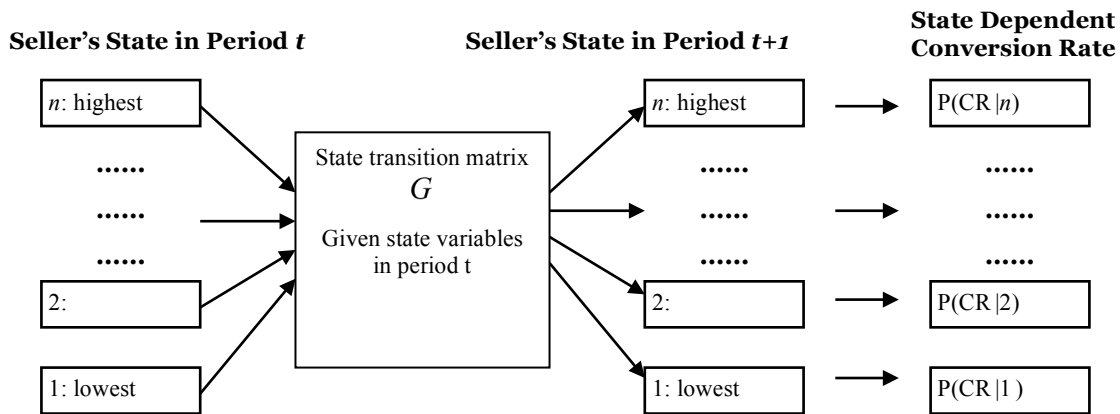


Figure 3 Hidden Markov Model of Sellers' Conversion Rate

### State-dependent Conversion Rate

In our model, constructing state-dependent outcome probability distribution  $F$  is such a task that concerns finding the most appropriate specification to describe randomness of the conversion rate. Various reports from the mass media suggest conversion rates among online retail stores often possess an extreme under-dispersion characteristic, in which most of the sellers' conversion rates are lower than 3%. Evidence from our dataset also supports this statement. These facts indicate that the conversion rate, though as a continuous measure, can hardly be assumed to follow either normal distribution or log-normal distribution. As it is arduous to find a proper continuous distribution for conversion rate, instead, we characterize it indirectly through the following discrete binomial distribution

$$\Pr(c_{it} | u_{it}, s_{it}, O_{it}) = \binom{u_{it}}{c_{it}} (p_{it})^{c_{it}} (1 - p_{it})^{u_{it} - c_{it}} .$$

$u_{it}$  is the total number of unique visitors for seller  $i$  in time period  $t$ .  $c_{it}$  is the total number of purchasers among  $u_{it}$ . Both  $u_{it}$  and  $c_{it}$  are observed in our dataset. The conversion rate, according to its definition, is

generated by  $c_{it} / u_{it}$ . Therefore, the expected value of conversion rate conditional on the observed  $u_{it}$ ,  $E(c_{it} / u_{it} | u_{it})$ , is equal to  $p_{it}$ . The probability  $p_{it}$  is modeled using logistic regression as follow.

$$\text{logit}(p_{it}) = \gamma_{s_{it}} O_{it} + v_i.$$

$\gamma_{s_{it}}$  is the state-dependent coefficient for  $O_{it}$  given seller  $i$ 's state  $s_{it}$  at the time  $t$ . Such state-dependent setting enables us to separately estimate different relationship between conversion rate and the outcome variables in different states.  $v_i$  is the random effect term capturing the unobserved individual seller's heterogeneity of the conversion rate at outcome level.

The adoption of binomial distribution results from a hypothetical construct for store visitors' decision making process. Define

$$U_{ijt} = \gamma_{s_{it}} O_{it} + v_i + \varepsilon_{ijt},$$

in which  $U_{ijt}$  represents visitor  $j$ 's net utility of purchasing a representative product in store  $i$  at time  $t$ . If we assume  $\varepsilon_{ijt}$  to be logistic distribution in  $(0,1)$  and the visitors make independent decision given they observe covariates  $O_{it}$  (for example, overall product rating given by the previous purchasers), the  $p_{it}$  will represent the probability that a single visitor make a purchase. Though this binomial distribution setting is parsimonious due to our aggregated construct of  $U_{ijt}$ , it provides more sensible logical explanation and more precise statistical approximation than a normal distribution assumption for  $p_{it}$ .

### State-transition Probabilities

For the two-state HMM model, the state transition matrix  $G$  is defined as

$$G = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

where  $q_{j,k} = \Pr(s_{i,t+1} = k | s_{i,t} = j), 1 \leq t \leq T_i$  and for each state  $j$ ,  $\sum_k q_{jk} = 1$ . We assume the state transition probability is governed by ordered logit model, which is a suitable choice for our context because (1) the high and low state have very clear economic interpretation that high state is more favorable than the low state in terms of conversion rate, (2) as stated in the theory development section, the threshold values are one of primary interests. Therefore,

$$q_{j,j+1} = 1 - \frac{\exp(\bar{\mu}_j - \beta_j R_{it} - \xi_i)}{1 + \exp(\bar{\mu}_j - \beta_j R_{it} - \xi_i)}, \quad q_{j,j-1} = \frac{\exp(\underline{\mu}_j - \beta_j R_{it} - \xi_i)}{1 + \exp(\underline{\mu}_j - \beta_j R_{it} - \xi_i)}, \quad q_{j,j} = 1 - q_{j,j-1} - q_{j,j+1}.$$

In the above function form,  $\bar{\mu}_j$  represents the threshold of transiting to next higher state given the current state  $j$ .  $\underline{\mu}_j$  represents the threshold of transiting to next lower state given the current state  $j$ . Applied to our low/high two states context,  $\bar{\mu}_1$  is the threshold from the low state to high state and  $\underline{\mu}_2$  is the one from the high state to low state.

$\beta_j$  denotes the state-dependent coefficients for state variables at given state  $j$ . As we explained in the theory development section, the prospect theory implies reversion of the path from low state to high state is not necessary the path from the high state to low state. The state-dependent setting enables us to estimate two paths separately. It also enables us to estimate different degrees of sensitivities for the same state variable given at different states.

$\xi_i$  is used to capture the seller's unobserved heterogeneity for state-transition. We assume two level's unobserved heterogeneity. The outcome level one  $v_i$  and the state level one  $\xi_i$  are potentially correlated.

To model the correlation, we assume they follow a bivariate normal distribution  $N(0, \Sigma)$  where the covariance structure is

$$\Sigma = \begin{bmatrix} \sigma_{\xi\xi} & \sigma_{\xi\nu} \\ \sigma_{\nu\xi} & \sigma_{\nu\nu} \end{bmatrix}.$$

It is worthwhile mentioning that the state transition happens at the end of a time period. Therefore the state variables at  $t$ ,  $R_{it}$ , together with the state  $s_{it}$  determine the state probability distribution at  $t+1$ . The state variables at  $t = T_i$  will not be used. This is because we sample the data and observe the values of state variables at the end of the time periods, instead of at the beginning.

The last component of the HMM is the initial state distribution  $\pi$ . Since in our data sample, all sellers are new born ones at the  $t = 1$ , we assume all of them start from the low state.

### Likelihood Function

In this subsection we build the overall likelihood function based on the component we proposed above. Denote  $S_i = \{s_{it} | 1 \leq t \leq T_i\}$  as the sequence of states for seller  $i$  during her entire lifespan,  $C_i = \{c_{it} | 1 \leq t \leq T_i\}$  as the sequence of numbers of purchasers for seller  $i$  during her entire lifespan,  $U_i = \{u_{it} | 1 \leq t \leq T_i\}$  as the sequence of number of visitors,  $O_i = \{O_{it} | 1 \leq t \leq T_i\}$  as the sequence of values of outcome variables,  $R_i = \{R_{it} | 1 \leq t \leq T_i\}$  as the sequence of values of state variables. First, we derive the likelihood for  $C_i$  as

$$\Pr(C_i | U_i, S_i, O_i) = \prod_{i=1}^{T_i} \Pr(c_{it} | u_{it}, s_{it}, O_{it}).$$

Then, the likelihood for  $S_i$  can be written as

$$\Pr(S_i | R_i) = \Pr(s_{i1}) \prod_{t=2}^{T_i} \Pr(s_{it} | s_{i,t-1}, R_{i,t-1}) = \pi_i \prod_{t=2}^{T_i} q_{s_{i,t-1}, s_{it}}.$$

Next, summing up all the possible paths of state evolution that seller  $i$  could take and taking the integral on the random effect, we derive the marginal distribution of  $C_i$  as

$$\Pr(C_i | U_i, O_i, R_i) = \int_{\xi} \int_{\nu} \left( \sum_{S_i} \Pr(C_i | U_i, S_i, O_i) \Pr(S_i | R_i) \right) \phi(0; \Sigma) d\nu d\xi.$$

Finally we obtain the overall likelihood function as

$$L(\beta, \gamma, \mu) = \prod_{i=1}^N \Pr(C_i | U_i, O_i, R_i)$$

where  $N$  is the total number of sellers in our dataset.  $\beta$  is the set of state variable coefficients at all states.  $\gamma$  is the set of outcome variable coefficients at all states.  $\mu$  is threshold set including both upper threshold  $\bar{\mu}$  and lower thresholds  $\underline{\mu}$  at all states.

### Data and Variable Description

The data were collected for a seven-month period from May 2011 to November 2011. All sellers opened their businesses in May 2011. Sellers could be divided into two mutually exclusive groups—ones that sell clothing and ones that sell prepaid refill cards for mobile phones. Some sellers exited in the middle of the seven-month period. The average lifespan of sellers is about 3.7 months. Table 1 provides variable descriptions and statistics.

*vstorap* is an important measure that indicates the overall attractiveness of the sellers' goods portfolio. It does not count the multiple visits to the same product page made by the same visitor. From a seller's perspective, high *vstorap* implies that the seller has formed a fine portfolio in which a number of products

capture consumers' interest. It doesn't count the multiple visits on the same product page made by the same visitor.

*goodsqualr* is a very common measure reflecting the degree of satisfaction consumers obtain from the product they purchased. It takes values from 0 to 5, in accord with the widely used 5-star rating scale system. In our study, the marketplace owner does not round the continuous average number to the nearest star or half star.

**Table 1 Variables and Descriptive Statistics**

State & outcome variables		Mean	StDev	Min	Max
<i>vstorap</i>	Average number of daily product page views per visitor	2.66	5.87	0.03	344
<i>goodsqualr</i>	Overall product rating score	1.32	2.18	0	5
<i>reputation</i>	Seller's reputation score	2.15	1.62	1	11
Outcome variables					
<i>imreplyspd</i>	Response ratio to visitors' inquiries	0.25	0.24	0	1
<i>mkttool</i>	Whether the seller uses marketing tools offered by platform owner	0.10	0.30	0	1
<i>numgoods</i>	Number of products	89.96	161.86	1	3,099
<i>avgprice</i>	Average product price of the seller	95.12	163.85	0.80	5,050
Control variables					
<i>numvstor</i>	Number of unique visitors	149.55	1163.06	1	42,780
<i>numbuyer</i>	Number of purchasers	5.16	39.10	0	1,592

*reputation* is also a very common and important score widely used in the online marketplace. It indicates the overall credibility and service quality of the seller. The consumer has the right to rate the seller's reputation after making a purchase. The score can be +1, 0, or -1, which represent good, fair, or bad, respectively. *reputation* is accumulated over time. A seller's accumulated total score is categorized into 11 levels in which level 0 is the lowest and level 11 is the highest. All stores start at level 0.

*imreplyratio* is the seller's response ratio to visitors' inquiries. In the online retail environment, consumers usually contact the seller for more details about the goods they are interested in. They often use the in-marketplace instant messaging tool to send inquiries and expect the seller to respond to them in a timely manner. *imreplyratio* is defined as the ratio of the total number of visitors who receive responses from the seller to the total number of visitors who inquire.

*mkttool* is a dummy variable indicating if the seller uses the marketing tools offered by market platform owner. Those marketing tools are basically designed for bringing more visitors to the seller. The platform owner charges the seller a considerable price for using those marketing tools.

*numgoods* reflects the seller's product variety. The sellers can adjust the number of good assortments listed in the store. They may delist some unpopular ones and enlist some popular ones according to the sales condition.

*vstorap*, *goodsqualr* and *reputation* are selected as both state variables  $R_{it}$  and outcome variables  $O_{it}$ . We believe those variables have both enduring and instantaneous impacts on the conversion rate. *imreplyspd*, *mkttool*, *numgoods* and *avgprice* are selected only as outcome variables  $O_{it}$  since we believe they do influence the conversion rate but do not incur regime change.

According to the correlation table shown in Table A1 in the appendix, multicollinearity is not a serious issue in our study.



## Estimation Results

Prior to estimating the parameters  $(\beta, \gamma, \mu)$ , we noticed in Table 1 that our variables exhibited long-tail properties. One possible reason is that our two groups of sellers, one selling clothes and the other selling prepaid refill phone cards, may have very different value ranges for the explanatory variables. We also suspect that the conversion rate dynamics may be so different across the two groups that solely adding fixed-effect intercepts may not be enough to take care of the heterogeneity. Therefore, we separate the clothing and prepaid card sellers into two groups and generate estimation results separately.

The literature endorses the identification of the general Markov regime switching model when the  $R_{it}$  and  $O_{it}$  variables overlap (Kim 2008). Therefore, as HMM is a special case of general Markov regime switching, any likelihood-based estimation method is suitable for our problem. Given that our likelihood function has no closed-form expression, we chose to estimate our problem using the Bayesian method via Monte Carlo Markov chain (MCMC) sampling. Since we have little prior knowledge about the parameter ranges, we adopt uninformative priors for all of our parameters. We draw multiple sets of initial values for the sampling process and run each chain for 40,000 iterations. The first 20,000 iterations, which are considered a “burn-in period,” are not used for parameter inferences. A representative sampling process is reported in Figure A1 in the Appendix to indicate the well convergence of our sampling process. It is essentially sampling process for prepaid refill cards’ state transition thresholds  $(\bar{\mu}_1, \underline{\mu}_2)$ .

The estimation results are listed in Table 2 for clothing sellers and Table 3 for prepaid refill cards.

**Table 2 Estimation Results (Product: clothing)**

		Low State		High State	
State variables		Mean	StDev	Mean	StDev
<i>vstorap</i>	Average num. of daily product page views per visitor	12.14 **	4.53	-9.94 **	5.80
<i>goodsqualr</i>	Overall product rating score	1.57 **	0.57	1.72	1.27
<i>reputation</i>	Seller’s reputation score	2.01 *	1.07	1.61	2.40
$\mu$	Thresholds between states (Low column is $\bar{\mu}_1$ ; high is $\underline{\mu}_2$ )	5.12 **	0.40	1.78 **	0.94
Outcome variables					
<i>imreplyspd</i>	Response ratio to visitors’ inquiries	-0.41 **	0.07	-5.06 **	0.22
<i>mkttool</i>	If using marketing tools offered by platform owner	0.14 **	0.03	-6.25 **	0.17
<i>numgoods</i>	Number of goods assortments	-4.21 **	0.14	4.45 **	0.30
<i>avgprice</i>	Average price of all goods	-7.10 **	0.60	-18.76 **	1.27
<i>vstorap</i>	Average num. of daily product page views per visitor	14.80 **	0.61	-0.65	0.54
<i>reputation</i>	Seller’s reputation score	0.89 **	0.08	-1.12 **	0.46
<i>goodsqualr</i>	Overall product rating score	2.40 **	0.07	3.42 **	0.37
<i>constant</i>	Constant term	-5.40 **	0.06	-1.849 **	0.29

\*  $p < 0.1$ ; \*\*  $p < 0.05$

**Table 3 Estimation Results (Product: prepaid refill cards)**

State variables		Low State		High State	
		Mean	StDev	Mean	StDev
<i>vstorap</i>	Average num. of daily product page views per visitor	4.78	4.65	2.50	9.16
<i>goodsqualr</i>	Overall product rating score	1.54 **	0.37	0.08	0.70
<i>reputation</i>	Seller's reputation score	3.57 **	0.64	3.13 **	1.42
$\mu$	Thresholds between states (Low column is $\bar{\mu}_1$ ; high is $\underline{\mu}_2$ )	3.31 **	0.19	0.97 **	0.48
Outcome variables					
<i>imreplspd</i>	Response ratio to visitors' inquiries	1.27 **	0.11	1.425 **	0.17
<i>mkttool</i>	If using marketing tools offered by platform owner	-0.07 *	0.04	1.531 **	0.05
<i>numgoods</i>	Number of goods assortments	-3.05 **	0.13	-2.175 **	0.14
<i>avgprice</i>	Average price of all goods	-19.03 **	1.49	-8.50 **	1.12
<i>vstorap</i>	Average num. of daily product page views per visitor	10.88 **	0.63	-0.52	0.42
<i>reputation</i>	Seller's reputation score	3.22 **	0.10	-11.81 **	0.18
<i>goodsqualr</i>	Overall product rating score	3.01 **	0.10	2.703 **	0.17
<i>constant</i>	Constant term	-3.79 **	0.06	1.567 **	0.11

\*  $p < 0.1$ ; \*\*  $p < 0.05$ 

Results from table 2 and table 3 show that most state variables and outcome variables are statistically significant. For the clothing sellers, it is not surprising that improving average number of daily product page views per visitor (putting more market popular goods into the seller's portfolio), overall product rating score (having higher product quality) and reputation score (promoting seller's credibility) will help seller switch from low state to high state. These facts are consistent with existing literature. For the prepaid refill card sellers, the *vstorap* doesn't have impact since the types of prepaid refill card available on the market are quite limited. Seller's goods portfolio are more homogenous.

However, compared to prepaid cards, it is interesting to note that when a clothing seller is in the high state, a high *vstorap* tends to switch the seller back to the low state, that is, to have a structural drawdown on the conversion rate. One potential explanation is the "too many choices" problem. After customers browse a number of products they are interested in but are not vertically differentiated, they may find it quite hard to make a purchase decision. Some of them may even end up making no purchase, because they could not figure out which product to buy. The prepaid refill card sellers do not have the same issue, since prepaid refill cards are often tied to the specific wireless carrier the consumer uses, which makes it much easier for the consumer to make a decision.

The constant terms in high state are consistently higher than those in the low states. This suggests that when the seller switches to the higher state, they will benefit the higher baseline conversion rate. Our estimation on  $\bar{\mu}_1$  provides implications for the seller that how far she is from transiting to a more favorable state. Notice that  $\bar{\mu}_1$  is not necessarily lower than  $\underline{\mu}_2$  because they are aggregated thresholds of all state variables' own ones.

For in-state relationship, when the seller is at low state, selling more market popular goods is a very effective way to boost the conversion rate. Marketing tools offered by market platform owner is also helpful. However, expanding the product variety at the low state is detrimental to conversion rate. This is

because when the seller is not sufficient credible, higher product variety is more only brings more visitors but not able to convince them to make the purchase. However when seller is in the high state, expanding the product variety becomes beneficial. This is because higher product variety will increase the likelihood of a consumer to find the product they want and high seller credibility will push them to make the purchase.

There also appears to be an interesting pitfall for sellers in both categories when they are in the high state: a high reputation score can actually depress the conversion rate. One possible reason is that when a seller's reputation score is very high, the seller's store will be ranked at the top of search results almost every time consumers input a query and ask to see the sellers with the best reputations first. This will bring a seller many casual visitors whose intrinsic purchasing propensities are generally lower than serious buyers'. However, if the seller is ranked at the upper middle of the first page of search results, the consumer who clicks that seller's link is more likely to be a serious buyer with a higher intrinsic purchase propensity.

## Conclusion

In this study, we constructed a hidden Markov model to study the conversion rate dynamics in online retail. We studied how the seller-level covariates affect the conversion rate dynamics and found that conversion rate dynamics are state dependent. We presumed two states for conversion rate dynamics—a low state and a high state—in which the high state represented a more favorable baseline conversion rate. The relationship between the conversion rate and the seller-level covariates differed across the two states. We estimated state-dependent relationships for both the high state and the low state and discussed the implications of the results. We also estimated the state-transition probability, which revealed how the seller-level covariates affected state transition. Our findings provide important guidance for online sellers of all statuses to improve their store conversion rates.

## References

- Moe, W. W., and Fader, P. S. 2004. "Dynamic Conversion Behavior at E-Commerce Sites," *Management Science* (50:3), pp. 326–335.
- Sismeiro, C., and Bucklin, R. E. 2004. "Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach," *Journal of Marketing Research* (41:3), pp. 306–323.
- Perdikaki, O., Kesavan, S., and Swaminathan, J. M. 2012. "Effect of Traffic on Sales and Conversion Rates of Retail Stores," *Manufacturing & Service Operations Management* (14:1), pp. 145–162.
- Mandel, N., and Johnson, E. J. 2002. "When web pages influence choice: Effects of visual primes on experts and novices," *Journal of Consumer Research* (29:2), pp. 235–245.
- Kahneman, D., and Tversky, A. 1979. "Prospect theory: An analysis of decision-making under risk," *Econometrica* (47:2), pp. 263–291.
- Kim, C. J., Piger, J., Startz, R. 2008. "Estimation of Markov regime-switching regression models with endogenous switching," *Journal of Econometrics* (143:2), pp. 263–273.

## Appendix

**Table A1 Correlation table**

	reputation	numgoods	avgprice	mkttool	goodsqualr	imreplspd	vstorap	numvstor	numbuyer
reputation	1								
numgoods	0.4163	1							
avgprice	-0.013	-0.0541	1						
mkttool	0.267	0.1335	0.043	1					
goodsqualr	0.5695	0.3263	-0.0663	0.1618	1				
imreplspd	0.2619	0.1131	-0.0152	0.097	0.2087	1			
vstorap	0.0415	0.1306	-0.0307	0.0204	0.0889	0.0069	1		
numvstor	0.3314	0.2461	0.0161	0.1769	0.1806	0.1277	-0.011	1	
numbuyer	0.3771	0.1994	-0.0132	0.1765	0.2097	0.122	0.0097	0.3418	1

**Figure A1 MCMC convergence graph**

