Model Selection Accuracy in Behavioral Game Theory: A Simulation

Paul J. Healy   Hyoeun Park

OSU

25 Years of QRE

A Warning:

A Warning:

I have no clue what I'm talking about.

Why I Don't Trust MLE in Behavioral Game Theory

2-person guessing game (Costa-Gomes and Crawford, 2006):
Pick $s_i \in S_i = [\underline{s}_i, \overline{s}_i]$
Target: $p_i s_j$
Payoff: $-|s_i - p_i s_j|$

Why I Don't Trust MLE in Behavioral Game Theory

2-person guessing game (Costa-Gomes and Crawford, 2006):
Pick $s_i \in S_i = [\underline{s}_i, \overline{s}_i]$
Target: $p_i s_j$
Payoff: $-|s_i - p_i s_j|$

Level-k Model: σ_i^0 is $U[\underline{s}_i, \overline{s}_i]$
$s_i^1 = BR_i(\sigma_j^0)$, $\quad s_i^k = BR_i(s_i^{k-1}) \;\; k = 2, 3, \ldots \qquad s_i^N = BR_i(s_j^N)$

Why I Don't Trust MLE in Behavioral Game Theory

2-person guessing game (Costa-Gomes and Crawford, 2006):
Pick $s_i \in S_i = [\underline{s}_i, \overline{s}_i]$
Target: $p_i s_j$
Payoff: $-|s_i - p_i s_j|$

Level-k Model: σ_i^0 is $U[\underline{s}_i, \overline{s}_i]$
$s_i^1 = BR_i(\sigma_j^0)$, $\quad s_i^k = BR_i(s_i^{k-1}) \;\; k = 2, 3, \ldots \quad s_i^N = BR_i(s_j^N)$
"Spike-Logit" Error: $\sigma_i^k = (1-\epsilon)\mathbb{1}_{\{s_i^k\}} + \epsilon LR(s_j^{k-1}|\lambda)$

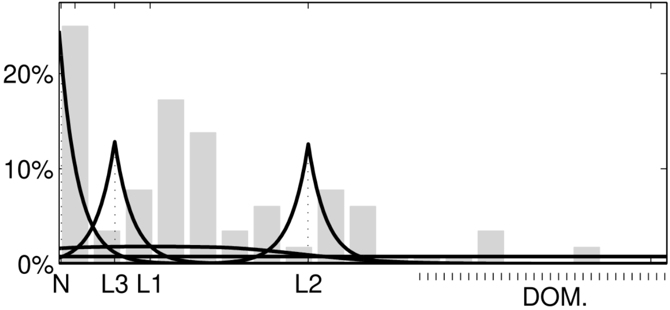
Why I Don't Trust MLE in Behavioral Game Theory

Level-k Model: σ_i^0 is $U[\underline{s}_i, \overline{s}_i]$

$s_i^1 = BR_i(\sigma_j^0)$, $\quad s_i^k = BR_i(s_i^{k-1}) \;\; k = 2, 3, \ldots \quad s_i^N = BR_i(s_j^N)$

"Spike-Logit" Error: $\sigma_i^k = (1-\epsilon)\mathbb{1}_{\{s_i^k\}} + \epsilon LR(s_j^{k-1}|\lambda)$

$LR(s_j^{k-1}|\lambda)$ for $\lambda = 1$:



Distrust of Model Selection

- Error structure may drive estimates of levels
  - Tail wagging the dog

Distrust of Model Selection

- Error structure may drive estimates of levels
  - Tail wagging the dog
- Error structure may drive misclassification

Distrust of Model Selection

- Error structure may drive estimates of levels
    - Tail wagging the dog
- Error structure may drive misclassification
- In an MLE horserace, can we really trust the winning model?
    - Is the *model* winning, or is the *error structure* winning?

Distrust of Model Selection

- Error structure may drive estimates of levels
  - Tail wagging the dog
- Error structure may drive misclassification
- In an MLE horserace, can we really trust the winning model?
  - Is the *model* winning, or is the *error structure* winning?
- *Hopefully*, cross-validation methods save the day

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games
3. For each model, generate fake game-play data from that model

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games
3. For each model, generate fake game-play data from that model
4. Fit all models to that fake data

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games
3. For each model, generate fake game-play data from that model
4. Fit all models to that fake data
5. Use various criteria to select a winning model

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games
3. For each model, generate fake game-play data from that model
4. Fit all models to that fake data
5. Use various criteria to select a winning model
6. See how frequently the "right" model wins

Our Motivation

Let's simulate a model-selection exercise!
(Inspiration: Salmon (2001) for learning models)

1. Pick several popular behavioral GT models
2. Pick a set of games
3. For each model, generate fake game-play data from that model
4. Fit all models to that fake data
5. Use various criteria to select a winning model
6. See how frequently the "right" model wins

Always a correct model. "Best-case" scenario.

Model Selection Criteria

MLE in-sample $\Rightarrow$ overfitting.

Penalty-based solutions:

1. **AIC** (Akaike, 1973): Log-Likelihood – (# params.)
2. **BIC** (Schwarz, 1978): Log-Likelihood – $\frac{1}{2}\ln(n)$(# params.)

Model Selection Criteria

MLE in-sample $\Rightarrow$ overfitting.

Penalty-based solutions:

1. **AIC** (Akaike, 1973): Log-Likelihood $-$ (# params.)
2. **BIC** (Schwarz, 1978): Log-Likelihood $- \frac{1}{2} \ln(n)$(# params.)

Cross-validation solutions:

1. Split the data (the games) into a "training" set and a "testing" set
2. Estimate parameters on training data
3. Measure likelihood on testing data
4. Repeat with different splits, take the average

Cross-Validation Methods

Our "subjects" play $n = 12$ games: $G = \{g_1, g_2, \ldots, g_{12}\}$

k-Fold Cross Validation:

- Randomly partition G into $\{G_1, \ldots, G_k\}$ of equal sizes
  - Split is *iid* across subjects
- In each "fold" $i \in \{1, 2, \ldots, k\}$...
  - Training data: $\bigcup_{j \neq i} G_j$
  - Testing data: G_i
  - Calculate log-likelihood LL_i
- Winning model: Maximum $\frac{1}{k} \sum_i LL_i$

Two-Fold Cross Validation (2FCV): $k = 2$
Leave-One-Out Cross Validation (LOOCV): $k = n = 12$

Connections

AIC & LOOCV:

- LOOCV → AIC as $n \to \infty$ (Stone, 1977)

Connections

AIC & LOOCV:

- LOOCV $\rightarrow$ AIC as $n \rightarrow \infty$ (Stone, 1977)

BIC & 2FCV:

- 2FCV $\rightarrow$ BIC for linear models, non-equal splits (Shao, 1997)

Connections

AIC & LOOCV:

- LOOCV → AIC as $n \to \infty$ (Stone, 1977)

BIC & 2FCV:

- 2FCV → BIC for linear models, non-equal splits (Shao, 1997)

Is there a consensus choice??

Methodology

- Twelve different symmetric 3×3 games. Games
  - 6 games: unique pure NE
  - 6 games: unique totally mixed NE
  - Based on Stahl and Wilson (1995)
- Seven competing models (e.g., Level-k, QRE)
- For each model, generate a dataset
  - Dataset = 3000 simulated subjects playing the 12 games
  - Play according to the model (model = "DGP")
  - Each subject's parameters are *iid* draws
- Pick winning model for each subject via AIC, BIC, 2FCV, and LOOCV

Q: Does the DGP win for the vast majority of subjects in its dataset?

Level-k (Nagel; Stahl & Wilson)

- Level 0: σ_i^0 is uniform random
- Level 1: $s_i^1 = BR_i(\sigma_i^0)$
- Level k: $s_i^k = BR_i(s_j^{k-1})$. Here: $k \in \{1, 2, 3, \text{N}\}$
- Base model is deterministic
- w/ spike-logit: $\sigma_i^k = (1-\epsilon)\mathbb{1}_{\{s_i^k\}} + \epsilon LR(s_j^{k-1}|\lambda)$

Double-Counting Error (LKD)

ϵLR_2	2nd-best response
$(1-\epsilon) + \epsilon LR_1$	Best response
ϵLR_3	3rd-best response

Single-Counting Error (LKS)

$\epsilon LR'_2$	2nd-best response
$(1-\epsilon)$	Best response
$\epsilon LR'_3$	3rd-best response

Poisson Cognitive Hierarchy (Camerer, Ho & Chong)

- Level 0: σ_i^0 is uniform random
- Level 1: $s_i^1 = BR_i(\sigma_i^0)$
- Level k: s_i^k = BR to Poisson dist'n over $\{\sigma_j^0, s_j^1, \ldots, s_j^{k-1}\}$. And $k \leq 3$
- Base model is deterministic
- w/ spike-logit: $\sigma_i^k = (1-\epsilon)\mathbb{1}_{\{s_i^k\}} + \epsilon LR(s_j^{k-1}|\lambda)$

Double-Counting Error (PCHD)

ϵLR_2	2nd-best response
$(1-\epsilon) + \epsilon LR_1$	Best response
ϵLR_3	3rd-best response

Single-Counting Error (PCHS)

$\epsilon LR_2'$	2nd-best response
$(1-\epsilon)$	Best response
$\epsilon LR_3'$	3rd-best response

Hierarchical Quantal Response (HQR)

- Level 0: σ_i^0 is uniform random
- Level 1: $\sigma_i^1 = LR_i(\sigma_i^0|\lambda)$
- Level k: $\sigma_i^k = LR_i(\sigma_i^{k-1}|\lambda)$
- Allow $k \leq 3$
- Base model is not deterministic. No spike-logit needed.

Quantal Level-k (QLK; Stahl & Wilson 1994)

- Level 0: σ_i^0 is uniform random
- Level 1: $\sigma_i^1 = LR_i(\sigma_i^0 | \lambda^1)$
- Level 2's Belief: $\sigma_i^{1(2)} = LR_i(\sigma_i^0 | \lambda^{1(2)})$
- Level 2: $\sigma_i^2 = LR_i(\sigma_i^{1(2)} | \lambda^2)$
- Only allow $k \leq 2$
- Base model is not deterministic. No spike-logit needed.

QRE (McKelvey & Palfrey 1995)

- You know it!
- Logit specification
- Principal branch
- Not deterministic. No spike-logit needed.

RESULTS (finally)

Result-Across Model Selection Comparison

Result 1 No model-selection criterion guarantees high accuracy

Fraction of 3,000 subjects for which the DGP wins.
(If models tie they share the win equally)

DGP (# Param.)	LOOCV	2FCV	BIC	AIC
LK Double (3)	21.48	22.62	0.72	0.93
LK Single (3)	50.12	40.81	57.6	59.1
PCH Double (4)	23.42	28.71	25.02	26.95
PCH Single (4)	20.95	16.45	35.57	40.17
QLK (4)	15.56	13.49	4.85	5.13
HQR (2)	21.88	22.45	92.5	92.87
QRE (1)	19.13	44.51	94.33	92.13

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK-Double	solo	**0.83**	0.73	0.40	0.37	0.63	0.47	2.6
	w/DGP	–	**92.83**	**62.13**	**62.10**	**62.10**	**92.93**	0.50
	w/Other	**92.97**	0.07	0.77	0.83	0.17	0.17	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK-Single	solo	3.43	**44.93**	2.63	4.23	2.2	3.97	10.23
	w/DGP	**21.87**	–	16.63	17.9	16.63	21.83	1.57
	w/Other	2.87	**23.13**	1.83	0.2	2.73	2.93	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
PCH-Double	solo	3.1	1.57	**3.7**	1.7	7.4	4.47	4.6
	w/DGP	47.07	47.03	–	**71.13**	48.13	47.1	1.33
	w/Other	1.53	0.07	**71.23**	0.07	0.7	2.1	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
PCH-Single	solo	5.4	19.87	4.6	**15.13**	6.3	8	14.17
	w/DGP	11.93	13.17	**19.47**	–	12.53	11.93	0.6
	w/Other	3.73	0.1	1.97	**20.7**	2.67	3.5	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QLK	solo	1.23	2.03	0.63	0.57	**1.73**	1.07	6.2
	w/DGP	82.6	82.47	84.03	84.03	–	**83.23**	0.4
	w/Other	1.13	0.43	0.2	0.47	**84.93**	0.97	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
HQR	solo	1.13	0.63	0.87	0.63	1.07	**2.43**	5.9
	w/DGP	**84**	83.1	54.9	54.83	55.99	–	0.17
	w/Other	0.17	0.43	1.97	2.23	0.03	**84.93**	0.00

Result (LOOCV)

The Problem: Ties.

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QRE	solo	**42.53**	0.9	0.53	0.67	1.6	1.73	**19.13**
	w/DGP	0.00	0.00	0.00	0.00	0.00	0.00	–
	w/Other	31.43	31.47	1.13	1.23	0.83	0.83	**0.00**

Result (LOOCV)

The Problem: Ties.

Result 2 When selecting among similar models it is important to verify the frequency of model non-identification (ties). BIC table 2FCV table

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QRE	solo	**42.53**	0.9	0.53	0.67	1.6	1.73	**19.13**
	w/DGP	0.00	0.00	0.00	0.00	0.00	0.00	–
	w/Other	31.43	31.47	1.13	1.23	0.83	0.83	**0.00**

The Cause

Consider DGP = LK-Double with $k = 2$

- Game payoffs are relatively high (\$0–\$100)
- Even modest $\lambda \Rightarrow$ near-perfect Level-2 play, even when trembling
- Models will estimate $\hat{k} = k = 2$ and $\hat{\epsilon} = 0$
- Also $\hat{\tau}$ and $\hat{\lambda}$ large
- All models get 100% likelihood

Possible Solution 1

Divide payoffs by 100

- Noisier play in DGPs $\Rightarrow$ better identification
- Noisier beliefs in QLK, HQR, and QRE can change base predictions

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
	solo	**8.07**						
LK-Double	w/DGP	–	**53.97**				**54.4**	
	w/Other	**56.3**						

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
	solo		27.37					
LK-Single	w/DGP	15.47	–		15.93		15.47	
	w/Other		20.43					

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
PCH-Double	solo			4.87				
	w/DGP			–	45.23			
	w/Other			47.33				

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
	solo				11.90			15.47
PCH-Single	w/DGP			15.53	–			
	w/Other				19.73			

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QLK	solo					**5.87**		**17.23**
	w/DGP					–	**28.10**	
	w/Other					**28.33**		

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
HQR	solo						**14.93**	**15.17**
	w/DGP						–	
	w/Other						**41.1**	

Table: Payoffs scaled by 1/100.

Result: Structural Change

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QRE	solo	**24.07**						**25.63**
	w/DGP							–
	w/Other							**0**

Table: Payoffs scaled by 1/100.

Result: Structural Change

Result 3 Structural changes to the games may not be enough to overcome the identification problems that arise when comparing similar models

BIC table 2FCV table

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
QRE	solo	**24.07**						**25.63**
	w/DGP							–
	w/Other							**0**

Table: Payoffs scaled by 1/100.

Possible Solution 2

- Omit similar models
    - 6 models other than QRE are level-based and tie frequently.
    - For each model, horserace that model against *only* QRE

Result: Two-horse horserace

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR	QRE
LK Double	95.13						4.87
LK Single		76.35					23.65
PCH Double			77.52				22.48
PCH Single				44.07			55.93
QLK					89.27		10.73
HQR						89.68	10.32
QRE DGP:	24.32	48.92	96.23	96.8	96.03	93.43	

Table: LOCCV winning frequency of each model versus only QRE in 3×3 games.

Result: Two-horse horserace

Result 4 Two-horse horserace fixes the non-identification problem (there are no ties), but model selection is still imperfect BIC table 2FCV table

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR	QRE
LK Double	95.13						4.87
LK Single		76.35					23.65
PCH Double			77.52				22.48
PCH Single				44.07			55.93
QLK					89.27		10.73
HQR						89.68	10.32
QRE DGP:	24.32	48.92	96.23	96.8	96.03	93.43	

Table: LOCCV winning frequency of each model versus only QRE in 3×3 games.

The Three “Mysteries”

Even without ties there are still three troublesome cases.
Why?

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR	QRE
LK Double							
LK Single							
PCH Double							
PCH Single				**44.07**			
QLK							
HQR							
QRE DGP:	**24.32**	**48.92**					

Table: LOCCV winning frequency of each model versus only QRE in 3×3 games.

Mystery #1: PCHS vs. QRE

PCHS: Varying τ varies pure BR's

- $k=2$: six different $(s_1^*,\ldots,s_{12}^*)$ BR vectors as τ varies

Case 1:

- Subject plays a perfect BR in 11 training games: $\hat{\epsilon}=0$
- Subject does not play the BR in testing game: $LL_i=-\infty$
- $(1/12)\sum_i LL_i=-\infty$

Case 2:

- Subject always plays non-BR in 11 training games: $\hat{\epsilon}=1$
- Subject plays BR in testing game: $LL_i=\infty$
- $(1/12)\sum_i LL_i=\infty$

Mystery #1: PCHS vs. QRE

Case 3:

- Subject plays "2nd-best" strategy in n_1 games and BR in $11 - n_1$
- $\hat{\epsilon} = \frac{n_1}{11}$ and $\hat{\lambda} = \infty$ since all trembles are on 2nd best
- Subject plays "3rd-best" in training game: $LL_i = -\infty$
- $(1/12)\sum_i LL_i = -\infty$

How often are there $-\infty$ problems?

Total PCHS subjects:	3,000
PCHS loses to QRE:	1,655 (55%)
PCHS gets $LL_i = -\infty$:	1,488 (90%)

Other models: failure rate low, but still high *fraction* of $-\infty$ problems

Mystery #2: QRE vs LK Models

- High payoffs ⇒ QRE subjects often play Nash
- LK estimated to be Nash type (95%), often noiseless (80%)
- QRE won't imitate this with $\hat{\lambda} = \infty$. Why?
    - Suppose mixed NE and $\sigma_i^*(s_i) < 1/3$
    - Subject plays s_i: better estimated as noise
    - Thus, $\hat{\lambda} < \infty$
- ⇒ LKS & LKD beat QRE

Mystery #3: QRE vs LK Double

- Suppose QRE subject with frequent Nash play
- QRE has fairly high (but finite) $\hat{\lambda}$
- LKD Nash type can have $\hat{\epsilon} > 0$
  - If $\sigma_i^*(s_i) < 1/3$ then s_i better fits as a tremble
- Suppose testing game has $\sigma_i^*(s_i) < 1/3$
- QRE likelihood $\approx \sigma_i^*(s_i) < 1/3$
- LKD likelihood $\approx (1-\hat{\epsilon})\sigma_i^*(s_i) + (\hat{\epsilon})\frac{1}{3}$
- LKD beats QRE

Summary

Two main problems for cross-validation:

1. $-\infty$ likelihoods
2. $\sigma_i^*(s_i) < 1/3$ scenarios

BIC and AIC shouldn't have these problems...

BIC Winning Frequencies

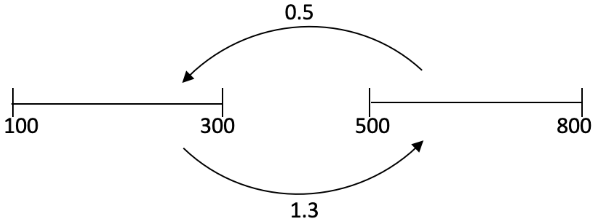
DGP\EST		LK Double	LK Single	PCH Double	PCH Single	QLK	HQR	QRE
LK Double		97.53						2.47
LK Single			93.07					6.93
PCH Double				92.17				7.83
PCH Single					76.7			23.3
QLK						93.5		6.5
HQR							93.97	6.03
QRE	QRE wins:	98.57	97.7	98.37	97.6	98.87	97.1	
	Model wins:	1.43	2.3	1.63	2.4	1.13	2.9	

Table: BIC winning frequency of each model versus only QRE in the 3 $\times$3 games with the original payoffs

Robustness

Would a larger strategy space improve model identification?

- 2-Person Guessing Games (CGC 2006):



- Fine strategy space: $\{100, 101, 102, \ldots, 300\}$
- Coarse strategy space: $\{100, 110, 120, \ldots, 300\}$

Robustness: Size of Strategy Space

LOOCV. Fine strategy set is Red and Coarse strategy set is Blue

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK Double	**86.67** **83.4**					
LK Single		**89.2** **83.97**				
PCH Double			82.07 82.2			
PCH Single				**81.8** **77.1**		
QLK					76 75.77	
HQR						72.5 71.93
QRE	**77.57** **70.97**	**84.97** **77.17**	94.27 95.17	94.43 94.3	94.7 94.87	93.6 94.3

Table: LOCCV winning frequency of each model versus only QRE in guessing games.

Robustness: Size of Strategy Space

BIC. Fine strategy set is Red and Coarse strategy set is Blue

DGP\EST	LK double	LK single	PCH double	PCH single	QLK	HQR
LK Double	**98.3** **96.9**					
LK Single		**98.2** **96.4**				
PCH Double			**87.4** **92.5**			
PCH Single				**87.1** **91.7**		
QLK					65.8 65.8	
HQR						69.3 69.4
QRE	**98.0** **97.0**	96.4 95.7	**99.7** **99.1**	**99.7** **99.0**	99.6 99.5	98.7 98.1

Table: BIC winning frequency of each model versus only QRE in guessing games.

Robustness: Size of Strategy Space

- LOOCV:
  - Fine is significantly better: 5 comparisons
  - Fine is significantly worse: 0 comparisons
- BIC:
  - Fine is significantly better: 5 comparisons
  - Fine is significantly worse: 2 comparisons

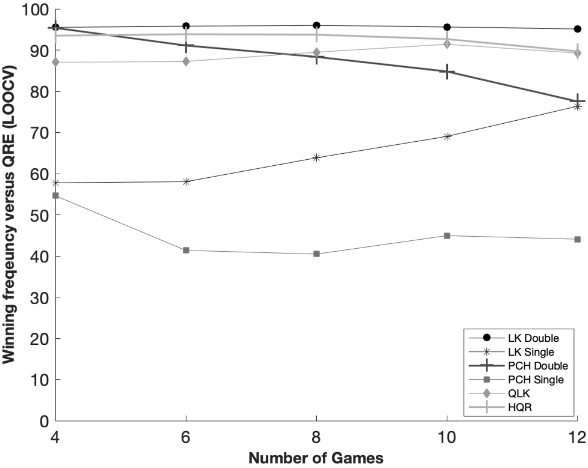
Robustness: Number of Games

What's the value of adding another game?

- Back to 3×3 games
- Original simulation: 12 games
- New simulations: random subsets of r games, $r \in \{4, 6, 8, 10, 12\}$

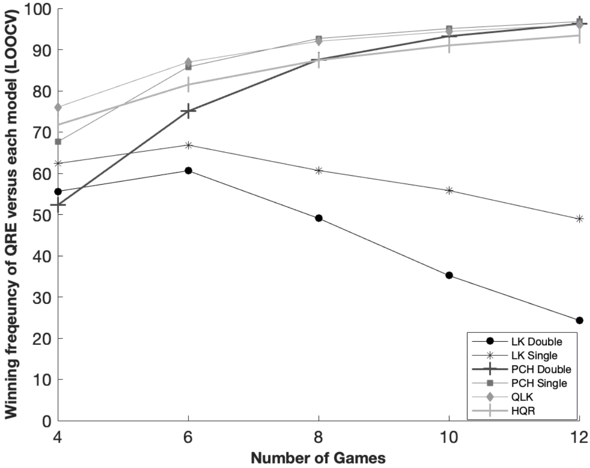
LOOCV: Number of Games

Levels models are DGPs:



LOOCV: Number of Games

QRE is DGP:

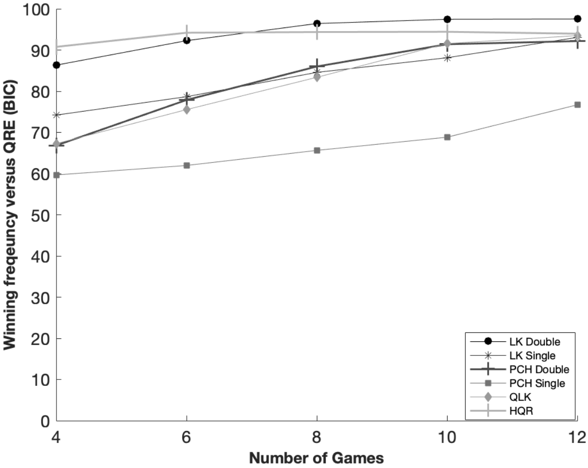


LOOCV: Number of Games

- Generally, more games helps
- When QRE is the DGP:
  - Lots of NE play
  - Small # games: LK has $-\infty$ problems
  - Large # games: LK's $\sigma_i^*(s_i) < 1/3$ advantage dominates

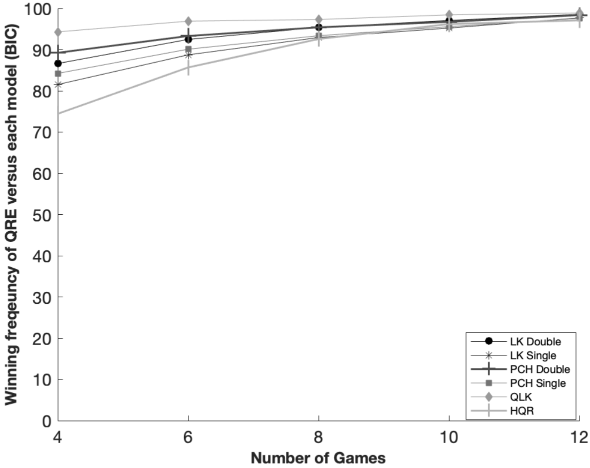
BIC: Number of Games

Levels models are DGPs:



BIC: Number of Games

QRE is DGP:



Conclusion

Lessons:

1. Don't include a bunch of similar models
2. Scaling payoffs/structural changes may not help much
3. Cross-validation can have problems!
    - Especially when models can become deterministic
4. Failures seem specific; few general lessons
5. BIC is less sensitive to such problems?
6. Finer strategy space: *mostly* better
7. More games: generally better, esp. for BIC

Want to run a model horserace? *Simulate it first*

Fin

Appendix - Tie BIC

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	0.7	2.1	0	0.47	0.4	95.07	1.17
	w/DGP	NaN	0	0	0	0.03	0	0
	w/Other	0.03	0	0.07	0.07	0	0	0
LK-Single	solo	1.1	57.6	0.37	4.87	0.73	31.1	3.97
	w/DGP	0	NaN	0	0	0	0	0
	w/Other	0.17	0	0.1	0.1	0.17	0	0
PCH-Double	solo	0.4	1.9	12.13	4.5	1.83	49.57	3.9
	w/DGP	0	0	NaN	25.77	0	0	0
	w/Other	0	0	25.77	0	0	0	0
PCH-Single	solo	0.97	21.83	3.17	31.33	1.57	21.07	11.5
	w/DGP	0	0	8.47	NaN	0	0	0
	w/Other	0.1	0	0	8.47	0.1	0	0
QLK	solo	0.17	1.33	0.1	0.73	4.8	89.03	3.7
	w/DGP	0.1	0	0	0	NaN	0	0
	w/Other	0	0	0.03	0.03	0.1	0	0
QR	solo	0.1	0.93	0	0.97	0.3	92.5	5.17
	w/DGP	0	0	0	0	0	NaN	0
	w/Other	0.03	0	0	0	0.03	0	0
QRE	solo	0.27	1.43	0.17	0.73	0.37	1.9	94.33
	w/DGP	0	0	0	0	0	0	NaN
	w/Other	0	0	0.8	0.8	0	0	0

Return

Appendix - Tie 2FCV

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	0.4	0.7	0.4	0.17	0.77	0.5	1.2
	w/DGP	NaN	93.83	63.3	63.07	63.07	93.87	0.03
	w/Other	94.1	0.17	1.3	1.47	0.3	0.3	0
LK-Single	solo	4.23	40.7	2.83	4.43	3.9	3.97	10.33
	w/DGP	19.5	NaN	14.8	18	14.8	19.5	0
	w/Other	4	22.7	3.9	0.33	3.63	3.27	0
PCH-Double	solo	0.87	0.63	1.97	0.9	2.27	0.7	3.93
	w/DGP	41.57	41.27	NaN	87.8	41.3	41.4	0.03
	w/Other	0.13	0.03	88.1	0.03	0.5	0.6	0
PCH-Single	solo	4.27	11.33	4.27	23.1	5.13	4.3	16.27
	w/DGP	10.93	13.73	20.4	NaN	10.93	10.97	0
	w/Other	4.07	0	3.8	23.2	4.43	4.4	0
QLK	solo	1.47	1	0.63	0.73	3.07	0.9	3.5
	w/DGP	86.4	86.13	86.3	86.13	NaN	87.57	0.03
	w/Other	0.7	0.17	0.83	0.3	87.7	0	0
QR	solo	1.33	0.73	0.53	1.13	0.3	1.33	3.87
	w/DGP	86.27	86.13	57.2	57.1	58.7	NaN	0.03
	w/Other	0.93	0.2	2.73	2.07	0.07	87.77	0
QRE	solo	7.33	1.5	3.6	3.83	2.63	4.47	72.77
	w/DGP	0	0	0	0	0	0	NaN
	w/Other	0.63	0.83	1.4	1.63	2.53	2.53	0

Return

Appendix - Tie BIC 1/100 Scale

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	7.37	15.5	0.9	3.8	1.83	57.23	12.7
	w/DGP	NaN	0	0	0	0.33	0	0
	w/Other	0.33	0	0.33	0.33	0	0	0
LK-Single	solo	4.73	36.77	0.43	7.87	1.23	30.13	18.4
	w/DGP	0	NaN	0	0	0	0	0
	w/Other	0.23	0	0.2	0.2	0.23	0	0
PCH-Double	solo	2.43	8.4	8.27	9.47	2.83	33.67	19.13
	w/DGP	0	0	NaN	15.43	0	0	0
	w/Other	0.37	0	15.43	0	0.37	0	0
PCH-Single	solo	2.13	16.7	3	23.4	2.03	21.4	22.87
	w/DGP	0	0	8.27	NaN	0	0	0
	w/Other	0.2	0	0	8.27	0.2	0	0
QLK	solo	0.97	6.73	0.9	5.8	17.23	29	39.07
	w/DGP	0.2	0	0	0	NaN	0	0
	w/Other	0	0	0.1	0.1	0.2	0	0
QR	solo	0.87	7.1	0.43	5.6	1	40.03	44.83
	w/DGP	0	0	0	0	0	NaN	0
	w/Other	0.07	0	0.07	0.07	0.07	0	0
QRE	solo	0.83	5.27	0.6	3.47	1.9	5.03	82.57
	w/DGP	0	0	0	0	0	0	NaN
	w/Other	0.07	0	0.27	0.27	0.07	0	0

Return

Appendix - Tie 2FCV 1/100 Scale

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	11.23	10.5	8.13	5.03	6.73	4.3	13.3
	w/DGP	NaN	33.3	25.07	23.73	24.13	33.5	0.23
	w/Other	35	2.27	0.47	2.53	3.23	3.23	0
LK-Single	solo	9.87	18.27	6.27	8.13	5.03	5.03	18.27
	w/DGP	19.63	NaN	14.93	17.63	14.97	19.67	0.13
	w/Other	3.03	22.37	2.8	0.17	4.57	4.17	0
PCH-Double	solo	6.97	8.17	6.57	8.63	6.57	5.33	17.47
	w/DGP	22.33	20.27	NaN	32.17	21.4	20.57	0.07
	w/Other	0.57	1.67	34.33	1.67	4.17	4.07	0
PCH-Single	solo	7.83	10.8	6.93	11.9	6.5	5.8	20.47
	w/DGP	12.43	14.77	20.23	NaN	12.97	12.43	0.17
	w/Other	3.13	0.03	2.97	22.57	4.93	4.83	0
QLK	solo	9	10.63	9.87	10.43	10.53	8.93	28.13
	w/DGP	2.43	0.13	1.4	0.1	NaN	8.03	0.07
	w/Other	2.57	1	1.93	1.03	8.93	0.7	0
QR	solo	9.93	8.83	8.77	8.73	9.03	14.8	26.5
	w/DGP	4.03	2.93	2.77	1.93	8.03	NaN	0
	w/Other	3	2.03	1.63	1.13	0.6	9.27	0
QRE	solo	7.73	6.3	8.23	8.07	8.73	8.9	46.93
	w/DGP	0	0	0	0	0	0	NaN
	w/Other	1.33	1.17	1.23	1.37	3.27	3.17	0

Return

Appendix - Two-horserace BIC

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	88.33						11.67
LK Single		97.2					2.8
PCH Double			92.17				7.83
PCH Single				76.7			23.3
QLK					93.5		6.5
QR						93.97	6.03
QRE	1.43	2.3	1.63	2.4	1.13	2.9	
	98.57	97.7	98.37	97.6	98.87	97.1	

Table: BIC winning frequency of each model versus only QRE in 3×3 games.

Return

Appendix - Two-horserace 2FCV

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	77.57						22.43
LK Single		97.33					2.67
PCH Double			91.57				8.43
PCH Single				55.72			44.28
QLK					95.05		4.95
QR						94.57	5.43
QRE	12.38	4	8.7	8.02	9	10.97	
	87.62	96	91.3	91.98	91	89.03	

Table: 2FCV winning frequency of each model versus only QRE in 3×3 games.

Return

Appendix - Tie LOOCV Guessing Games with Finer Strategy Sets

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	22.63	30.1	7.83	5.17	1.93	2.63	3.4
	w/DGP	–	17.73	16.93	13.47	0	0.03	0
	w/Other	21.23	2.43	0.03	2.47	2.63	2.63	0
LK-Single	solo	20.67	33.83	7.13	5.4	1.73	2.63	3.4
	w/DGP	16.63	–	13.2	15.93	0	0	0
	w/Other	3.27	19.37	3.27	0	2.57	2.57	0
PCH-Double	solo	17.63	11.93	15.23	14.9	1.73	2.87	7.47
	w/DGP	16.67	12.1	–	17.5	0.8	0.83	0
	w/Other	0.03	3.27	21.97	3.27	3	3	0
PCH-Single	solo	17.17	12.8	14.23	15.7	1.7	2.8	7.33
	w/DGP	12.13	15.47	17.33	–	0.43	0.43	0
	w/Other	4.3	0	4.43	20.87	3.47	3.53	0
QLK	solo	0.9	3.63	2.23	20.2	27.03	23.87	15.8
	w/DGP	0.17	0	1.63	0.37	–	4.83	0
	w/Other	0.27	0.6	0.7	0.6	4.83	0.87	0
QR	solo	1.27	12.63	1	5	9.2	45.23	17.33
	w/DGP	1.5	0.17	1.8	0.5	2.7	–	0
	w/Other	0.17	4.67	0.17	4.67	0	3.5	0
QRE	solo	12.43	10.83	0.17	1.9	1	2.67	69.57
	w/DGP	0	0	0	0	0	0	–
	w/Other	0.43	0.47	0.93	0.67	0.87	1.1	0

Return

Appendix - Two-horserace BIC Guessing Games with Finer Strategy Sets

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	98.2						1.8
LK Single		98.23					1.77
PCH Double			87.37				12.63
PCH Single				87.07			12.93
QLK					65.77		34.23
QR						69.27	30.73
QRE	2.03	3.57	0.33	0.33	0.4	1.33	
	97.97	96.43	99.67	99.67	99.6	98.67	

Table: 2FCV winning frequency of each model versus only QRE in 3×3 games.

Return

Appendix - Two-horserace 2FCV Guessing Games with Finer Strategy Sets

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	46.87						53.13
LK Single		91.4					8.6
PCH Double			75.6				24.4
PCH Single				81.02			18.98
QLK					72.03		27.97
QR						72.9	27.1
QRE	31.53	0.9	8.83	9.83	9.93	9.33	
	68.47	99.1	91.17	90.17	90.07	90.67	

Table: 2FCV winning frequency of each model versus only QRE in 3×3 games.

Return

Appendix - Tie LOOCV Guessing Games with Coarser Strategy Sets

DGP \EST		LK Double	LK Single	PCH Double	PCH Single	QLK	QR	QRE
LK-Double	solo	18.37	29.3	6.47	4.97	2.17	4.37	3.03
	w/DGP	–	22.83	12.53	8.83	0	0.03	0
	w/Other	26.43	2.03	0	2.03	2.87	2.87	0
LK-Single	solo	15.9	40.63	5.2	5.27	2	2.5	3.4
	w/DGP	16.93	–	8.4	10.77	0	0	0
	w/Other	3.33	19.4	3.33	0	2.37	2.4	0
PCH-Double	solo	8.27	20.53	18.07	12.13	2.8	3.57	5.17
	w/DGP	16.63	11.5	–	15.9	0.27	0.33	0
	w/Other	0.6	2.73	22.63	2.67	3.63	4.1	0
PCH-Single	solo	7.43	21.23	15.77	16.1	2.7	3.77	5.2
	w/DGP	9.37	13.43	14.77	–	0	0	0
	w/Other	6.13	1.63	5.87	18.83	2.97	3.2	0
QLK	solo	4.5	11.33	4.1	6.23	23.93	17.67	12.5
	w/DGP	4.5	0	3.3	0	–	7.4	0
	w/Other	8.3	4	2.9	4	7.4	7.2	0
QR	solo	3.77	11.3	2.57	7.1	8.53	39.57	14.83
	w/DGP	3.3	0	1.57	0	3.5	–	0
	w/Other	3.67	6.5	0.8	3.37	0	5.3	0
QRE	solo	12.1	8.6	0.83	2.6	0.97	1.73	60.7
	w/DGP	0	0	0	0	0	0	–
	w/Other	11.9	11.9	0.1	0.1	0.47	0.47	0

Return

Appendix - Two-horserace BIC Guessing Games with Coarser Strategy Sets

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	96.43						3.57
LK Single		97					3
PCH Double			92.53				7.47
PCH Single				91.73			8.27
QLK					65.77		34.23
QR						69.4	30.6
QRE	1.63	2.73	0.33	0.53	0.4	1.17	
	98.37	97.27	99.67	99.47	99.6	98.83	

Table: 2FCV winning frequency of each model versus only QRE in 3×3 games.

Return

Appendix - Two-horserace 2FCV Guessing Games with Coarser Strategy Sets

DGP\EST	LK double QR	LK single QR	PCH double QR	PCH single QR	QLK	QR	QRE
LK Double	83.9						16.1
LK Single		97.48					2.52
PCH Double			75.85				24.15
PCH Single				83.15			16.85
QLK					71.23		28.77
QR						72.17	27.83
QRE	36.1	30.9	8.37	9.77	11.4	10.63	
	63.9	69.1	91.63	90.23	88.6	89.37	

Table: 2FCV winning frequency of each model versus only QRE in 3×3 games.

Return

Games

Return

Game1	T	M	B
T	25	30	100
M	40	45	65
B	31	0	40

Game 2	T	M	B
T	30	50	100
M	40	45	10
B	35	60	0

Game3	T	M	B
T	10	100	40
M	0	70	50
B	20	50	60

Game 4	T	M	B
T	30	100	50
M	40	0	90
B	50	75	29

Game5	T	M	B
T	30	100	22
M	35	0	45
B	51	50	20

Game 6	T	M	B
T	40	15	70
M	22	80	0
B	30	100	55

Game7	T	M	B
T	25	30	100
M	40	0	65
B	31	45	40

Game 8	T	M	B
T	10	100	40
M	0	70	60
B	20	50	50

Game9	T	M	B
T	39	15	70
M	40	80	0
B	30	100	55

Game10	T	M	B
T	30	50	100
M	40	60	10
B	[illegible]	[illegible]	[illegible]

Game11	T	M	B
T	30	100	22
M	35	0	20
B	[illegible]	[illegible]	[illegible]

Game12	T	M	B
T	40	80	60
M	23	25	0
B	[illegible]	[illegible]	[illegible]

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: 2nd International Symposium on Information Theory, Akadémiai Kiadó, Budapest, Hungary. pp. 267–281.

Costa-Gomes, M., Crawford, V.P., 2006. Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. American Economic Review 96, 1737–1768.

Salmon, T.C., 2001. An evaluation of econometric models of adaptive learning. Econometrica 69, 1597–1628.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464. doi:10.1214/aos/1176344136.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica sinica , 221–242.

Stahl, D.O., Wilson, P.W., 1995. On Players' Models of Other Players: Theory and Experimental Evidence. Games and Economic Behavior 10, 218–254.

Stone, M., 1977. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical

Society: Series B (Methodological) 39, 44–47.
doi:10.1111/j.2517-6161.1977.tb01603.x.