데이터의 시장가격에 대한 알고리즘적 이해

임요한[1]

2023년 6월 22일

[1] 도움주신분: 조민준,이슬기(서울대 통계학과), 이수형(서울대 국제대학원), DGP 연구회 등

발표 개요

1. Algorithmic Solution to Data Valuation by Argawal et al. 2019, ACM Conference on Econ. and Comp. (EC' 19) 에 대한 리뷰

2. 가상실험의 배경 문제 소개

3. 두 가지 가상실험

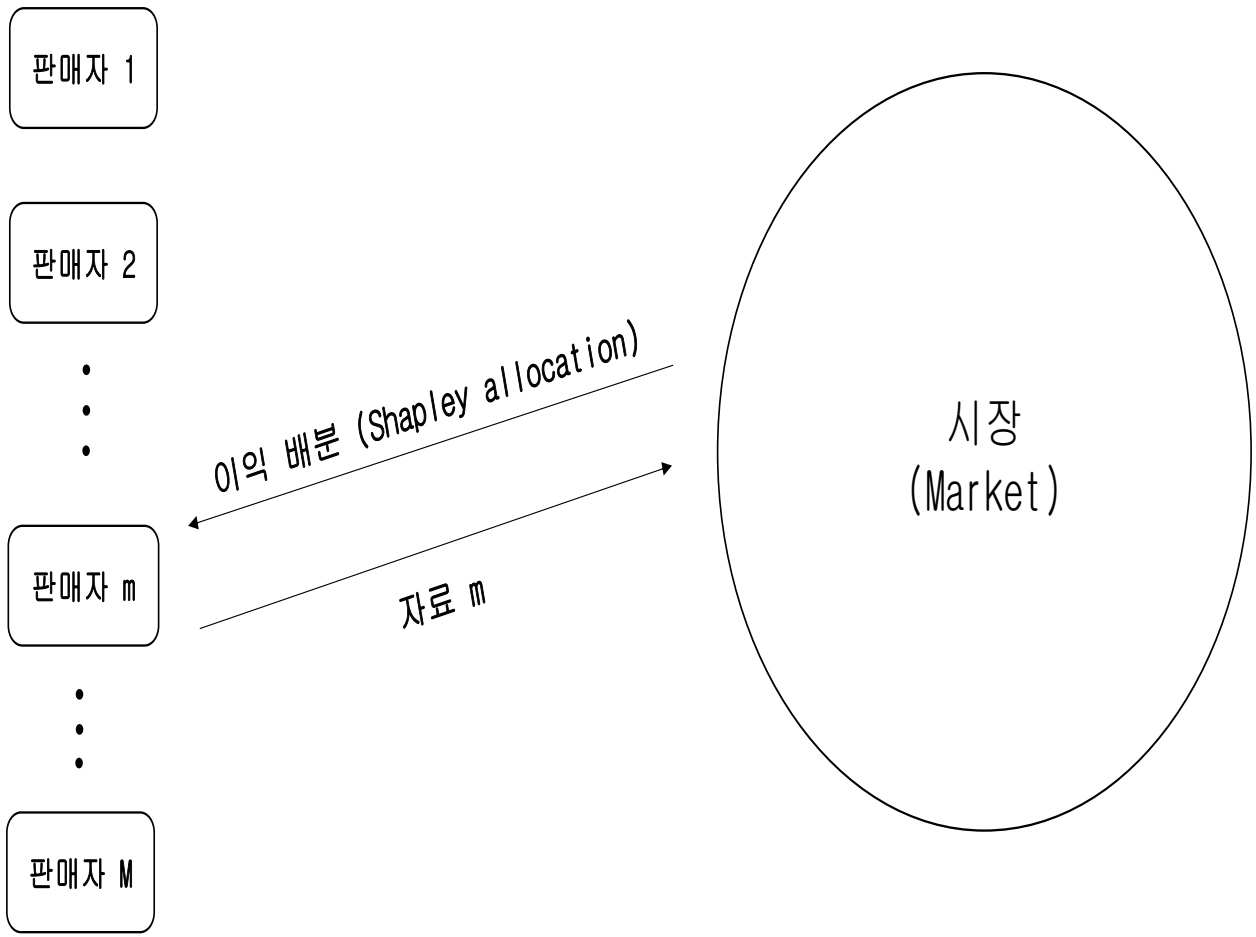
4. 논의

   일부 그림들을 위 논문으로 부터 특별한 인용 없이 가져왔음.

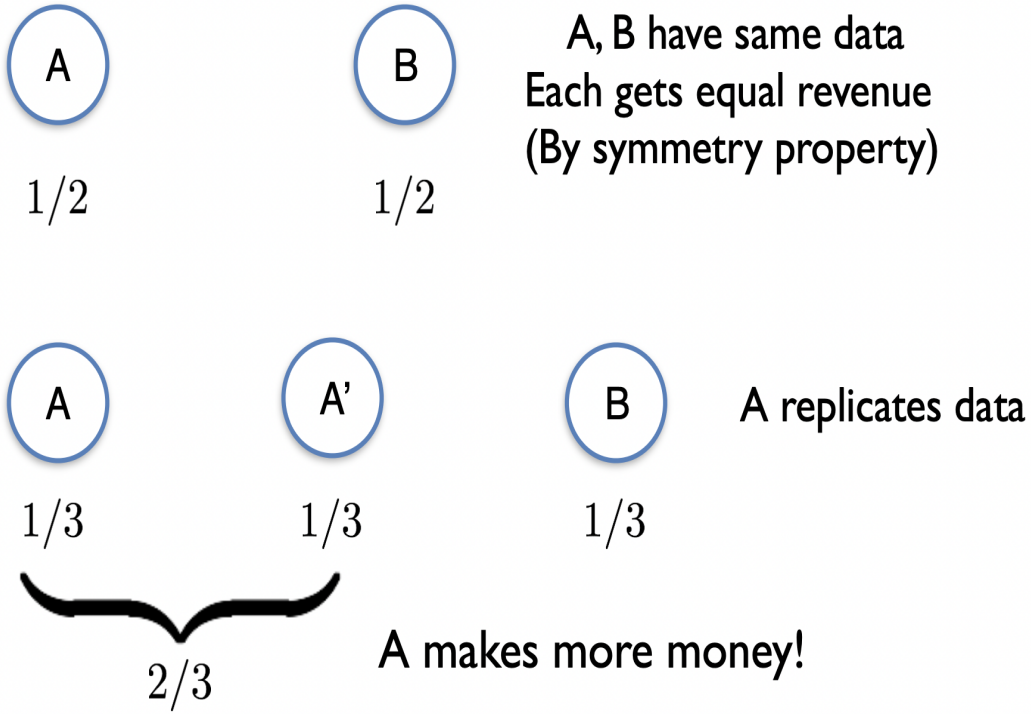
데이터 시장의 몇 가지 특성

- 복제의 용이성, Replication at zero cost
- 데이터-구매자 조합적(combinatorial) 특성
- 구매자의 데이터 사용목적이 복잡,다양함
- 데이터의 시의성
- ...

1. 시장 구조: 판매자와 시장

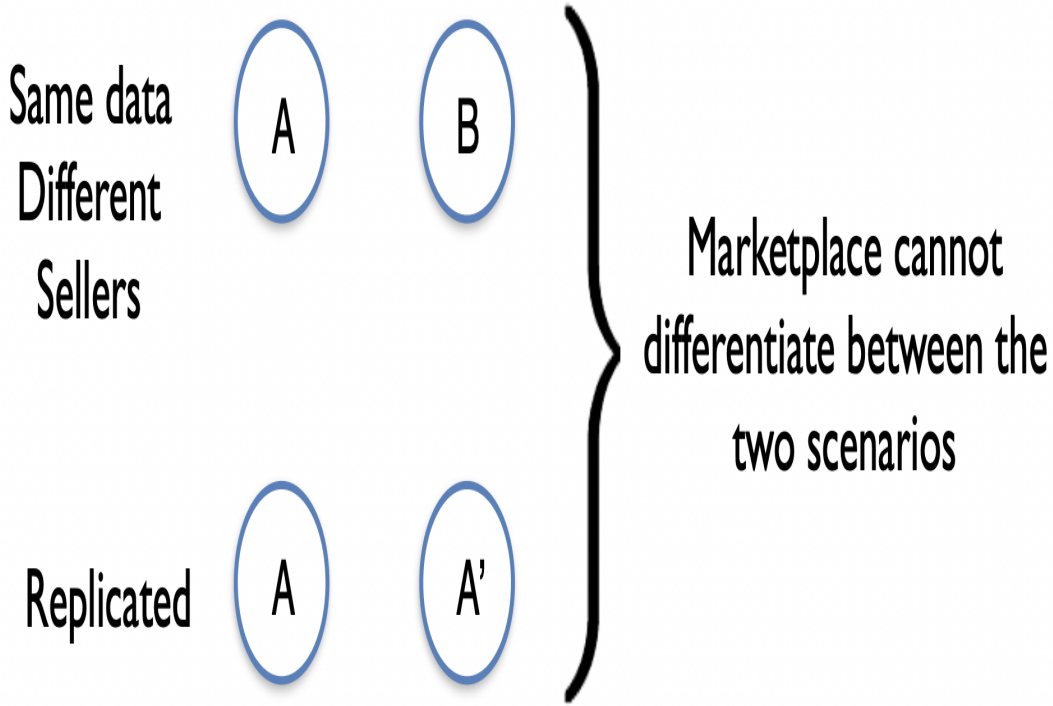


Shapley allocation



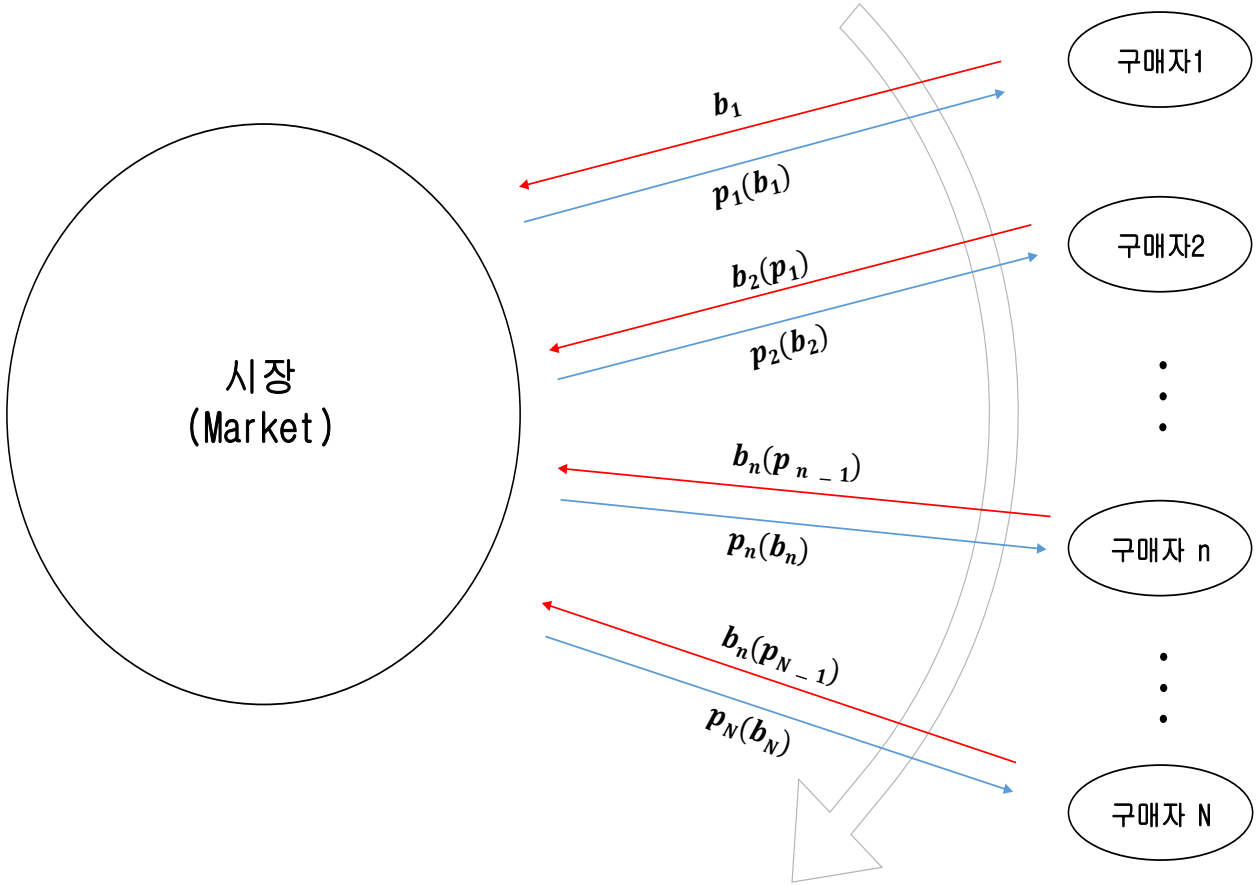
- 다른 자료들과의 pairwise 거리 기반, 유사자료에 대한 penalty 적용

Shapley allocation



- 반영 불가

1. 시장 구조: 시장과 구매자



1.2 시장의 구성, 용어

- X_n: 구매자 n이 기 보유한 자료
- Y_n: 구매자 n이 구입하고자 하는 자료
- b_n: 구매자 n의 (1 gain 당) bidding 가격
- p_n: (1 gain 당) 시장가격
- $\tilde{Y}_n(p_n, b_n)$: bidding가격이 b_n, 시장가격이 p_n일 때 구매자 n일 실제 얻게되는 자료. 편의상 종종 p_n과 b_n을 생략하고 적는다.
- $\mathcal{G}(Y_n, X_n)$: X_n을 가진 판매자가 추가로 Y_n을 얻게 되었을 때의 gain, 치역을 $[0, 1]$로 제한한다.
- μ_n: 구매자가 실제 얻는 1 gain에 대한 이익. 아래 실험에서 20으로 설정.

1.2 시장의 구성, 효용함수와 bidding 가격

- 효용함수

$$\mathcal{U}(b; p_n, Y_n, X_n) = \mu_n \mathcal{G}\left(\tilde{Y}_n(p_n, b), X_n\right) - \mathcal{RF}\left(p_n, b, \tilde{Y}_n(p_n, b), X_n\right),$$

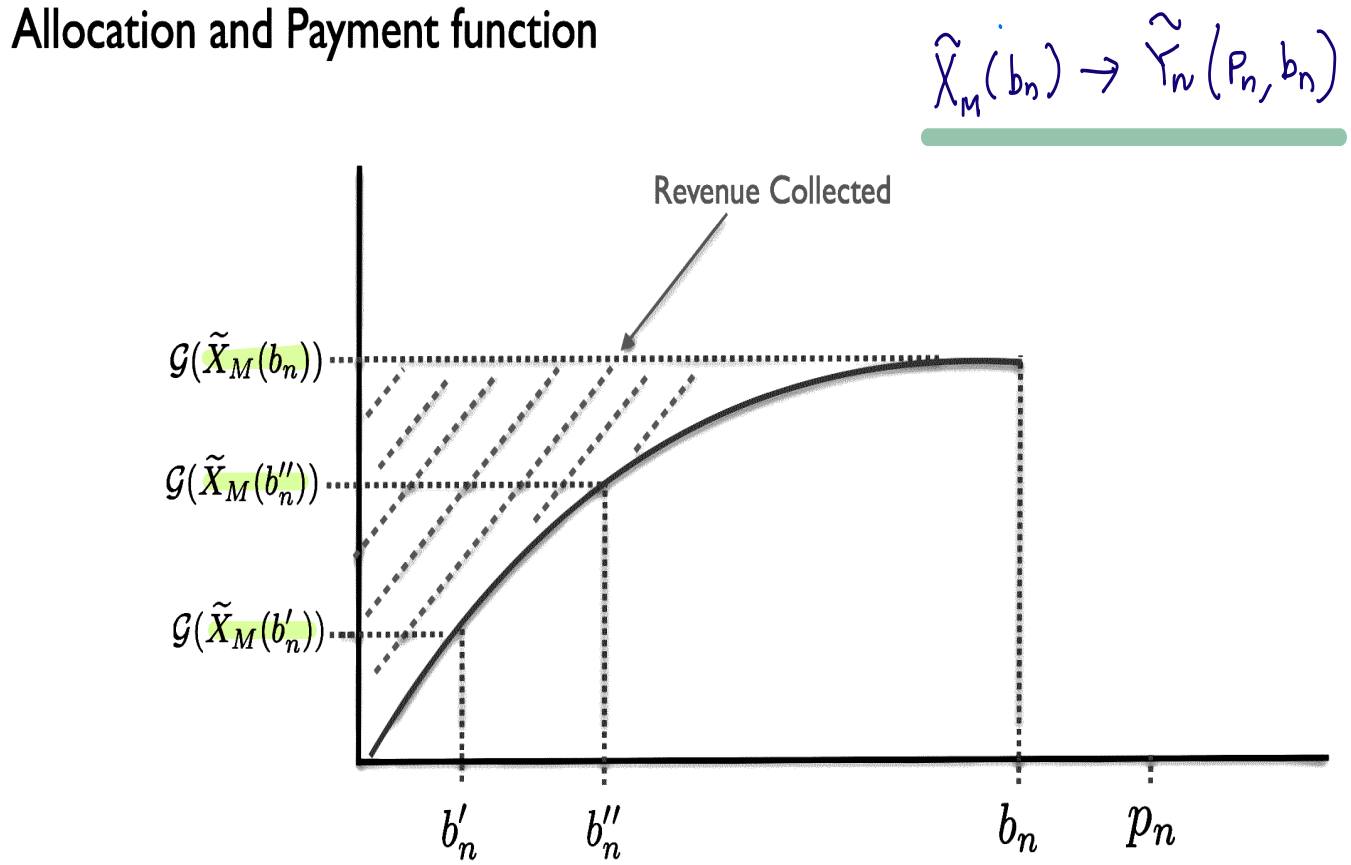
이고 여기서

$$\mathcal{RF}\left(p_n, b, \tilde{Y}_n(p_n, b), X_n\right) = b \cdot \mathcal{G}\left(\tilde{Y}_n(p_n, b), X_n\right) - \int_0^b \mathcal{G}\left(\tilde{Y}_n(p_n, t), X_n\right)\, dt$$

- Bidding 가격:

$$b_n = \text{argmax}_{z \in \mathbb{R}^+} \mathcal{U}\left(z; p_n, Y_n, X_n\right)$$

Revenue 함수, Payment 함수



"Pay only for additional marginal utility of each allocated feature"

1.2 시장의 구성, Revenue 함수와 시장가격

Price Update Algorithm

Algorithm 1 PRICE-UPDATE: $\mathcal{PF}^*(b_n, Y_n, \mathcal{B}, \epsilon, \delta)$

Let $\mathcal{B}_{\text{net}}(\epsilon)$ be an ϵ-net of $\mathcal{B}$
for $c^i \in \mathcal{B}_{\text{net}}(\epsilon)$ **do**
    Set $w_1^i = 1$ ▷ initialize weights of all experts to 1
end for
for $n = 1$ to N **do**
    $W_n = \sum_{i=1}^{|\mathcal{B}_{\text{net}}(\epsilon)|} w_n^i$
    Let $p_n = c^i$ with probability w_n^i / W_n ▷ note p_n is not a function of b_n
    for $c^i \in \mathcal{B}_{\text{net}}(\epsilon)$ **do**
        Let $g_n^i = \mathcal{RF}^*(c^i, b_n, Y_n)/\mathcal{B}_{\max}$ ▷ revenue gain if price c^i was used
        Set $w_{n+1}^i = w_n^i \cdot (1 + \delta g_n^i)$ ▷ Multiplicative Weights update step
    end for
end for
return p_n

1.2 시장의 구성, 시장가격의 이론적 근거

- Regret 함수:

$$\mathcal{R}(N, M) = \left[\sup_{(b_n, Y_n), n \in [N]} \left(\sup_{P^* \in \mathbb{R}_+} \sum_{n=1}^{N} \mathcal{RF}(P^*, b_n, Y_n) - \sum_{n=1}^{N} \mathcal{RF}(P_n, b_n, Y_n) \right) \right]$$

- Regret minimization:

$$\frac{1}{N} \mathcal{R}(N, M) \to 0.$$

2. 가상실험의 배경

- 선거에서 지지율을 예측을 목적으로, 서베이 자료를 거래.
  20-30대, 40-50대, 60대 이상의 세 그룹의 자료를 가정.
- 구매자가 사전에 $\mathbf{n} = (n_1, n_2, n_3)$ (X)를 가지고 시장이(또는 판매자가) 자료 $\mathbf{m} = (m_1, m_2, m_3)$ (Y)를 거래하고자 한다.
- 동일한 형태의 자료를 가진 구매자가 독립적으로 그리고 연속적으로 시장에 들어오는 상황을 가정하여 $\mathbf{n}$을 가진 구매자가 $\mathbf{m}$의 자료를 거래할 경우의 bidding가격과 시장가격을 살펴본다.

- Gain 함수:

$$G(\mathbf{n},\mathbf{m}) = \left\{2\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}\right\} - \left\{2\frac{1}{(n_1+m_1)}+\frac{1}{(n_2+m_2)}+\frac{1}{(n_3+m_3)}\right\}.$$

$$G(\mathbf{n},\mathbf{m}) \leftarrow G(\mathbf{n},\mathbf{m})\frac{1}{\left\{2\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}\right\}}$$

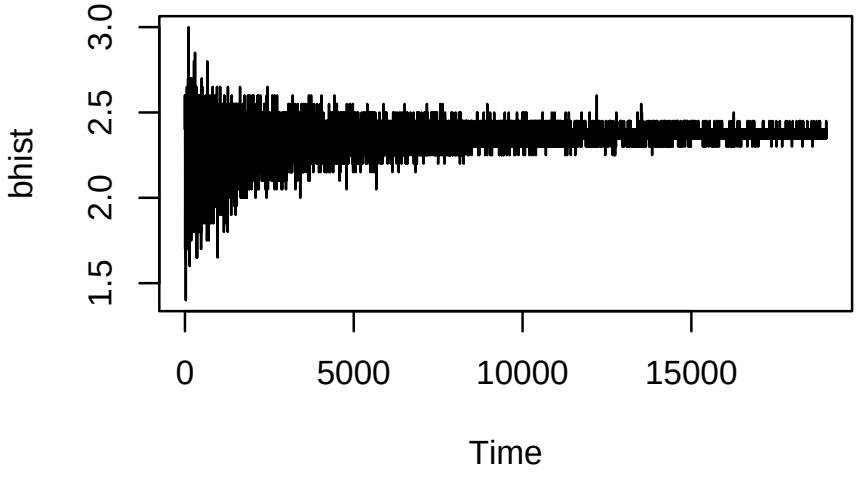
- p_n과 b_n의 차이에 대한 희석화(dilution) 또는 흐림(blurring):

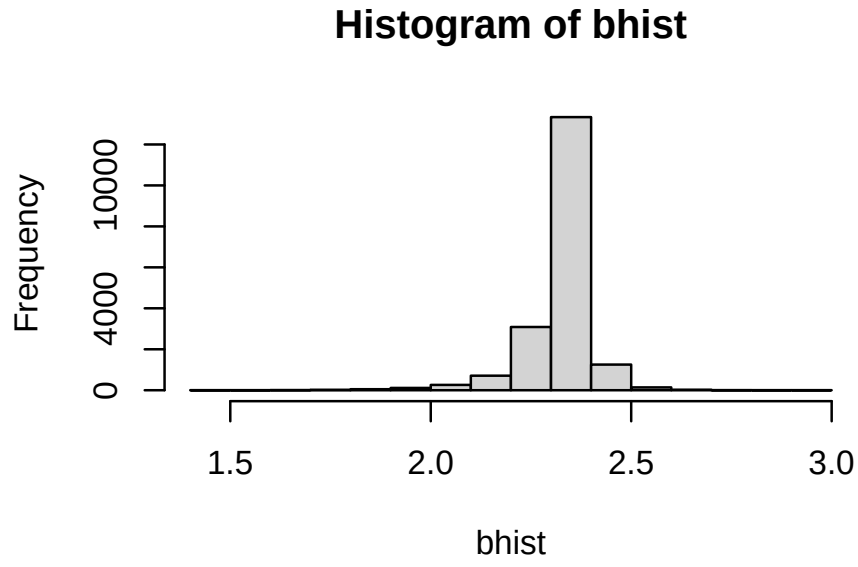
$$\mathbf{m} \rightarrow \frac{b_n}{p_n}\mathbf{m}.$$

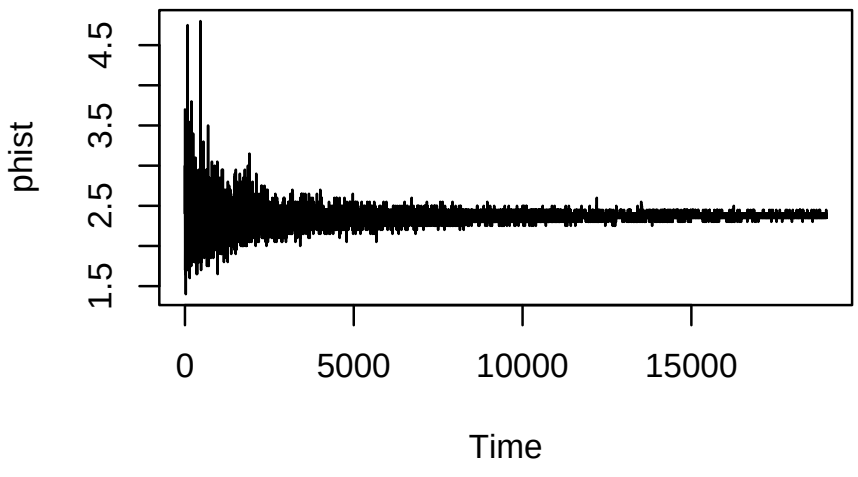
3. 가상실험들

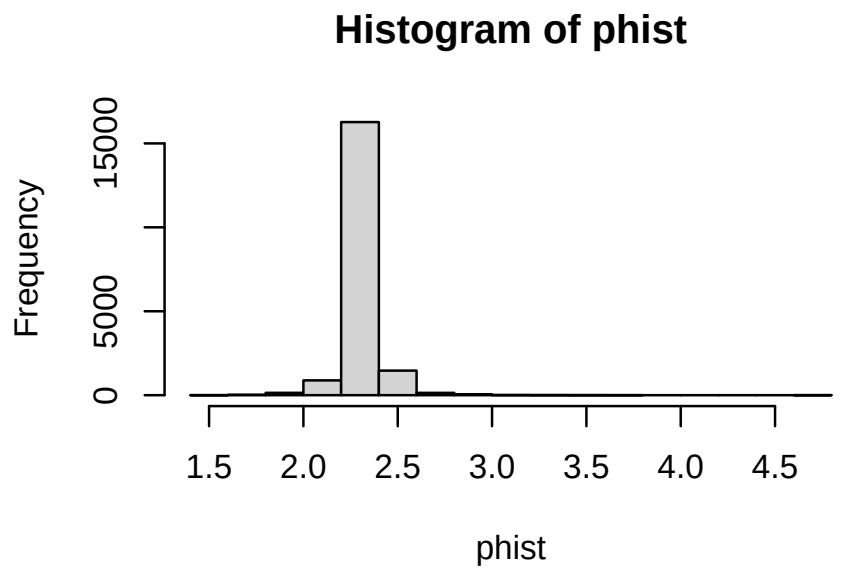
3.1 알고리즘 작동의 이해

$\mathbf{n} = (n_1, n_2, n_3) = (500, 1200, 1300)$을 가진 독립적인 구매자들이 크기 $\mathbf{m} = (700, 400, 400)$인 자료를 계속적으로 구매하는 시장.


bhist
1.5
2.0
2.5
3.0
0
5000
10000
15000
Time


Histogram of bhist
Frequency
0
4000
10000
1.5
2.0
2.5
3.0
bhist


phist
1.5
2.5
3.5
4.5
0
5000
10000
15000
Time


Histogram of phist
Frequency
0
5000
15000
1.5
2.0
2.5
3.0
3.5
4.0
4.5
phist

3. 가상실험의 결과

3.2 실험 1: m의 크기가 커짐에 따른 변화

앞과 동일한 실험에서 구매하고자 하는 자료의 크기를

- $\mathbf{m}_1 = K \cdot (700, 400, 400)$ (informative)
- $\mathbf{m}_2 = K \cdot (100, 700, 700)$ (less informative)

으로 하고 $K = 1, 3, 10$으로 변화 시키며 형성되는 가격의 변화를 살핌.

n	**m**	b_n	p_n
$(500, 1300, 1200)$	$1 \times (700, 400, 400)$	2.5110	2.5110
	$3 \times (700, 400, 400)$	4.3303*	4.3371*
	$10 \times (700, 400, 400)$	19.7675*	19.8114*
$(500, 1300, 1200)$	$1 \times (100, 700, 700)$	1.6512	1.6512
	$3 \times (100, 700, 700)$	1.9822	1.9822
	$10 \times (100, 700, 700)$	3.9274	3.9299

- "*" 는 아직 최종 수렴까지는 조금 더 반복하여야 함.
- 위의 bidding가격, 시장가격은 1 gain당 가격임

3. 가상실험의 결과

3.3 실험 2: n의 크기가 커짐에 따른 변화

n	**m**	b_n	p_n
$1 \times (500, 1300, 1200)$	$(700, 400, 400)$	2.5110	2.5110
$3 \times (500, 1300, 1200)$		1.9148	1.9148
$10 \times (500, 1300, 1200)$		1.8702	1.8702

3. 가상실험의 결과

3.4 실험 3: 거래자료가 이질적인 경우

- 거래되는 자료가 이질적인 경우: 시장에서 거래되는 자료가 $1-\alpha$의 확률로 $\mathbf{m}_1$ (informative), 그리고 α의 확률로 $\mathbf{m}_2$ (less informative)인 경우.
- 구매자는 시장에 대한 아무런 정보 없이 독립적으로 선택된 자료를 거래한다.

$\mathbf{n} = (500, 1300, 1200)$

$\mathbf{m}_1$	$\mathbf{m}_2$	α	b_n	p_n
$(700, 400, 400)$	$(100, 700, 700)$	0	2.5110	2.5110
		0.02	1.9333	1.9333
		0.05	1.9422	1.9422
		0.10	1.8501	1.8501
		0.20	1.9004	1.9004
		0.30	1.9004	1.9004
		0.50	1.9686	1.9686

시장가격에 $\alpha = 0$에서 discontinuity가 존재하는 것 처럼 보임.

4. 논의

- 고려한 모든 경우에서 $b_n = p_n$ 또는 $b_n \approx p_n$.
- 구매하고자 하는 자료의 정보(Y_n)가 보유하고 있는 자료의 정보(X_n)보다 월등한 경우 $p_n = \mu_n$ 관측.
- 구매하고자 하는 자료의 크기가 커짐에 따라 gain당 가격이 증가하는 현상 설명 못함.

- 상호경쟁적/협동적 시장을 모형화 하기가 힘들어 보임.

실험 1 with unnormalized gain function

n	**m**	b_n	p_n
$(500, 1300, 1200)$	$1 \times (700, 400, 400)$	4.69	4.69
	$3 \times (700, 400, 400)$	5.61	5.61
	$10 \times (700, 400, 400)$	7.49	7.49
$(500, 1300, 1200)$	$1 \times (100, 700, 700)$	4.35	4.35
	$3 \times (100, 700, 700)$	4.34	4.34
	$10 \times (100, 700, 700)$	5.35	5.35