JETEM

Journal of Economic Theory and Econometrics

The Journal of the Korean Econometric Society

(계 량 경 제 학 보)

June 2008

Volume 19. Issue 2. pp.21-43

Learning and Herding: An Account of Hwang Woo-Suk Scandal[*]

In Ho Lee[†]**·Yeon Joon Lee**[‡]

Abstract A model of observational learning is constructed to explain the evolution of social opinion. In contrast to existing models, the model allows for uninformed agents who nevertheless make rational choices in view of their individual preferences. We apply the model to explain the time path of the accumulation of research funds in the case of the recent scandal involving a biomedical scientist, Hwang Woo-Suk. Empirical tests indicate that the evolution of social opinion represented by the research fund committed to the support of Hwang was induced by a low information arrival rate combined with private information of high precision when the agent is informed.

Keywords observational learning, Hwang Woo-Suk, simulated method of moments

JEL Classification C72, C73, D82

[*] We are grateful to helpful comments of John Geweke, Keunkwan Ryu, Junghoon Lee, and Kyoo-il Kim. Remaining errors are ours. The authors participate in 'The Second Brain Korea 21 Department of Economics, Seoul National University' Project in 2007. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2006-321-B00305).

[†] Seoul National University, ihl@snu.ac.kr

[‡] Deutsche Bank, aceofbass@hanmail.net

Received June 13, 2007, Revised August 20, 2007, Accepted June 26, 2008

1. Introduction

Hwang Woo-Suk[1] (born on January 29, 1953) was once a prominent South Korean biomedical scientist. He was a professor of theriogenology and biotechnology at Seoul National University who rose to fame after claiming a series of remarkable breakthroughs in the field of human stem cell research. Until November 2005, Hwang was considered to be one of the pioneering experts in stem cell research, best known for two articles he published in the journal "Science" in 2004 and 2005. The popularity of these articles was so widespread that the South Korean government was soon convinced that the entire Korean biomedical industry would benefit if the government provided him with a significant financial sponsorship.

However, Hwang's popularity took a dramatic plunge when it was discovered that he fraudulently reported his research findings. Both papers have been editorially retracted after being found to contain a large amount of fabricated data, while Hwang admitted to numerous charges of deceit and fraud. When the news reported that there were no human stem cells, the world was completely caught by surprise.

The present paper attempts to account for the evolution of the event from the view point of social learning. While his scientific achievement was examined through the editorial procedure at the journal "Science," - his fame, at least in Korean society, was largely spread through word-of-mouth. As he gained more fame, the Korean government provided greater research support for him. While there are other Korean scientists who published in the prestigious journals like "Science," the amount of financial support he was provided was much larger. We regard the research fund committed to his support representing the social opinion.

We construct a model of observational learning in which individuals make choices among two alternatives, riskless one and risky one. The risky choice[2] represents support for the authenticity of Hwang's research achievements while the riskless one repre-

[1]From Wikipedia, the free encyclopedia.
[2]The support for Hwang's research is risky in the sense that its authenticity is uncertain.

sents making a claim otherwise. Individuals may have private information surrounding authenticity; each individual is regarded as having an information about the authenticity of Hwang's research although the information may be just the observation of other individuals' claims. In case the individuals do not have private information, they are uninformed agents who nevertheless make choices among the alternatives. The probability that an individual has private information is an unknown parameter which we try to estimate using a statistical procedure. The precision of the private information is another parameter which we estimate.

In addition, an individual may get an extra payoff from having the research achievement proven to be authentic. We imagine that a family member of a patient who may benefit from stem cell research prefers supporting Hwang's research. The parameter capturing this feature of preferences in our model is a continuous random variable distributed on an interval.

Individuals make choices by sequentially observing the choices of predecessors. The debate on the Internet about authenticity is a real world example of the dynamic model in which individuals make choices over time.

Once we construct the analytic model, we solve for the equilibrium choice of individuals. We take the analytic result as the basis for the empirical test in the next stage. The estimation procedure used in this paper is a simulation estimator based on Pakes and Pollard (1989) and McFadden (1989). We use the data of the amount of research funding committed to Hwang's support by the Korean government. The government officers are willing to commit research funding as social opinions turn positive toward Hwang's research. In particular, we take the amount of research funds committed as representative of the public belief supporting the research.

The empirical test reveals that the event was best explained by the social conditions with a low probability of private information combined with a high precision of the information when it exists.

In the literature of herding since two seminal works of Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1992), numerous papers have shown that models in which there is sequential learning from other agents' action choices produces non-fully revealing informational cascades. These models show that the action choices of agents who make decisions after observing others' choices may depend entirely on those earlier choices i.e., later movers ignore entirely their own private information. The theoretical model which the present paper builds has the property that the learning never stops in contrast to the previous research. In particular adding individual cost to the model allows only full information aggregation in the sense that the economy always converges to the true state after a long sequence of action choices.

While there have been numerous theoretical attempts to extend the basic model of Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1992), there have been few attempts to test the analytic result using data. The present paper fills this gap by empirically testing the theoretical model using the data from the recent events.

The rest of the paper is organized as follows. Section 2 formalizes the intuition into a model which analyzes the theoretical construct. Section 3 provides an illustrative example. Section 4 investigates Hwang's scandal by applying the model using the simulation. Section 5 shows how robust the analysis is. Finally, Section 6 concludes.

2. The Model

The society consists of a scientist, N individuals $i = 1, \ldots, N$, and the government. The scientist works on a scientific project which may benefit society or may only serve to waste resources. The common state variable surrounding the authenticity of this research achievement, θ, is either 0 or 1, which we call L and H. L means the research is a failure and H denotes that the research is successful. N individuals have signals about the research and take actions after interpreting them. Each of the individuals chooses an action x_i from a binary action space, $\{0, 1\}$. Action 0 is riskless in that it guarantees the individual zero payoff. On the other hand, action 1 is risky in the sense that the utility from action 1 is $\theta - c$ where θ is an unknown state of the research common to all individuals, and c is the cost of choosing it, which is idiosyncratic among the individuals. The idiosyncratic cost is distributed independently and identically according to the uniform distribution on $[0, 1]$. In addition, there is a government which sponsors the research project. After observing each individual's action, the government chooses its level of cumulative sponsorship $M_i \in [\underline{M}, \overline{M}]$ where $\underline{M} = 0$.

Individuals make choices sequentially in increasing order, i.e., from $i = 1$ to N. Before they make a move, they receive at most two signals: first a noisy signal about the research, s_i, and second the precise information about c_i. However, some individuals may not receive a signal about the authenticity of the research, and are thus identified as uninformed agents. These individuals still get precise information about the idiosyncratic costs which they use in the decision making. The probability with which an agent gets the noisy signal about the authenticity of the research is denoted as λ. This is interpreted as the information arrival rate.

There are two signals about the research, $s_i \in \{G, B\}$, which are symmetric in the sense that

$$\Pr[G|H] = \Pr[B|L] = p, \quad p > \frac{1}{2}.$$

where the last inequality denotes informativeness of a signal. Initially, the society starts with a common prior, $\mu_0 = \Pr[\theta = 1]$, which is common knowledge among individuals.

Individual i solves the following optimization problem:

$$\max_{x_i} \mathbb{E}[(\theta - c_i)x_i | s_i, c_i, \mu_{i-1}] \tag{1}$$

where μ_{i-1} denotes the publicly available prior belief when individual i makes the decision. Note that the individual may have no signal s_i in which case we write $s_i = \emptyset$. The solution to the optimization problem (1) is: $x_i = 1$ if and only if $\mathbb{E}[\theta |\, s_i, \mu_{i-1}] \geq c_i$.

It is easy to see that

$$\mathbb{E}[\theta |\, s_i, \mu_{i-1}] = \pi_i(s_i) \cdot 1 + (1 - \pi_i(s_i)) \cdot 0 = \pi_i(s_i, \mu_{i-1}). \tag{2}$$

where $\pi_i(s_i, \mu_{i-1})$ is the posterior belief of the research's authenticity H conditional on the public prior and the noisy signal s_i:

$$\pi_i(s_i, \mu_{i-1}) = \frac{\mu_{i-1} \Pr[s_i|H]}{\mu_{i-1} \Pr[s_i|H] + (1-\mu_{i-1}) \Pr[s_i|L]}. \tag{3}$$

where $s_i \in \{G, B\}$. If individual i does not receive the signal about the research project, $s_i = \emptyset$, the posterior remains the same as the public prior μ_{i-1}.

We can summarize the findings so far as the individual chooses the risky action when the assessment of the research's authenticity is positive enough relative to the idiosyncratic cost. In particular, the uninformed agent relies heavily on the public belief in the decision making.

Next we analyze how to update the public beliefs after observing the choice of individual i. If individual i chooses the risky action, we know that $\mathbb{E}[\theta|s_i, \mu_{i-1}] = \pi_i(s_i, \mu_{i-1}) \geq c_i$. However, we cannot directly observe the signal of the individual even though we know that the signal is either G or B or $\emptyset$. There are 6 possibilities which induce the individual to take a risky action: i) the state is H, the individual gets signal G, and the cost is low enough; ii) the state is H, the individual gets signal B and the cost is low enough; iii) the state is L, the individual gets signal G, and the cost is low enough; iv) the state is L, the individual gets signal B, and the cost is low enough; v) the state is H, the individual gets signal $\emptyset$, and the cost is low enough; and vi) the state is L, the individual gets signal $\emptyset$, and the cost is low enough. For each possibility, the cost which is low enough for a risky action choice is determined based upon the beliefs that the individual demonstrates at the moment of decision making.

Define the set $C(s_i)$ as the set for which the idiosyncratic cost of individual i is less than the expected value of the authenticity of the research θ conditional on the observation of signal s_i:

$$C(s_i, \mu_{i-1}) = \{c_i | \pi_i(s_i, \mu_{i-1}) \geq c_i\}. \tag{4}$$

Also define $\gamma(s_i, \mu_{i-1})$ as the probability that the idiosyncratic cost of individual i is less than the expected value of the state θ conditional on s_i.[3]

$$\gamma(s_i, \mu_{i-1}) = \Pr[c_i \in C(s_i, \mu_{i-1})]. \quad (5)$$

Among the six possibilities, the first, the second and the fifth ones happen under the state H. Hence we can update the public belief conditional on observation of action 1 as follows using the definitions:

$$\mu_i(x_i = 1) = \frac{\mu_{i-1}\left\{\lambda[p\gamma(G) + (1-p)\gamma(B)] + (1-\lambda)\gamma(\emptyset)\right\}}{\lambda\left\{\mu_{i-1}[p\gamma(G) + (1-p)\gamma(B)] + (1-\mu_{i-1})[(1-p)\gamma(G) + p\gamma(B)]\right\} + (1-\lambda)\gamma(\emptyset)}. \quad (6)$$

where $\gamma(s_i)$ is a simplified notation of $\gamma(s_i, \mu_{i-1})$ for convenience. Similarly, we can update the public belief conditional on the observation of action 0 as follows:

$$\mu_i(x_i = 0) = \frac{\mu_{i-1}\left\{\lambda[p\gamma^c(G) + (1-p)\gamma^c(B)] + (1-\lambda)\gamma^c(\emptyset)\right\}}{\lambda\left\{\mu_{i-1}[p\gamma^c(G) + (1-p)\gamma^c(B)] + (1-\mu_{i-1})[(1-p)\gamma^c(G) + p\gamma^c(B)]\right\} + (1-\lambda)\gamma^c(\emptyset)}. \quad (7)$$

where $\gamma^c(s_i)$ denotes the probability that the idiosyncratic cost of individual i is higher than the expected value of the state θ: $\gamma^c(s_i) = 1 - \gamma(s_i)$[4].

After updating the public beliefs on the authenticity of Hwang's project, the government chooses its amount of funds. The cumulative governmental support after individual i's action, M_i, is defined as follows:

[3] To avoid the cluttering of notations we sometimes suppress the reference to the public prior belief when no ambiguity arises.

[4] Note that when $\theta = 1$, $\gamma(\emptyset) = \mu_{i-1}$, $\gamma(G) = \frac{\mu_{i-1}p}{\mu_{i-1}p + (1-\mu_{i-1})(1-p)}$, and $\gamma(B) = \frac{\mu_{i-1}(1-p)}{\mu_{i-1}(1-p) + (1-\mu_{i-1})p}$.

$$M_i = \mu_i \cdot \overline{M} \tag{8}$$

where $\overline{M}$ denotes the maximum of M_i, which means the maximum amount of governmental funds allocated to the scientific research. By definition, the government provides additional funds of the amount $(\mu_i - \mu_{i-1}) \cdot \overline{M}$ when $x_i = 1$. Similarly, the government withdraws a part of previously provided funds of the level $(\mu_{i-1} - \mu_i) \cdot \overline{M}$ when $x_i = 0$.

The behavioral rule of the government in our model can be regarded as the result of decision making procedure where the likelihood of the success of competing projects determines the amount of the fund support for each project. The outcome of scientific research depends on many probabilistic events. Moreover each project is so specialized that the funding committee at the government may not have precise knowledge to make correct judgement as to the likelihood of its success. The government may instead choose to rely on the "wisdom of crowds" since the agents who constitute the society may collectively have a better information about the likelihood of the success of projects.

The model of information aggregation employed in the present paper is a variation of the herd behavior model pioneered by Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) (hereafter referred as BHW). The standard result in this literature is the development of non fully-revealing informational cascades. The economy fails to aggregate the information dispersed in the economy as the consequence of sequential decision making. However the present model departs from this literature in that the economy eventually aggregates the information correctly in the sense that the true parameter about which the agents in the economy own collectively is discovered in the long run. The reason behind this surprising result is that the individual cost of each agent from the alternative choice induces the agents to make choices which depend on the private information they own.

The role of individual cost can be understood by the following simple exercise. Let us relax the assumption imposed on cost distribution. If we set a support of cost distribution as a subset of $[0, 1]$, then the uninformative information cascade happens with the positive probability since actions convey no information about private signals when μ_i is beyond the cost distribution support. When all individuals have the same cost, $c_i = \frac{1}{2}$, alike Banerjee (1992) or BHW (1992) model, herding prevents the aggregation of information of numerous individuals even though there exist uninformed rational players. Therefore, with idiosyncratic costs being introduced, our general model covers the phenomenon mentioned in BHW (1992). Furthermore, our basic assumption on the cost function presents the case where no herding arises which clearly distinguishes our model from the previous works.

Before we proceed into the estimation, we provide a short example to explain our model. The next table shows the evolution of the social consensus when common state H, $\mu_0 = 0.5$, $p = 0.75$ and $\lambda = 0.67$. After individual 1 gets signal G, she updates her posterior belief to 0.75 from 0.5. Since the cost 0.53 is low enough, she chooses $x_1 = 1$ and the public updates μ_1 to 0.58 from 0.5. On the other hand, individual 2 updates her posterior belief to 0.32 from 0.58 receiving signal B. Since her cost 0.85 is higher than her belief 0.32, she chooses $x_2 = 0$ and the public updates μ_2 to 0.48 from 0.58. Individual 6 is the uninformed agent. Her posterior belief remains the same as the prior belief 0.70 since she had no signal. Nevertheless, she has information on her idiosyncratic cost. Therefore she chooses $x_6 = 1$ because her cost is low enough.

i	s_i	c_i	π_i	x_i	μ_i
1	Good	0.53	0.75	1.0	0.58
2	Bad	0.85	0.32	0.0	0.48
3	Good	0.35	0.74	1.0	0.57
4	Good	0.59	0.80	1.0	0.64
5	Good	0.62	0.84	1.0	0.70
6	No	0.59	0.70	1.0	0.74
7	Bad	0.26	0.49	1.0	0.78

Table 1: (i, s_i, c_i, π_i, x_i, μ_i)

3. Estimation

Given the theoretical framework in the previous section and the data from the Hwang Woo-Suk scandal, we estimated λ and p using the empirical approach of simulated method of moments. A public prior belief μ_i is simulated using a Monte Carlo procedure and is used to form unbiased estimates of the computationally intractable predicted values from the model. These simulated μ_i values are then used to form 1^{st} and 2^{nd} moment conditions that are solved to find parameter values, λ and p.

3.1. Data

The data used in this study came from the judgment of the South Korean prosecution. On May 12, 2006, Hwang was indicted on charges of fraud, embezzlement, and breach of the bioethics laws of Korea, without physical detention. Prosecutors also brought fraud charges against the three stem cell researchers. Hwang embezzled 2.8 billion won ($3 million) out of some 40 billion won in research funds for personal purposes and the illegal purchase of ova used in his experiments.

According to the prosecution announcement, the Korean government supported Hwang through 6 channels, resulting in a total of 16.444 billion won ($17.6 million) in monetary support. Considering the fact that 6 different supports by the government were executed at different periods, we draw a graph showing relations between month i and M_i from Jan 2001 to May 2006.[5] Jan 2001 was the first time that the government provided 405 million won for Hwang's research, and May 2006 was the period when Hwang was indicted by the prosecution for his frauds. Only two days after the announcement of the prosecution, the government heralded that it would withdraw 15.18 billion won of the research funds that the government previously provided to Hwang. The government could withdraw a majority of its financial support because it planned to build research facilities which have not yet been constructed, while private sector funds were used for immediate research costs

[5] Month i is assumed as member i in the model which is to be relaxed in the section 'How Robust?'

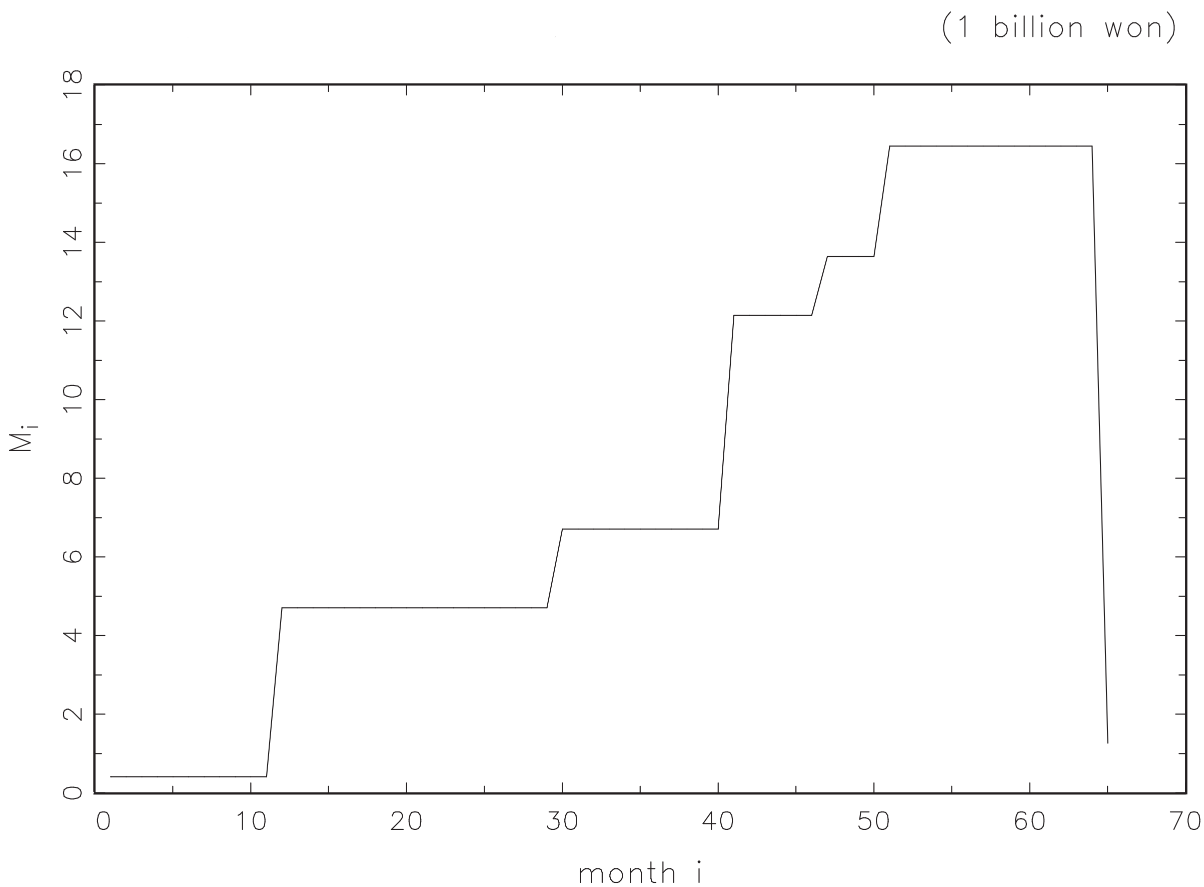

Figure 1: (month i, M_i)

including TA salaries. Based on the data, Figure 1 shows M_i through 65 months.

Because the government would not support Hwang once it regarded Hwang's research as unprofitable, the minimum of M_i, $\underline{M} = 0$. On the opposite side, the maximum of M_i, $\overline{M}$ equals the maximum support for 65 months within the government's budget. To figure out the maximum support from the government, we searched through the annual governmental budgets. We discovered that the maximum financial support provided by the South Korean government to a single scientific project was 6.936 billion won in 2004. Since the budget is planned annually, we extended the fund to 65 months to get 37.57 billion won. Therefore, $\overline{M}$ = 37.57 was based on the actual expenditure executed by the government. Hence, $M_i \in [0, 37.57]$. Furthermore, we know the true state $\theta = 0$ ex post.

3.2. Identification

Before we can proceed with estimation, we must show that the parameters λ and p can be identified. Since we have M_i and $\overline{M_i}$ = 37.57, we get μ_i by the equation (8) which was used as the data to estimate the model. Let μ_i^d denote μ_i derived from the data and σ_d, standard deviation of μ_i^d. Similarly, μ_i^s denotes simulated μ_i and σ_s, standard deviation of μ_i^s. Since μ_i^s and σ_s are the functions of λ and p, we use the following equations:

$$\sum_i \{\mu_i^s(\lambda, p) - \mu_i^d\}^2 = 0 \tag{9}$$

$$\{\sigma_s(\lambda, p) - \sigma_d\}^2 = 0 \tag{10}$$

A sufficient condition to the identification of two parameters is satisfied when we can find λ and p which solve equations (9) and (10).

3.3. Simulated Method of Moments

The estimation method used is the simulated method of moments (SMM) developed by Pakes and Pollard (1989) and McFadden (1989). Following this method, we estimated λ and p which explains Figure 1. Since λ ranges from 0 to 1, we partition $(0, 1)$ into 0.05-

sized subintervals to obtain 21 numbers from subintervals such as 0.01, 0.05, 0.1, 0.15, ..., 0.95, 0.99. Note that 0 is replaced by 0.01, and 1 is replaced by 0.99 because λ is strictly bigger than 0 and strictly smaller than 1. Similarly, in the perspective of p, we partition $(\frac{1}{2}, 1)$ to obtain 11 numbers such as 0.51, 0.55, 0.6, 0.65, ..., 0.95, 0.99. Thus, we have total 231 cells from 21 grids on the horizontal axis (λ) and 11 grids on the vertical axis (p).

Then we solve the following minimization problem of loss function $l_{\lambda,p}$:

$$\min_{\lambda,p} \; l_{\lambda,p} = \frac{l^1_{\lambda,p} - \nu^1_l}{\sigma^1_l} + \frac{l^2_{\lambda,p} - \nu^2_l}{\sigma^2_l} \tag{11}$$

where $l^1_{\lambda,p} = \sum_i (\mu^s_i - \mu^d_i)^2$, $l^2_{\lambda,p} = (\sigma_s - \sigma_d)^2$, ν^j_l, mean of $l^j_{\lambda,p}$, and σ^j_l, standard deviation of $l^j_{\lambda,p}$ for j = 1, 2. $l^1_{\lambda,p}$ is a measure of distance between the first moment of simulation and the first moment of the data. Simulating μ^s_i, we calculated $l^1_{\lambda,p}$ for each cell having 231 values. $l^2_{\lambda,p}$ is a measure of distance between the second moment of simulation and the second moment of the data moment. Similarly, with μ^s_i calculated previously, we obtained 231 values of $l^2_{\lambda,p}$. Next we use identity matrix as a weighting scheme. Since $l^1_{\lambda,p}$ and $l^2_{\lambda,p}$ have different scales, simply adding up two terms would not be an accurate measure of loss function. To provide equal explanatory power, we need to normalize $l^j_{\lambda,p}$ by subtracting the mean (ν^j_l) and dividing by the standard deviation (σ^j_l) for j = 1, 2.

3.4. Results

We made simulations to obtain a loss function for each cell, then repeated these 300,000 times to get an average of loss functions. Given μ^d_i, σ_d, λ and p, Table 2 shows $l_{\lambda,p}$ in each cell.

p \ λ	0.01	0.05	0.10	0.15	0.20	0.25	0.30
0.51	0.495766	0.495641	0.495486	0.495333	0.495176	0.495024	0.494870
0.55	0.495016	0.491914	0.488037	0.484133	0.480280	0.476376	0.472510
0.60	0.492599	0.479801	0.463886	0.448094	0.432121	0.416375	0.400964
0.65	0.488206	0.458182	0.420788	0.384253	0.347541	0.312247	0.276937
0.70	0.481256	0.423908	0.354138	0.285373	0.218600	0.151072	0.087318
0.75	0.470595	0.371305	0.252674	0.136835	0.027472	-0.084195	-0.185169
0.80	0.453856	0.291185	0.097739	-0.082577	-0.255185	-0.416767	-0.571493
0.85	0.425524	0.159769	-0.147130	-0.422867	-0.680449	-0.880345	-1.055806
0.90	0.370987	-0.084101	-0.574327	-0.961255	-1.223018	**-1.304462**	**-1.300043**
0.95	0.228773	-0.654529	**-1.368015**	**-1.563675**	**-1.364272**	-1.089560	-0.766701
0.99	-0.322902	**-2.026133**	**-1.837764**	-0.864152	-0.018262	0.521141	0.880296

($\lambda = 0.05$, p = 0.99) has the minimum value of the loss function and 7 numbers in bold font including ($\lambda = 0.05$, p = 0.99) indicate the minimum value of the loss function corresponding to the top 3$\%$. To test whether ($\lambda = 0.05$, p = 0.99) is statistically significant, we benchmarked ($\lambda = 0.01$, $p = 0.51$) since one can expect the short-run information cascade like the Hwang Woo-Suk scandal to happen because of lack of information and low accuracy of information. With the benchmark, we used the sign test. After simulation, if ($\lambda = 0.05$, p = 0.99) yielded a smaller loss function than ($\lambda = 0.01$, p = 0.51) did, we assigned (+), and (-) otherwise. The estimates and their standard errors are reported in Table 3.

Table 3 shows that ($\lambda = 0.05$, p = 0.99) is statistically significant, with a p-value of 0.098$\%$. Note that the numbers in parentheses are standard errors. Therefore, from Table 2 and 3, we conclude that the best estimate is ($\widehat{\lambda}$ = 0.05, $\widehat{p}$ = 0.99), which minimized the loss function among statistically significant values.

The empirical result shows that in the Hwang Woo-Suk scandal, there was extremely little information but the signal was exceptionally precise. In short, while a majority of members had no private information ($\widehat{\lambda}$ = 0.05), once they had private information, it was virtually perfect information ($\widehat{p}$ = 0.99). Our result is surprisingly consistent with the evolution of the Hwang Woo-Suk scandal. Until the truth was revealed, nearly all of the

p \ λ	0.35	0.40	0.45	0.50	0.55	0.60	0.65
0.51	0.494720	0.494557	0.494407	0.494254	0.494098	0.493941	0.493792
0.55	0.468689	0.464803	0.460980	0.457043	0.453275	0.449400	0.445620
0.60	0.385834	0.370179	0.355188	0.339480	0.325033	0.309282	0.294814
0.65	0.240249	0.206796	0.171906	0.138302	0.104098	0.071170	0.038564
0.70	0.022843	-0.040668	-0.096368	-0.157249	-0.221378	-0.276002	-0.333031
0.75	-0.286177	-0.383707	-0.477185	-0.572202	-0.656751	-0.732259	-0.809335
0.80	-0.716715	-0.836154	-0.943608	-1.006736	-1.077826	-1.143174	-1.145476
0.85	-1.148318	-1.206698	-1.222671	-1.209380	-1.169682	-1.116190	-0.289861
0.90	-1.237359	-1.132220	-0.954157	-0.828543	-0.709714	-0.553707	-0.440607
0.95	-0.146982	-0.219526	-0.004258	0.175160	0.331711	0.480880	0.592864
0.99	1.170299	1.353038	1.503376	1.637661	1.733185	1.843098	1.916260

Korean newspapers and mass media applauded Hwang's achievements without demonstrating any expertise in biomedical topics. On the other hand, the young but prospective scientists in the Biological Research Information Center (BRIC) continuously raised questions regarding the reliability of Hwang's papers from an expert's view.

4. How Robust?

In this section, we vary our assumptions to check the robustness of the model and the results. First, we vary the level of $\overline{M}$. Second, we change the time intervals of the data into equi-distances. Third, we estimate λ and p after interpolating the relation between month i and M_i. In every case, we prove its robustness respectively.

4.1. $\overline{M}$

In the previous section, we set $\overline{M}$ = 37.57 based on the observation of the actual expenditure executed by the government. However, we changed the assumption by letting $\overline{M} = 18$. This assumes that the highest μ_i from the data lies nearby $\mu_i = 1$ to yield ($\widehat{\lambda} = 0.05$, $\widehat{p} = 0.99$) through the same methodology used in the previous section. Hence,

p \ λ	0.70	0.75	0.80	0.85	0.90	0.95	0.99
0.51	0.493632	0.493474	0.493330	0.493183	0.493028	0.492868	0.492737
0.55	0.441811	0.437823	0.434118	0.430284	0.426296	0.422854	0.419902
0.60	0.278659	0.263595	0.250656	0.234947	0.219891	0.205960	0.194354
0.65	0.004263	-0.028620	-0.056635	-0.092853	-0.121830	-0.152517	-0.176253
0.70	-0.390127	-0.443630	-0.494106	-0.548731	-0.594517	-0.647359	-0.678460
0.75	-0.861631	-0.924814	-0.981973	-1.021345	-1.055909	-1.089658	-1.089476
0.80	-1.168019	-1.175539	-1.149402	-1.130503	-1.106818	-1.068302	-1.047472
0.85	-0.961704	-0.916957	-0.832098	-0.758563	-0.680913	-0.604588	-0.523756
0.90	-0.327352	-0.208517	-0.108762	0.008143	0.076413	0.165551	0.227098
0.95	0.732820	0.814870	0.923860	0.983065	1.083844	1.159591	1.208172
0.99	2.003575	2.063031	2.115856	2.179775	2.231630	2.283722	2.309822

Table 2: (λ, p, $l_{\lambda,p}$)

a variation on $\overline{M}$ yielded the same value of the estimates, thus demonstrating its robustness.[6]

4.2. Time Scale

In Figure 1, there are flat intervals such as the interval between month 1 and month 11. We interpreted the interval as the period when people are not able to move unilaterally because there is no decisive information. However, one may critique our interpretation because M_i should move upward or downward in each period in our model. Hence, we changed the previous time scale into time intervals with equal distances, i.e. only 7 months which reflect 7 observations. We gained ($\widehat{\lambda} = 0.20$, $\widehat{p} = 0.99$) which also suggests that our result is consistent even though we changed the time scale.

[6]The empirical results with this section including tables and graphs are available upon request from the authors.

$l_{0.05,0.99}$	$l_{0.01,0.51}$	sign
-2.0258608 (0.014540577)	0.49576600 (0.0000000843)	(+)
-2.0283629 (0.014514875)	0.49576592 (0.0000000847)	(+)
-2.0003367 (0.014471409)	0.49576596 (0.0000000844)	(+)
-2.0051341 (0.014471411)	0.49576602 (0.0000000843)	(+)
-2.0246640 (0.014560188)	0.49576591 (0.0000000845)	(+)
-2.0355598 (0.014501227)	0.49576602 (0.0000000843)	(+)
-2.0117498 (0.014390099)	0.49576593 (0.0000000845)	(+)
-2.0168784 (0.014431708)	0.49576576 (0.0000000847)	(+)
-2.0215708 (0.014566133)	0.49576592 (0.0000000846)	(+)
-2.0243600 (0.014447008)	0.49576609 (0.0000000846)	(+)

Table 3: ($l_{0.05,0.99}$, $l_{0.01,0.51}$, sign)

4.3. Interpolation

In Figure 1, the given data represents $(1, 0.405)$, $(12, 4.705)$, $(30, 6.705)$, $(41, 12.151)$, $(47, 13.651)$, $(51, 16.454)$, and $(65, 1.274)$. Between these points, we interpreted intervals as the time period in which no decisive information led people toward actions. However, in this subsection, we interpreted each interval as follows alternatively. For example, let us take a look at the interval between $(1, 0.405)$ and $(12, 4.705)$. In the interval, the government decided to support Hwang's research by committing 4.2 billion in month 12. We may interpret this as the government continuously gathering information about Hwang's research until it finally made a decision to support him rather than making a decision after observing only one public action in month 12. This interpretation enables us to interpolate missing points.

We chose the cubic spline method to interpolate the data which connects the two neighboring points with cubic polynomials. An explanation that an estimation with polynomial splines is superior to an estimation with exponential splines is well discussed in Shea (1985).

Let $(x_1, y_1) = (1, 0.405)$, $(x_2, y_2) = (12, 4.705)$, $(x_3, y_3) = (30, 6.705)$, $(x_4, y_4) = (41, 12.151)$, $(x_5, y_5) = (47, 13.651)$, $(x_6, y_6) = (51, 16.454)$, and $(x_7, y_7) = (65, 1.274)$. To estimate with cubic spline, we need to define first-order differentials at (x_1, y_1) and

(x_7, y_7). Since the overall graph slopes upward and it goes downward after reaching (x_6, y_6), we define two variables as follows; the first-order differential at (x_1, y_1) is defined as the slope between (x_1, y_1) and (x_6, y_6), which equals 3.2. In the same manner, the first-order differential at (x_7, y_7) is defined as the slope between (x_6, y_6) and (x_7, y_7), which equals -10.8. And we need

$$h_1 = 11, \quad h_2 = 18, \quad h_3 = 11, \quad h_4 = 6, \quad h_5 = 4, \quad h_6 = 14 \tag{12}$$

$$dy_1 = 0.391, \quad dy_2 = 0.111, \quad dy_3 = 0.495, \quad dy_4 = 0.25, \quad dy_5 = 0.701, \quad dy_6 = -1.084 \tag{13}$$

where $h_k = x_{k+1} - x_k$ and $dy_k = \frac{y_{k+1} - y_k}{h_k}$. Define a 7×4 matrix S where (i, j) element of S denotes a coefficient of $(j-1)$ th-order term of a cubic polynomial explaining i th interval. We get the following equation by cubic spline method.

$$\begin{bmatrix} 2h_1 & h_1 & 0 & \dots & 0 & 0 \\ h_1 & 2(h_1+h_2) & h_2 & \dots & 0 & 0 \\ 0 & h_2 & 2(h_2+h_3) & \dots & 0 & 0 \\ 0 & 0 & h_3 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & h_5 & 0 \\ 0 & 0 & 0 & \dots & 2(h_5+h_6) & h_6 \\ 0 & 0 & 0 & \dots & h_6 & 2h_6 \end{bmatrix} \begin{bmatrix} S_{(1,3)} \\ S_{(2,3)} \\ S_{(3,3)} \\ S_{(4,3)} \\ S_{(5,3)} \\ S_{(6,3)} \\ S_{(7,3)} \end{bmatrix} = \begin{bmatrix} 3(dy_1 - S_{(1,2)}) \\ 3(dy_2 - dy_1) \\ 3(dy_3 - dy_2) \\ 3(dy_4 - dy_3) \\ 3(dy_5 - dy_4) \\ 3(dy_6 - dy_5) \\ 3(S_{(7,2)} - dy_6) \end{bmatrix} \tag{14}$$

With equation (14) and these formulae

$$S_{(k,1)} = y_k, \; S_{(k,2)} = dy_k - \frac{h_k}{3}(S_{(k+1,3)} + 2S_{(k,3)}), \; S_{(k,4)} = \frac{S_{(k+1,3)} - S_{(k,3)}}{3h_k}, \tag{15}$$

we have S.

$$S = \begin{bmatrix} 0.40500000 & 0.31766667 & 0.02800000 & -0.00193939 \\ 4.70500000 & 0.29100000 & -0.03600000 & 0.00144444 \\ 6.70500000 & 0.39600000 & 0.04200000 & -0.00300000 \\ 12.1510000 & 0.22800000 & -0.05700000 & 0.01011111 \\ 13.6510000 & 0.63700000 & 0.12500000 & -0.02725000 \\ 16.4540000 & 0.33000000 & -0.20200000 & 0.00721428 \\ 1.27400000 & 0.00000000 & 0.10100000 & 0.00000000 \end{bmatrix} \tag{16}$$

Hence we have the following 6 polynomials. Each of them explains an interval. The polynomials are depicted graphically in Figure 2.

$$s_1(x) = -0.0019(x-1)^3 + 0.0280(x-1)^2 + 0.3177(x-1) + 0.4050 \qquad (1 \leq x < 12)$$
$$s_2(x) = 0.0014(x-12)^3 + -0.0360(x-12)^2 + 0.2910(x-12) + 4.7050 \quad (12 \leq x < 30)$$
$$s_3(x) = -0.0030(x-30)^3 + 0.0420(x-30)^2 + 0.3960(x-30) + 6.7050 \quad (30 \leq x < 41)$$
$$s_4(x) = 0.0101(x-41)^3 + -0.0570(x-41)^2 + 0.2280(x-41) + 12.151 \quad (41 \leq x < 47)$$
$$s_5(x) = -0.0273(x-47)^3 + 0.1250(x-47)^2 + 0.6370(x-47) + 13.651 \quad (47 \leq x < 51)$$
$$s_6(x) = 0.0072(x-51)^3 + -0.2020(x-51)^2 + 0.3300(x-51) + 16.454 \quad (51 \leq x \leq 65)$$

With Figure 2 and the above 6 polynomials, we estimated λ and p where we had ($\widehat{\lambda} = 0.05$, $\widehat{p} = 0.99$). This was also the same outcome achieved from the previous section.

5. Concluding Remarks

The rapidly changing Internet age help people to observe other people's decision makings and to express their opinions in open spaces in which collective learning occurs. In this paper, we have analyzed a model of observational learning and made an empirical approach using observations from Hwang Woo-Suk scandal. Our results show that the evolution of social opinion represented by the research fund committed to the support of Hwang was induced by a low information arrival rate combined with private information with high precision when the agent is informed.

Although theoretical aspects of observational learning are frequently analyzed, empirical attempts to explain social herding behavior using data are seldom made. We believe this empirical research will provide additional lights into the understanding of how social learning takes place.

Finally we make a few remarks about the extension of the present model. First we can allow the agents to choose the timing of decision making. It is conceivable that agents who have different information may choose their timing for decision making to take advantage of the information owned by other agents. The present model is too simple to allow for full account of the extension in this direction since there are informed agents with the same precision only and other agents without any information. However if we introduce many different information with differing precisions, the standard result where agents with higher precision moves first seems to follow without major difference in the result.

We can also recognize explicitly the fact that the success of the scientific research is not independent of the financial support but instead may depend on it. This will introduce endogeneity of the success probability and the financial support. While this direction suggests an interesting venue for further extension, it appears beyond the scope of the

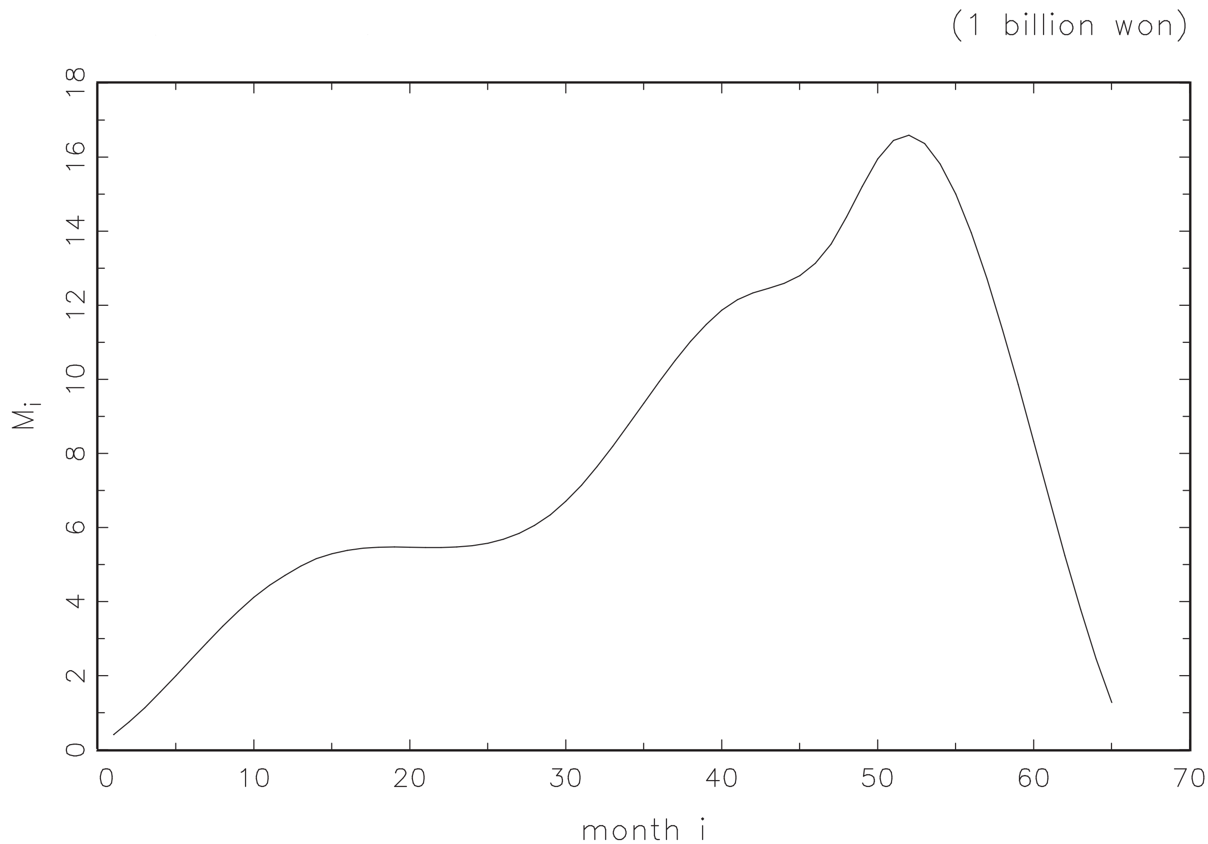

Figure 2: (month i, M_i using cubic spline method)

model since we need an explicit feedback mechanism from the financial support to the success probability. We leave this as a possible future project.

References

Banerjee, A. V., "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107, 797-817 (1992).

Bikhchandani, S., D. Hirshleifer, and I. Welch, "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100, 992-1026 (1992).

McFadden, D., "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57, 995-1026 (1989).

Pakes, A., and Pollard, D., "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057 (1989).

Shea, Gary S. (1985): "Interest Rate Term Structure Estimation with Exponential Splines: A Note," *The Journal of Finance*, 40, 319-325.

Smith, Lones, and Sϕrenson, Peter, "Pathological Outcomes of Observational Learning," *Econometrica*, 68, 371-398 (2000).