Comparing Theories of One-Shot Play Out of Treatment

Philipp Külpmann — University of Vienna
Christoph Kuzmics — University of Graz

July 3, Keio University

Motivation

Use of game theory in social science research:

- observe behavior that one would like to explain
- identify key individuals in the interaction
- identify what they could do
- identify their goals
- this constitutes a game (players, strategies, payoffs)
- identify appropriate "solution concept" (a theory): set of predictions, a mapping from games to predicted behavior

the explanation can only be "good" if the solution concept predicts well

Testing predictive power

we want to test the predictive power of such theories with lab experiments

many theories have parameters

how should we choose these parameters?

we don't estimate them with our own data!

why? we are worried about overfitting and getting game specific subject pool specific parameter estimates.

recall: when using such theories we probably use them for new games and new "subjects"

so we estimate parameters "out of treatment"

Advantages of "out of treatment" testing

conceptually appropriate for our motivation

allows a direct likelihood comparison of theories

without having to "punish" theories for the number of parameters

What we do specifically, theories

test theories of one-shot play

Nash equilibrium (NE)
level k reasoning (LK) - Stahl and Wilson (1994, 1995), Nagel (1995)
cognitive hierarchy (CH) - Camerer, Ho, and Chong (2004)
quantal response equilibrium (QRE) - McKelvey and Palfrey (1995)
noisy introspection (NI) - Goeree and Holt (2004)
quantal level k (QLK) - Stahl and Wilson (1994)
quantal cognitive hierachy (QCH) - Camerer, Nunnari, and Palfrey (2016)

all theories are used with and without risk aversion ("-RA" added to their abbreviation)

What we do specifically, parameter estimates

all parameter estimates from meta-analysis of Wright and Leyton-Brown (2017)

risk aversion coefficient (CRRA) from Hey and Orme (1994) and Harrison and Rutström (2009)

What we do specifically, games

we look at representative selection of 2 by 2 games with unique and mixed strategy predictions

additional advantage: we do not need to have a model of mistakes:

no pure strategy prediction and mixed strategy observation

can directly apply a Vuong (1989) test (based on log-likelihood comparison)

experimental design, games

	U	D
U	0, 0	x, 1
D	1, x	y, y

(a) Hawk-Dove Game

	U	D
U	z, 0	0, 1
D	0, 1	1, 0

(b) Matching Pennies

Figure: Payoff matrices for our hawk dove and matching pennies games

with $(x, y) \in \{(1,0),(2,0),(3,0),(5,0),(10,0),(3,2),(5,2),(10,2),(10,3),(10,5)\}$

first five T1-T5: anti-coordination games; second five T6-T10: "proper" hawk-dove games

and $z \in \{1,2,3,5,10\}$ each once as player 1 (T11-T15) and 2 (T16-T20)

experimental design, subjects

conducted at the DR@W Laboratory at the University of Warwick using zTree (Fischbacher, 2007)

147 subjects recruited using Warwick's SONA System without placing any restrictions on the subject pool

experimental design, play and pay

each subject was asked to play each game once (20 games)

subjects were for each round randomly matched with some other subject in the subject pool

subjects never received any feedback about their opponent or their opponent's strategy choice until the very end when all they were told is how much money they received

at the very end of the experiment, one of the 10 rounds of hawk-dove and one of the 10 rounds of matching pennies was randomly selected and paid out in *GBP*

after the games were played, we also elicited risk aversion and level k reasoning skills (in the 11/20 game developed by Arad and Rubinstein, 2012) which were not used in this paper

predictions

theory i makes prediction $p_{i,t}$ for treatment t, where $p_{i,t}$ is the proportion of action U

theory i is thus identified by the vector $p_i = (p_{i,1}, \ldots, p_{i,20})$ of predictions

p is the true probability vector of choices of our subjects

$\bar{p}_t$ is the observed proportion of U in treatment t in our sample

Vuong test

log-likelihood ratio between any two theories i and j is given by

$$\log LR(\bar{p}, p_i, p_j) = \sum_t \left[n\bar{p}_t \log\left(p_{i,t}/p_{j,t}\right) + n(1-\bar{p}_t)\log\left((1-p_{i,t})/(1-p_{j,t})\right)\right]$$

"true" variance of this log-likelihood is then given by

$$\sum_t np_t(1-p_t)\left[\log\left(p_{i,t}/p_{j,t}\right) - \log\left((1-p_{i,t})/(1-p_{j,t})\right)\right]^2,$$

estimated by replacing p_t by its maximum likelihood estimator $\bar{p}_t$ for each treatment t

Vuong statistic (or z-score) given by the log-likelihood divided by the square root of its estimated variance

under the true model, the Vuong statistic is asymptotically standard normally distributed

results of Vuong test, hawk-dove

	NE	NE-RA	LK	CH	CH-RA	QRE	QRE-RA	QLK	QLK-RA	QCH	QCH-RA	NI	NI-RA	RND
NE	0	5.86	4.31	7.98	6.36	7.73	5.19	9.3	2.72	5.76	6.27	6.47	4.67	3.49
NE-RA	-5.86	0	-3.71	-1.21	0.47	-1	-2.26	-1.57	-5.76	-2.05	1.54	-1.43	-3.17	-4.98
LK	-4.31	3.71	0	3.05	5.61	4.35	6.8	1.76	-1.4	6.89	5.63	6.13	5.03	-6.49
CH	-7.98	1.21	-3.05	0	1.77	0.92	-0.25	-1.04	-2.72	-0.82	1.91	0.1	-1.4	-3.77
CH-RA	-6.36	-0.47	-5.61	-1.77	0	-1.63	-3.99	-2.12	-5.56	-3.36	1.34	-2.4	-5.07	-7.11
QRE	-7.73	1	-4.35	-0.92	1.63	0	-0.86	-2.42	-3.01	-2.34	1.8	-1.27	-2.3	-5.05
QRE-RA	-5.19	2.26	-6.8	0.25	3.99	0.86	0	-0.33	-2.92	-0.81	4.54	0.59	-8.93	-11.2
QLK	-9.3	1.57	-1.76	1.04	2.12	2.42	0.33	0	-2.37	0.07	2.21	0.97	-0.62	-2.57
QLK-RA	-2.72	5.76	1.4	2.72	5.56	3.01	2.92	2.37	0	2.5	5.69	2.84	2.17	0.74
QCH	-5.76	2.05	-6.89	0.82	3.36	2.34	0.81	-0.07	-2.5	0	3.46	4.32	-2.11	-7.62
QCH-RA	-6.27	-1.54	-5.63	-1.91	-1.34	-1.8	-4.54	-2.21	-5.69	-3.46	0	-2.55	-5.41	-7.23
NI	-6.47	1.43	-6.13	-0.1	2.4	1.27	-0.59	-0.97	-2.84	-4.32	2.55	0	-2.81	-6.81
NI-RA	-4.67	3.17	-5.03	1.4	5.07	2.3	8.93	0.62	-2.17	2.11	5.41	2.81	0	-12.38
RND	-3.49	4.98	6.49	3.77	7.11	5.05	11.2	2.57	-0.74	7.62	7.23	6.81	12.38	0

results of Vuong test, MP asymmetric

	NE	NE-RA	LK	CH	CH-RA	QRE	QRE-RA	QLK	QLK-RA	QCH	QCH-RA	NI	NI-RA	RND
NE	0	0	13.8	10.12	10.12	8.79	12.95	9.67	9.28	12.76	9.32	11.75	14.05	0
NE-RA	0	0	13.8	10.12	10.12	8.79	12.95	9.67	9.28	12.76	9.32	11.75	14.05	0
LK	-13.8	-13.8	0	9.28	9.28	7.54	11.26	8.47	8.46	9.46	8.53	9.39	9.11	-13.8
CH	-10.12	-10.12	-9.28	0	0	0.55	-5.95	0.61	5.39	-6.67	5.2	-5.06	-8.02	-10.12
CH-RA	-10.12	-10.12	-9.28	0	0	0.55	-5.95	0.61	5.39	-6.67	5.2	-5.06	-8.02	-10.12
QRE	-8.79	-8.79	-7.54	-0.55	-0.55	0	-5.67	-0.23	1.09	-6.65	2.33	-6.05	-7.15	-8.79
QRE-RA	-12.95	-12.95	-11.26	5.95	5.95	5.67	0	6.79	5.99	-10.05	6.61	1.89	-11.54	-12.95
QLK	-9.67	-9.67	-8.47	-0.61	-0.61	0.23	-6.79	0	1.31	-7.69	2.77	-7.19	-8.13	-9.67
QLK-RA	-9.28	-9.28	-8.46	-5.39	-5.39	-1.09	-5.99	-1.31	0	-6.52	3.19	-5.41	-7.51	-9.28
QCH	-12.76	-12.76	-9.46	6.67	6.67	6.65	10.05	7.69	6.52	0	7.04	8.97	-9.58	-12.76
QCH-RA	-9.32	-9.32	-8.53	-5.2	-5.2	-2.33	-6.61	-2.77	-3.19	-7.04	0	-6.17	-7.81	-9.32
NI	-11.75	-11.75	-9.39	5.06	5.06	6.05	-1.89	7.19	5.41	-8.97	6.17	0	-9.29	-11.75
NI-RA	-14.05	-14.05	-9.11	8.02	8.02	7.15	11.54	8.13	7.51	9.58	7.81	9.29	0	-14.05
RND	0	0	13.8	10.12	10.12	8.79	12.95	9.67	9.28	12.76	9.32	11.75	14.05	0

results of Vuong test, MP symmetric

	NE	NE-RA	LK	CH	CH-RA	QRE	QRE-RA	QLK	QLK-RA	QCH	QCH-RA	NI	NI-RA	RND
NE	0	2.48	-0.37	-0.26	-0.26	-0.3	-0.19	-0.8	2.93	-1.07	0.56	-1.06	-1.05	-1.1
NE-RA	-2.48	0	-5.01	-4.86	-4.86	-5.11	-4.94	-5.61	1.02	-5.9	-3.68	-5.88	-5.87	-5.93
LK	0.37	5.01	0	8.53	8.53	1.35	3.6	-7.99	4.54	-9.36	7.69	-9.32	-9.32	-9.45
CH	0.26	4.86	-8.53	0	0	-0.59	1.52	-8.16	4.42	-9.25	7.57	-9.21	-9.21	-9.32
CH-RA	0.26	4.86	-8.53	0	0	-0.59	1.52	-8.16	4.42	-9.25	7.57	-9.21	-9.21	-9.32
QRE	0.3	5.11	-1.35	0.59	0.59	0	7.16	-8.15	4.37	-8.61	6.4	-8.6	-8.56	-8.63
QRE-RA	0.19	4.94	-3.6	-1.52	-1.52	-7.16	0	-8.7	4.26	-8.95	6.27	-8.95	-8.92	-8.97
QLK	0.8	5.61	7.99	8.16	8.16	8.15	8.7	0	4.96	-9.49	7.94	-9.51	-9.41	-9.48
QLK-RA	-2.93	-1.02	-4.54	-4.42	-4.42	-4.37	-4.26	-4.96	0	-5.27	-3.52	-5.26	-5.25	-5.3
QCH	1.07	5.9	9.36	9.25	9.25	8.61	8.95	9.49	5.27	0	8.41	8.69	10.27	-9.41
QCH-RA	-0.56	3.68	-7.69	-7.57	-7.57	-6.4	-6.27	-7.94	3.52	-8.41	0	-8.39	-8.38	-8.45
NI	1.06	5.88	9.32	9.21	9.21	8.6	8.95	9.51	5.26	-8.69	8.39	0	8.79	-9.21
NI-RA	1.05	5.87	9.32	9.21	9.21	8.56	8.92	9.41	5.25	-10.27	8.38	-8.79	0	-9.87
RND	1.1	5.93	9.45	9.32	9.32	8.63	8.97	9.48	5.3	9.41	8.45	9.21	9.87	0

summary of results, general

one theory is significantly better (or worse) than another if the z-score for their comparison is less than -2 (greater than $+2$)

one theory is weakly better (or worse) than another when the z-score for their comparison is negative (positive)

all theories, except Nash equilibrium without risk aversion in hawk dove games and for matching pennies games for the asymmetric player position, are on the whole significantly better than random guessing

all theories have some predictive power (based on simple χ^2-test, all p-values < 0.000001)

no universally best theory

Nash equilibrium with risk aversion is pretty good

summary of results, hawk-dove

For the ten hawk-dove treatments (T1-T10),

1. the overall best theory without considering risk aversion is QRE, which is significantly better than NE, LK, QLK, and QCH (and weakly better than CH and NI);
2. Nash equilibrium with risk aversion (NE-RA) is weakly better than QRE and weakly worse only compared to CH-RA and QCH-RA.

summary of results, matching pennies asymmetric

For the five matching pennies treatments for the asymmetric own payoff player position (T11-T15),

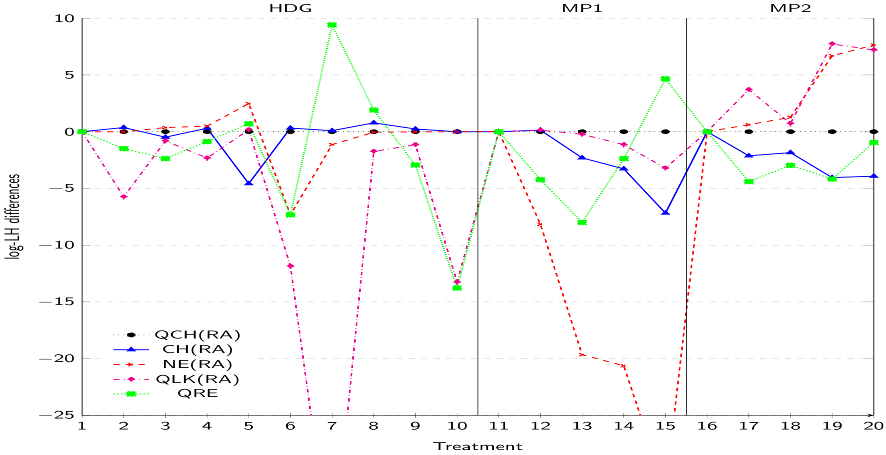
1. the two overall best (and essentially equally good) theories without considering risk aversion are QRE and QLK, which are significantly better than NE, LK, QCH, NI (and weakly better than CH);
2. the best theory overall is QCH-RA, which is significantly better than all other theories.

summary of results, matching pennies asymmetric

For the five matching pennies treatments for the symmetric own payoff player position (T16-T20),

1. the best theory without considering risk aversion is NE, which is, however not significantly better than any other theory without risk aversion;
2. Nash equilibrium with risk aversion (NE-RA) is significantly better than all other theories, except QLK-RA, which is weakly better than NE-RA.

treatment by treatment comparison



Conclusion

“out of treatment” testing methodology

for one-shot games:

no universally best theory

Nash equilibrium with risk aversion is among best theories in two out of three treatment groups

only bad in asymmetric player position in matching pennies games

Specific Results

findings agree fairly well with those of Wright and Leyton-Brown (2017): "winning" theories in their meta-analysis: cognitive hierarchy model (CH) and quantal level k model (QLK)

only here with risk aversion!

QRE implicitly incorporates risk aversion

biggest problem for Nash equilibrium with risk aversion is the asymmetric own payoff player position in the matching pennies treatments

Omitted theories

some theories make pure strategy predictions: maximin play, ambiguity aversion according to Eichberger and Kelsey (2011) (that they used to explain the data of Goeree and Holt (2001)), "level-1 with risk aversion" of Fudenberg and Liang (2019)

some theories make predictions identical to those of other theories: Nash equilibrium with a fraction of fairness-minded individuals of Fehr and Schmidt (1999) (with calibrations taken from Fehr and Schmidt (2004)), also Bolton and Ockenfels (2000)

Predictions, hawk-dove

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Data	0.55	0.63	0.69	0.69	0.84	0.34	0.58	0.65	0.56	0.41
x	1.00	2.00	3.00	5.00	10.00	3.00	5.00	10.00	10.00	10.00
y	0.00	0.00	0.00	0.00	0.00	2.00	2.00	2.00	3.00	5.00
NE	0.50	0.67	0.75	0.83	0.91	0.50	0.75	0.89	0.88	0.83
NE-RA	0.50	0.57	0.61	0.66	0.73	0.20	0.39	0.57	0.52	0.40
LK	0.50	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
CH	0.50	0.66	0.66	0.66	0.66	0.50	0.66	0.66	0.66	0.66
CH-RA	0.50	0.59	0.60	0.66	0.66	0.34	0.40	0.59	0.59	0.40
QRE	0.50	0.54	0.57	0.62	0.71	0.50	0.57	0.68	0.66	0.62
QRE-RA	0.50	0.53	0.54	0.57	0.60	0.43	0.47	0.52	0.51	0.47
QLK	0.50	0.55	0.59	0.67	0.79	0.50	0.59	0.76	0.74	0.67
QLK-RA	0.50	0.76	0.78	0.78	0.70	0.17	0.21	0.76	0.63	0.22
QCH	0.50	0.52	0.53	0.56	0.62	0.50	0.53	0.60	0.59	0.56
QCH-RA	0.50	0.57	0.61	0.65	0.70	0.31	0.40	0.57	0.52	0.41
NI	0.50	0.52	0.54	0.58	0.66	0.50	0.54	0.63	0.61	0.58
NI-RA	0.50	0.52	0.53	0.54	0.57	0.47	0.48	0.51	0.50	0.49
RND	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Predictions, matching pennies

	T10	T12	T13	T14	T15
Data	0.63	0.67	0.76	0.76	0.84
z	1.00	2.00	3.00	5.00	10.00
NE	0.50	0.50	0.50	0.50	0.50
NE-RA	0.50	0.50	0.50	0.50	0.50
LK	0.50	0.53	0.53	0.53	0.53
CH	0.50	0.66	0.66	0.66	0.66
CH-RA	0.50	0.66	0.66	0.66	0.66
QRE	0.50	0.55	0.59	0.67	0.82
QRE-RA	0.50	0.53	0.55	0.59	0.64
QLK	0.50	0.55	0.60	0.67	0.78
QLK-RA	0.50	0.68	0.69	0.69	0.69
QCH	0.50	0.52	0.53	0.56	0.63
QCH-RA	0.50	0.64	0.70	0.72	0.73
NI	0.50	0.52	0.54	0.58	0.68
NI-RA	0.50	0.52	0.53	0.55	0.58
RND	0.50	0.50	0.50	0.50	0.50

	T16	T17	T18	T19	T20
Data	0.52	0.33	0.36	0.27	0.27
z	1.00	2.00	3.00	5.00	10.00
NE	0.50	0.33	0.25	0.17	0.09
NE-RA	0.50	0.43	0.39	0.34	0.27
LK	0.50	0.47	0.47	0.47	0.47
CH	0.50	0.47	0.47	0.47	0.47
CH-RA	0.50	0.47	0.47	0.47	0.47
QRE	0.50	0.49	0.48	0.47	0.44
QRE-RA	0.50	0.49	0.48	0.46	0.44
QLK	0.50	0.50	0.49	0.49	0.48
QLK-RA	0.50	0.33	0.31	0.31	0.31
QCH	0.50	0.50	0.50	0.50	0.50
QCH-RA	0.50	0.44	0.43	0.43	0.43
NI	0.50	0.50	0.50	0.50	0.50
NI-RA	0.50	0.50	0.50	0.50	0.50
RND	0.50	0.50	0.50	0.50	0.50