A Predictive Evaluation of Econometric Models of Adaptive Learning

Gregory Chernov
Ivan Susin
Sergey Cheparuhin

April 2, 2020

Abstract

Game-theoretic models of learning are hard to study even in the laboratory setting due to econometric and practical concerns (such as limited length of the experimental session). In particular, as (Salmon, 2001) simulations shows, in a “blind testing” of several models data generated by those models does not correspond correctly to the estimated parameters. Thus, even when the real data generation process is known we cannot distinguish correct and incorrect models by looking at the estimates. However, we demonstrate that under the same conditions models are clearly distinguishable if instead of the parameters of the models we compare predictions that the models make. We also provide a rationale for why predictive quality is a particularly relevant metric for a learning model.

1 Introduction

Game theory allows for both deep theoretical and very applied experimental work: we can study games both by examining their mathematical structure and by looking at how people play it in the laboratory. However, there is a clear rift between those two approaches. On the one hand, theory makes heavy use of the notion of equilibrium, while empirically it is sometimes elusive. On the other hand, empirically processes of updating, learning, even teaching and experimenting are often very recognizable, but are difficult to formalize. There is a considerable amount of literature
devoted to this problem, both theoretical and empirical, but the major hurdle is in between - it is econometric. Development of theoretical model demands thorough pondering which models are close to the reality and what are the main features left out by the existing models. It is a challenging problem.

Seminal work of (Salmon, 2001) used ingenious approach to highlight those difficulties. Its idea is similar to the blind testing process used, for example, in wine tasting. This implies that we can identify a particular specimen among different unlabeled ones, and can specify it with its own correct label. In this case it transforms to using of several variations of a learning algorithm. Learning algorithm takes the previous history of the play and produces the next action. There are infinite ways to select an action if we allow randomization, so there is potentially a whole space of possible algorithms. But within this space there are certain basic types that reflect theoretical foundations (Cournot best response, fictitious play, reinforcement learning and others). At the very least we should be able to distinguish between those basic ones.

However, simulations in that paper show that classic econometric tool (hypothesis testing in a maximum likelihood model) fails in this blind testing challenge. Surely, expanding the sample size would eventually overcome this problem, however, whole idea of Salmon was to reflect the natural limitations of the laboratory experiments. So it could be taken as an implication that laboratory experiments are naturally underpowered to study learning.

This work shows that above-mentioned interpretation is incorrect and too narrow. We first replicate (Salmon, 2001) to show that indeed, standard econometric methods are problematic in this setting. However, we propose a cross-validation procedure based on a prediction of behavior instead of in-sample fitting measure that overcomes this problem. Thus, under the same conditions unlabeled data from different data generating processes can be recognized by the quality of predictions that different models (including the true one) give.

1.1 Literature review

There are two major generations of the models of learning in games, roughly divided by the publishing of the (Fudenberg, Levine, 1998) textbook. First generation of models was developed independently in various sciences (e.g. reinforcement learning stems from biology) and culminates in “Experience-Weighted Attraction” model (Camerer, Ho, 1999) that incorporates major prior models as special cases. This generation is characterized by the hope to specify a single model that can represent all varieties of learning behavior. Second generation largely abandons this
hope and tries to reflect some specific aspects of the learning process, such as experimenting behavior or limited memory. We briefly overview two major classes of models that form the backbone of the field and EWA model in particular.

The first approach is known as “reinforcement learning” and has plentiful biological and computer science applications. Consider the player that reflects on the success of their own action and reinforces the propensity to take action that is more fruitful. Second approach, called “fictitious play”, disregards those successes. Their player builds a fictitious version of the opponent (hence the name) by taking previous actions by the opponent as a statistical sample and chooses best response to such distribution of probable actions.

Experience-weighted Attractions (EWA) model (Camerer, Ho, 1999) was a major breakthrough. It can flexibly account both for fictitious play and reinforcement learning elements of learning by incorporating them as edge cases into the model. Note, however, that being an edge case is not a desirable property to fit a maximum likelihood model. In this case the estimator will necessarily be biased, because points away from the edge are always more likely than the edges. However, without extensive simulation it is hard to say how big is this bias. It is highly unlikely that this problem has analytic solution.

Empirical forecasting literature consists mostly of purely experimental studies, but among methodological contributions are Peysakhovich and Naecker (2017) and Kleinberg, Liang and Mullainathan (2017) that use the percentage of the possible improvement over random guessing. (Xie, 2019) studies stability of learning models and identification issues under varying the learning environment. With a nested hybrid models it demonstrates that EWA is unstable when the form of monetary payoffs changes.

(Fudenberg, Liang, 2019) predicts the modal action in a variety of games using machine learning techniques. The focus of this paper is on the across-game robustness of the machine learning predictor, while we study a predictive quality of tailored models that may be (but are not necessarily) game-specific.

The first example of comparing experimental metadata and a wide range of theoretical and heuristic approaches of models of learning is (Mathevet, Romero, 2014). It is based on prediction of average payoffs counted under several rounds. Moreover, it provides a thorough and meticulous study of forecasting metrics.
2 Main part

2.1 The EWA model

Main model of learning in games that we and (Salmon, 2001) use is called Experience-Weighted Attraction (EWA) (Camerer, Ho, 1999). It allows to describe theoretically very different rules through different values of parameters of the same empirical model. Essentially when the values of parameters are set to different values, we obtain samples with different “true” theoretical models. Those include “fictitious play” (FP) that model that plays a best response to the empirical distribution to the opponent’s actions and “reinforcement learning” that increases probability to play more successful actions (reinforces those actions). The basic version of EWA contains 6 parameters:

- $\rho$ and $\phi$ - discounting factors
- $N(0)$ - strength of initial experiences
- $A^j_i(0)$ - form of initial experiences
- $\delta$ - relative weight of hypothetical and actual payoffs
- $\lambda$ - sensitivity to attraction.

Those parameters form “observation-equivalents” (function $N()$), then “attractions” (function $A()$) and, finally, attractions map into behavior through logit probabilities of choices:

\[
N(t) = \rho N(t-1) + 1
\]

\[
A^j_i(t) = \frac{\phi N(t-1) A^j_i(t-1) + [\delta + (1-\delta)\mathbb{I}(s^J_i, s_i(t))] \pi_i(s^J_i, s_{-i}(t))}{\rho N(t-1) + 1}
\]

\[
P^j_i(t+1) = \frac{\exp(\lambda A^j_i(t))}{\sum_{k=1}^{m_i} \exp(\lambda A^k_i(t))}
\]

While the first and third formulas are straightforward discounting and multinomial logit functions respectively, attraction function for a particular action consists of two term (normalized by $N(t)$ function): first is just a discounting of previous
attraction, the second is a weighted (by $\delta$) sum of payoffs $\pi_i(s^t_i, s_{-i}(t))$ from that action generally (like in fictitious play) and from that action conditional on whether this action was actually taken ($\mathbb{I}(s^t_i, s_i(t))$, like in reinforcement learning). Thus $\delta$ controls similarity to the fictitious play versus reinforcement learning, while $\rho$ and $\phi$ control the length of memory that is used in decision-making.

We consider four basic populations similarly to (Salmon, 2001)\[1\]

<table>
<thead>
<tr>
<th>population</th>
<th>Parameter In DGP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho$</td>
</tr>
<tr>
<td>EWA_FP (fictitious play)</td>
<td>1</td>
</tr>
<tr>
<td>EWA_CB (Cournot best response)</td>
<td>0</td>
</tr>
<tr>
<td>EWA_RL (reinforcement learning)</td>
<td>0</td>
</tr>
<tr>
<td>EWA_mdl (middle)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

So far we simulate each of four populations by creating 500 play-sets. Each play-set contains 500 couples of players interacting during 40 rounds.

Consider, for example, the EWA_FP population (that is, the parameters $[\rho, \delta, \phi, N(0), \lambda]$ take the values $[1, 1, 1, 0, 1]$). 40 pairs of the EWA_FP players play a repeating game of 2x2 (matching pennies) - it forms 1 play-set. The game is repeated for each of the 500 play-sets and for each of three matrices we produce 500 play-sets.

### 2.2 What we do

Basic logic of simulations in our work follows those in (Salmon, 2001). First, we simulate the data of two learning algorithms playing against each other. By generating our data we know the true data generation process (DGP). To compare different models of learning now we have to estimate the parameters of those models on the empirical data that we had created. We expect the correct model to fit better than an incorrect one and those estimated parameters should point out the right model. The main question is how significant is the difference between them. So the workflow is the following: DGP selection $\rightarrow$ data generation $\rightarrow$ estimation $\rightarrow$ comparison of estimated data to DGP.

\[1\]There are two more populations in the paper, one is just a different flavor of FP algorithm and the last one is a mixed population. Mixed populations add additional layer of complexity (see (Wilcox, 2006)) and will be the focus of future work.
(Salmon, 2001) concludes that statistical analysis is not powerful enough to sort though the simulated data. Analysis is based on the confidence intervals: if the data was generated with the same value of the parameter that it is tested against, test should be rejected very rarely. Stronger claim is that actual percentage should correspond to the selected significance level.

Our main contribution is the way we evaluate our candidate DGPs: we compare not the values of the parameters, but relative success of the predictions that those parameters help to produce. So, while we do not statistically differentiate hypotheses about the parameters themselves, we are able to confidently reject alternative models based on their relative performance. Note that even if our criterion fails to separate the models (and it is always possible to achieve by creating minimally different models and a small sample), it happens for a good reason: because those models produce similar predictions.

We also provide more reliable simulation suit (it seems that quite a number of original simulation runs failed to actually calculate estimates which may have introduced additional bias) and somewhat streamline the estimation by using likelihood ratio tests.

2.3 Simulations

We create 4 populations of EWA over 3 games and estimate parameters for each set. Thus, first criterion identification conclusions of EWA analyses the distribution of parameters of each of the 500 play-sets for each game matrices.

Parameters are estimated with the maximum likelihood estimator:

\[
LL(A(0), N(0), \phi, \rho, \delta, \lambda) = \sum_{t=1}^{T} \sum_{i=1}^{N} \ln \left( \sum_{j=1}^{m_i} I(s_i^j, s_i(t)) \cdot P_i^j(t) \right)
\]

\[
= \sum_{t=1}^{T} \sum_{i=1}^{N} \ln \left( \sum_{j=1}^{m_i} I(s_i^j, s_i(t)) \cdot \frac{e^\lambda A_i^j(t-1)}{\sum_{k=1}^{m_i} e^\lambda A_i^k(t-1)} \right)
\]

Parameters are restricted to \( \rho \in [0, 1], \delta \in [0, 1], \phi \in [0, \infty), N_0 \in [0, 1], \lambda \in [0, \infty) \) (following (Camerer, Ho, 1999, p.18.)). For each of the 5 parameters and each play-set, the program evaluates 5 numerical values.

\[2\text{Matrices are in the appendix} \]

\[3\text{via Python 3.7's optimize and scipy minimize packages with BFGS algorithm.} \]
Our statistical fit exercise deviates from (Salmon, 2001) by using the LR test. We compare likelihoods of the unrestricted model to the model with the parameter set to some fixed value.

We consider three types of restrictions: the one where 4 of 5 parameters (all except \(\lambda\)) correspond to the true DGP, the \(\delta\) set to 0 or to 1, and \(\rho\) as either 0 or 1. For each population and each matrix we calculate ratio indicating the percentage of play-sets where the restrictions are significant (e.g. for 5 percent significance).

For a predictive measure we use a slightly different principle of data evaluation. Note that EWA is stochastic and not deterministic algorithm. Each pair of players “looks” at the game history, estimates the attractions of each available action and play-sets those attractions as a probability of playing this action. Thus, actions are realisations of a discrete random variable.

To compare predictions for a specific round, “hypothetical players” from all analysed populations (that is, different values of EWA parameters) are given the game path up to that round and their probabilities for choosing each action are recorded as their predictions. Since we know the realized outcomes, we can compare those predictions to the fact.

Basically, all such measures at the core are counting the weighted aggregate of mistakes, their difference lies either in what aggregate we take (sum, product or something else) or in what weighting function we apply to those mistakes (whether all mistakes are treated equal or some mistakes are worse than other, for example). Given that all three game matrices in (Salmon, 2001) have 0 or 1 payoffs, any additive measure just sums up the number of errors in a particular model.

We use a multinomial version of a standard metric called Brier score. Let

\[ \text{Score} = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{R} (P_{ti} - F_{ti})^2 \]

4In the appendix we show that logarithmic loss metric produces qualitatively the same results
It increases in number of mistakes, thus lower the the score – better the model. Because it depends on the number of possible actions, we also want to adjust it for comparisons on game matrices of different size.

3 Results

The correctness of the code itself (whether it actually calculates what it should) was checked with two types of tests. First, we manually calculated the attraction values in a 2x2 game for 40 periods and compared them with the code calculations. Second test is a replication of the results from a experimental econometrics textbook (Moffatt, 2015).

Histograms here present density of different estimates of the parameters of the 2x2 matrix \( [2] \) (in other cases, the picture only gets worse and 6x6 histograms are placed in appendix). Value after the name of parameter in the legend is the true value. It is evident that the values of most estimated parameters obtained by MLE are close to the true values, however, \( \rho \) for EWA_mdl and N(0) have a clear skewness or are distributed almost uniformly.

Our conclusions here confirm (Salmon, 2001) in that parameters of the model are habitually not what they should be. This instability of estimation affects the performance of statistical tests.

The main statistical test results are listed in the table \( [3] \) In the columns with the four parameters, test fixed values according to the values in true DGP (as an illustration consider EWA_FP \( [\rho[1], \delta[1], \phi[1], N_0[0]] \) and EWA_mdl \( [\rho[0.5], \delta[0.5], \phi[0.5], N_0[0.5]] \)). Here only \( \lambda \) is a free parameter and fraction of populations that passed the test is far from encouraging.

Only for EWA_CB the test demonstrates some self-recognition, though far from the expected 95%. Remaining columns contain results of tests for one fixed parameter. Theoretically, in EWA_FP \( \delta = 1 \) and for most play-sets test shows it. But correct \( \phi \) is rejected almost everywhere.

Unlike MLE, Brier score does not rely o a combination of parameters that is “plausible” for all individuals within a single play-set conditional on observations. For each individual average of the model’s predictions with a fixed set of parameters can be directly compared to the alternatives. Results for 2x2 matrix are in Fig. \( [1] \). The true model for each population is on the right, while all the tested potential models are on the left by the y-axis. The average Brier scores are on the-x axis with green and red signifying over- and under-performance relative to the random forecast. Width of the horizontal bars indicates an interval of two standard deviations. Finally, a
Kruskal-Wallis rank sum test shows how significant is the difference between two “neighboring” average Brier values (for instance for [EWA_CB, EWA_CB] cell test compares averages of EWA_CB and EWA_mdl groups). As is evident already from the standard deviations themselves, all those differences are statistically significant.

In all cases true models show a better result in self-prediction than alternative models for all three game matrices (see the appendix). This is not just because Brier score uses more information about the probability of using alternative strategies than MLE which uses only the strategies that was actually used, because the result stays the same if logarithmic loss function (as it is predictive metric which uses only the implemented forecasts, most similar to MLE, see in the appendix Fig. [6]) is considered.

![Figure 1: Average brier score for 2x2 matrix](image)

Two standard deviation from average brier better than random *** <0.001

worse than random ** <0.01

Kruskal-Wallis test for two group with closest mean brier

Figure 1: Average brier score for 2x2 matrix
3.1 Why classic statistical measures of fit are not very suitable for evaluation the models of learning?

First, note that most problems in classical statistics concern cross-section or at least stationary time series. However, adaptive learning is not only a nonstationary process, it is highly reactive. Each observation in adaptive learning potentially may mean very different things depending on what model is used and what was the previous play history. Models should be sensitive enough to capture those reactions. This means predicting the next move well even when the context is new.

Second, model-identifying information is by construction more scarce in the learning data than in an ordinary time series. We can differentiate models only when they predict different things and can be evaluated on their performance. However, learning process usually results in an equilibrium, a very predictable state once it is reached. All models will predict the same actions for pure strategy (and the same probability distribution for mixed) equilibrium, so once models capture equilibrium dynamics (as they should do and usually do), it is impossible to distinguish between them. Similar problem arises in the first rounds - though all models use available information, in the initial, "burn-in" period this information is scarce. Without a significant "common ground" of previous history players have to rely on their priors (for example, history of playing similar games applied by analogy and so on) that cannot be captured by the model. The initial play is naturally very randomized and we cannot distinguish models when they are almost random number generators. So both the beginning and the endpoint is not informative - given a stable behavior of the opponent, most models will agree on the same (correct) prediction and will be equally bad at predicting random numbers. Therefore, in a standard experimental setting only the middle part, where the actual "learning" happens, can be used to distinguish different models.

The final problem with ordinary statistical measure of fit is a conceptual one. What is the final goal when we model learning? We want to know what the player will play. In order to know what the player will play in a new environment (and because games are reactive, it is very hard not to have a new environment) we want to understand why he played what he did in the past. Then our inference about the future will be primarily based on the changes to those reasons. While predictive metrics are well-suited to fit this reasoning, classic statistical approach is, evidently, not so much.

---

5 Even when we want, for example, to understand the learning process, prediction can be used to evaluate the quality of this understanding.
3.2 Limitations

While our results are inspiring, there are clear obstacles and limitations on the way forward. Let’s summarize some of them.

**Other models.** By construction, our analysis is only valid for the models that we explicitly compare (and possibly some nested within them ones). It is possible that the method that we propose allows to reject EWA altogether in favor of another, more sophisticated or simpler model. Due to this pseudo-out-of-sample prediction, it is more robust to common “overfitting” criticism of EWA than in-sample fit. Here we concentrate on developing this comparison method itself.

**Heterogeneity.** (Wilcox, 2006) shows that if the model is misspecified in assuming the homogeneity of heterogeneous population, reinforcement learning models will implicitly capture that heterogeneity and will be mistakenly taken as more fitting. Given that both (Salmon, 2001) and we operate with truly homogeneous populations, we don’t have to address this concern, however it should be further investigated using the measure of fit that we employ here.

**Sensitivity to game matrix.** Game matrices used in experiments are not representative of all possible game matrices or, more importantly, of the economically relevant ones. It is entirely possible that there are games where evaluating models is significantly more challenging that in those that (Salmon, 2001) and we used. It is an empirical question and we provide a new tool to answer it with.

**Sensitivity to behavioral factors.** Experimental scientist cannot exclude the possible effect of factors which influence they ex ante disregard (framing of the experiment, time of the day, specific characteristics of experimental pool and so on). We don’t really know which factors are actually important for predicting behavior in the experimental games, but we are doing our best to understanding it and to account for it.

4 Conclusion

In a simulated environment it is possible to perform a “blind tasting” task if we use predictive quality of models and not goodness of their fit. This ability to correctly identify the correct model in a simulation setting is only a first step. The real world rarely presents us with econometric problems where the actual data generation process is known in advance. However, previous literature gave a pessimistic impression that even in this case we cannot do much. While this method is sure to have serious limitations somewhere, it is likely to expand possibilities everywhere else.
The reason we need models is exactly that we don’t know the real rules for real behavior. But we can compare the behavior generated by the model to the same model fitted on that generated data, and thus identify the true model.

If we are unsure what is the true model, we can simulate all or some of the candidates and compare their behavior. This horse-race of the models not only shows what model is superior, but also how much of a difference does misspecification make. Ideally we want to be able not only to selectively breed better models, but also to quantify costs and likelihood of errors given the inescapable model uncertainty.

Why does prediction in our task outperform classic statistical theory? It seems that predictive statistics is more tailored to learning processes that are reactive and nonstationary time series than the standard statistics that was developed to handle iid samples. In particular, when parameters in actual DGP are on the edge of their domain, we cannot expect them to be unbiased. We also have no good reasons to expect their distribution to be something from the standard toolbox. Thus, classic tests are bound to fail.

Predictive statistics, however, relies only on the outputs of the model, not on the distribution of its’ inner parts. Thus, even if the model is slightly misspecified or has unexpected distribution of errors, we still can reasonably compare it to other models. This robustness also helps us to derive useful implications from this comparison as we don’t have to get the model absolutely right when we just want it to fit better than other considered models.

With the development of machine learning and the automation of economic processes, mechanisms, institutions and regulations, reliable and explicitly formulated learning models will be invaluable.

References


Appendix

**MATRIX 1**

This matrix, often referred to as 'Matching Pennies,' has a single mixed strategy Nash equilibrium of (.5, .5) for both players.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>4.0</td>
</tr>
</tbody>
</table>

**MATRIX 2**

This matrix has several mixed strategy Nash equilibria that are various combinations of the strategies (.25, .25, .25, .25), (.5, 0, 0, .5) and (0, .5, .5, 0) for both players.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0</td>
<td>2.0</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
MATRIX 3

This matrix has a mixed strategy Nash equilibrium at (.4, .2, .2, .2, 0, 0) for the row player and (.4, .2, .2, .2, 0, 0) for the column player and one at (.4, .2, .2, .2, 0, 0) for the row player and (.4, .2, .2, .2, 0, 0) for the column player.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>2.0</td>
<td>2.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
<td>2.0</td>
<td>2.0</td>
<td>0.2</td>
</tr>
<tr>
<td>6</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
<td>0.2</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Table 2: Procedural correctness verified by (Moffatt, 2015) data

<table>
<thead>
<tr>
<th></th>
<th>ρ</th>
<th>δ</th>
<th>φ</th>
<th>N(0)</th>
<th>λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>True parameters</td>
<td>0.97</td>
<td>0.6</td>
<td>0.94</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>Moffatt’s Stata estimation</td>
<td>0.98</td>
<td>0.54</td>
<td>0.94</td>
<td>0.82</td>
<td>0.88</td>
</tr>
<tr>
<td>Our estimation</td>
<td>0.97</td>
<td>0.61</td>
<td>0.94</td>
<td>1</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Table 3: Fraction of populations for which the restrictions are met (should be either 95% or 5%)

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Model</th>
<th>rho; delta; phi; N(0) equal to true DGP</th>
<th>mean p value</th>
<th>delta = 1 mean p value</th>
<th>delta = 0 mean p value</th>
<th>delta = 0 mean p value</th>
<th>this one param hold'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EWA_FP</td>
<td>0.0</td>
<td>0.31</td>
<td>0.7</td>
<td>0.94</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>EWA_CB</td>
<td>0.25</td>
<td>0.65</td>
<td>0.875</td>
<td>0.98</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>EWA_RL</td>
<td>0.05</td>
<td>0.36</td>
<td>0.0</td>
<td>0.0</td>
<td>0.85</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>EWA_mdl</td>
<td>0.0</td>
<td>0.15</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**2x2**

|        | EWA_FP    | 0.0                                    | 0.29         | 0.8                     | 0.96                   | 0.0                    | 0.0                  |
|        | EWA_CB    | 0.4                                    | 0.73         | 0.875                   | 0.98                   | 0.0                    | 0.0                  |
|        | EWA_RL    | 0.0                                    | 0.46         | 0.0                     | 0.75                   | 0.9                    | 0.0                  |
|        | EWA_mdl   | 0.0                                    | 0.17         | 0.0                     | 0.0                    | 1.0                    | 1.0                  |

**4x4**

|        | EWA_FP    | 0.1                                    | 0.4          | 0.725                   | 0.85                   | 0.0                    | 0.0                  |
|        | EWA_CB    | 0.325                                   | 0.69         | 0.8                     | 0.94                   | 0.0                    | 0.0                  |
|        | EWA_RL    | 0.025                                   | 0.45         | 0.0                     | 0.85                   | 0.96                   | 0.0                  |
|        | EWA_mdl   | 0.0                                    | 0.14         | 0.0                     | 0.0                    | 1.0                    | 1.0                  |

**6x6**

|        | EWA_FP    | 0.1                                    | 0.4          | 0.725                   | 0.85                   | 0.0                    | 0.0                  |
|        | EWA_CB    | 0.325                                   | 0.69         | 0.8                     | 0.94                   | 0.0                    | 0.0                  |
|        | EWA_RL    | 0.025                                   | 0.45         | 0.0                     | 0.85                   | 0.96                   | 0.0                  |
|        | EWA_mdl   | 0.0                                    | 0.14         | 0.0                     | 0.0                    | 1.0                    | 1.0                  |
Figure 3: Parameter values histograms for 6x6 matrix
Figure 4: Average Brier score
Figure 5: Average Brier score
Figure 6: Average Logarithmic score