Testing game-theoretic comparative statics using Bayesian model selection*

Ozan Aksoy† University of Oxford

Jeroen Weesie‡ Utrecht University

February 20, 2015

Abstract

Aksoy & Weesie (2013) propose a formal-game theoretic model which yields comparative statics predictions about behavior of actors and actors' expectations about behaviors of others in non-repeated Prisoner's Dilemma games. Aksoy & Weesie (2013) consider five competing specifications of this model. In this note, we present a Bayesian statistical method to test these five competing specifications, taking the hierarchical nature of the data into account. While this paper is a follow-up to Aksoy & Weesie (2013), the method we discuss is suitable to test multiple comparative statics and is, thus, useful to asses the fit of other game-theoretic models.

1 Introduction: problem formulation

Game-theoretic models typically yield a number of comparative statics predictions. Testing these comparative statics often require several (mean) comparisons. It is not trivial to find a statistical procedure that combines the evidence from a number of such comparisons. If the aim is to test and compare several theoretical models each of which yields a number of comparative statics predictions, the task is even more difficult. In this note, we demonstrate the benefits of recently developed Bayesian tools, particularly the Deviance Information Criterion (DIC) and posterior predictive p-values (PPP) (Gelman et al. 2013, Ch6) for checking and comparing the fit of competing game-theoretic models.[1]

Aksoy & Weesie (2013), henceforth AW13, propose a formal model to explain cooperation in non-repeated social dilemmas. This model has three components: non-selfish motives, expectations about others' non-selfish motives, and a game-theoretic decision model. AW13 analyze, in total, five alternative specifications of their formal model: the social orientation model, a "normative"' model, and three versions of an inequality aversion model. The three versions of the inequality aversion model differ in how expectations are modeled. Because the inequality aversion specifications yield highly complex and non-monotonic predictions, AW13 follow a two-step empirical strategy. AW13 first test the macro-level comparative statics, ignoring the nested structure of the data (decisions

*Forthcoming in *Journal of Mathematical Sociology*.

†Corresponding author, Department of Sociology and Nuffield Centre for Experimental Social Sciences, University of Oxford; Nuffield College, New Road OX1 1NF, Oxford, UK (e-mail: ozan.aksoy@sociology.ox.ac.uk).

‡ICS/Department of Sociology, Faculty of Social Sciences, Utrecht University, Padualaan 14, 3584 CH, NL.

[1]There is also a frequentist, albeit not widely used literature on "combining p-values", see e.g., (Won et al. 2009).

are nested in subjects, see next section). This test rejects all of the three inequality aversion specifications. In the second step, AW13 fit the remaining specifications to data, acknowledging the nesting. Below we demonstrate how Bayesian tools can be used to test and compare the five theoretical specifications, taking the nesting in the data into account.

2 Data

AW13 used the Asymmetric Investment Game (AIG) framework that had the outcome structure of the Prisoner's Dilemma (PD) (also see Aksoy & Weesie 2009). The experimental design included nine AIGs each of which had a different set of game outcomes. Each of the 134 subjects played each of these nine AIGs. Furthermore, one game–the full symmetric AIG–was played twice by each subject, thus the number of games played by each subject was 10. The order in which these 10 games were played was varied in two factors, and in each of these 10 games the partner was a new subject, i.e. stranger matching. After subjects played the first five AIGs without feedback, their expectations about the behavior of their interaction partners in these five AIGs were asked after which they received feedback on the actual decisions of their partners. Then, subjects played the remaining five AIGs.

3 Statistical analysis

The five alternative specifications of the formal model of AW13 predict choices and expectations in the 10 AIGs. The predictions are in the following comparative statics form. Each theoretical specification predicts a threshold non-selfish motive parameter value for each of the 10 AIGs, such that a subject with a non-selfish motive parameter value exceeding this threshold cooperates and defects otherwise. For the details of the predictions and threshold values see Table 4 and Table 5 on pages 40 and 42, respectively in AW13. At the macro level, lower levels of cooperation are predicted in games with higher threshold values. At the individual level, a theoretical specification predicts a Guttman pattern: given the vector of 10 AIGs, ordered based on the threshold values, a subject is expected to switch from cooperating to defecting at most once. For example, assume that we ordered the 10 AIGs under the social orientation specification. The behavioral patterns 1111100000 and 1100000000 where 1's denote cooperation and 0 defection, are perfectly consistent with the social orientation specification, whereas the pattern 1100000011 is not. The behavioral patterns of subjects who cooperate or defect in all 10 games are consistent with any of the five theoretical specifications. The same as choices, the games can be ordered with respect to expectations.

To model statistically the cooperative choices of subjects, we use the following Rasch-like model in which the probability that subject $i = 1, ..., 134$ cooperates in game $j = 1, ..., 10$ in period $k = 1, ..., 10$ is:

$$\text{logit}(\Pr(\text{C}_{\text{ijk}} = 1)) = \frac{\delta_i - t_j}{\exp(\beta_0 + \beta_1(k-1))} \tag{1}$$

where δ_i is the individual non-standard utility parameter ("ability") and t_j is the threshold non-selfish utility value given by the theoretical specification. t_j is similar to the "item difficulty" parameter in the standard Rasch model. Different from Rasch models, here t_j is not a parameter but a known value given by the theoretical model, hence denoted by a Roman letter. The term

in the denominator captures "decision error". Decisions become more erratic as the denominator increases. Here, the decision error is "exogenous" to the theory, i.e.,, subjects do not take error into account when making their decisions. For an alternative approach where error is endogenized, see e.g., Aksoy & Weesie (2014). Because the order in which the games were played was not totally randomized, but varied in two factors, it is important to control for a possible period effect. We, thus, make decision error depend on period via β_1.

We fit (1) using the free Bayesian software OpenBugs (Spiegelhalter et al. 2009). The estimation routines and datasets are available from the authors. For the social orientation and normative specifications, we assume that δ is normally distributed in the subject pool. For the three inequality aversion specifications, a log-normal distribution is assumed for δ because in theory inequality aversion is assumed to be positive. Sensitivity analyses (not reported) assuming normal distributions for the inequality specifications yield qualitatively the same results. We use the following rather uninformative priors. For β_0, β_1, and the mean of δ (for the inequality aversion specifications, the mean of the natural logarithm of δ) we use normal priors with zero mean and 1000 variance. For the variance of δ (for the inequality aversion specifications, the variance of the natural logarithm of δ), we use a Gamma(1000,1000) prior.[2] For all specifications, the posterior distribution of 60.000 draws is estimated after running 30.000 burn-in Markov Chain Monte Carlo (MCMC) iterations from two chains. Convergence is ascertained by visual inspection of the history of the two chains. For a detailed description of this Bayesian procedure for a similar set-up, see Aksoy & Weesie (2014).

In this note, we are interested only in the *overall fit* of the theoretical specifications and do not discuss individual parameter estimates. We use two methods to asses fit, Deviance Information Criterion (DIC) and Posterior Predictive P-values (PPP) which we discuss below in detail.

3.1 Deviance Information Criterion

DIC is a Bayesian counterpart of Akaike's Information Criterion. DIC is particularly suitable for multilevel models and is used to compare non-nested models where smaller values indicate better fit. Informally, DIC = mean deviance + $2p_D$ where mean deviance is the deviance (-2 times the log-likelihood) averaged over the simulated parameter values in the MCMC process. p_D is the *effective* number of parameters, thus DIC penalizes less parsimonious models. In multilevel models determining the effective number of parameters is not straightforward and calculation of p_D is complex. Readers can refer to Gelman & Hill (2007), Ch24 and Fox (2010), Ch3 for a summary and discussion on DIC.

3.2 Posterior predictive p-values

PPPs are used to check whether a model differs systematically from the observed data. In a typical setup, the researcher defines a discrepancy statistics T (in the current study T is the number of Gutmman errors as explained below). Then, for each MCMC draw, a new dataset is replicated from the fitted model so that once the MCMC process is completed an ensemble of replicated datasets, conditional on the model parameters, is obtained. A distribution of T under the null can easily be obtained from this ensemble. A comparison of the observed value of T with this distribution gives

[2]The t_j values are rescaled to be between 0 and 10 for all of the five specifications such that the above priors remain uninformative.

a posterior predictive p-value. If the discrepancy between the model and data is too high, PPPs are expected to be small.

This procedure is quite appealing as it is very flexible and can be used for any T. Moreover, PPPs can be broken down to assess the fit of different parts of the model, as we demonstrate below in the results section. On the other hand, there are some limitations of PPPs. A common issue is related to the double use of data: the observed data is first used to obtain the posterior distribution of parameters and used once again to compute the PPP. Because of this double use, PPPs can behave in unnatural ways, e.g., the distribution can be non-uniform on [0,1] under the null concentrating around 0.5. Effectively, this means that the power of PPPs can be low. Another potential issue with PPPs is that while the posterior distribution of model parameters depend only on the likelihood but not on, e.g., the data collection protocol, PPPs, as frequentist p-values, do depend on the data collection protocol. PPPs should be interpreted bearing these caveats in mind, not as definitive tests of null hypotheses, but as useful summaries of model misfit (see Fox (2010), Gelman & Hill (2007), and Gelman et al. (2013) for a detailed discussion). This is the reason why we report DICs next to PPPs.

We obtain PPPs in the following way. Under each theoretical specification, we calculate the total number of Guttman errors, i.e., the number of pairs that violate the Guttman pattern as our deviance statistics T. For example, in the aforementioned behavioral pattern of 1100000011 there are 12 Guttman errors.[3] The distribution of Guttman errors under the null is obtained by the replication procedure described above. Very small or very large PPPs, e.g, smaller than 0.05 or larger than 0.95 indicate bad fit. We also fit models predicting expectations and obtain separate DICs and PPPs for expectations.[4]

4 Results and discussion

Table 1 shows the results. DIC and PPP scores point to very similar conclusions. When subjects' own behavior is considered all but the normative model is rejected. When expected behavior is considered all three versions of the inequality aversion specification are rejected but the normative and social orientation models fit equally well.

An important feature of posterior predictive checking is that separate fit statistics can be obtained for different elements of the model. This feature is particularly useful to understand where the model fails. In the current case, for example, one can obtain a separate PPP for each subject by comparing the Guttman errors observed for a particular subject with the distribution replicated for that subject. We followed this procedure. Under the best fitting inequality aversion model (USE, $F = U[0, 2]$), when cooperative behavior is considered 77 out of 134 subjects have PPP scores smaller than 0.05 or larger than 0.95. Under the normative model, on the other hand, only 49 subjects (37%) have such extreme PPP scores. Note that, as we discussed above, one should be careful in interpreting PPPs. Also, one should, in principle, be worried about the inflated type-I error when one performs that many tests (in this case 134, one per subject). Yet, we are not using

[3] For the games with the same threshold values, any switch is counted as a Gutmann error. The more lenient approach of not counting switches in tied games as errors yielded the same conclusions.

[4] We experienced convergence problems when predicting expectations, especially for parameters related to error. Plugging in estimates from the models predicting behavior did not solve the convergence issue. As a result, for expectations we fitted "$\text{logit}(\Pr(C_{ijk} = 1)) = \widetilde{\delta}_i - \gamma \cdot \widetilde{t}_j \quad \text{with} \quad \gamma > 0$". To ensure $\gamma > 0$, we imposed a Gamma(1000,1000) prior for γ. This specification is a reparametrized version of the specification in equation 1 where β_1 is assumed to be zero. Further alternative specifications, such as assuming a log-normal distribution for γ, yielded the same results.

Table 1: Bayesian fit measures for five theoretical specifications obtained for behavior and expectations of others' behavior. N(subject)=134, N(behavior)=10, N(expectation)=5.

	Own behavior		Expected behavior	
Specification	DIC	PPP	DIC	PPP
Social orientation with CSE/USE	1174	0.002	657	0.193
Normative model with CSE/USE	1159	0.510	658	0.216
Inequality aversion with CSE	1660	0.000	887	0.001
Inequality aversion with USE, $F = U[0, 2]$	1455	0.000	787	0.025
Inequality aversion with USE, $F = U[0, 10^2]$	1457	0.000	787	0.025

these PPPs as definite tests of null hypotheses but to see where the inequality aversion models seem to fail. Information on which models fail for which subjects is potentially very useful. For example, using these subject-level PPPs, one can classify subjects into distinct social motive categories. Similarly, it is possible to calculate separate PPPs for each pair of games so that one can track down which particular comparative statics is not well predicted by which theory. Such extensions are beyond the scope of this short note.

The Bayesian statistical arsenal includes powerful tools to asses model fit. We believe that these tools are particularly suitable for testing formal models which predict several comparative statics. In this note, we demonstrate the benefits of these Bayesian tools with an application to the experimental data of Aksoy & Weesie (2013). In their original paper, the model selection method Aksoy & Weesie (2013) used as a part of their analysis assumed that individual observations were independent when observations were, in fact, nested in subjects. The improved method we present here does take this nesting into account and corroborates the conclusions of Aksoy & Weesie (2013). Moreover, the method is suitable to track down for which particular subjects or which particular comparative statics the models fail.

References

Aksoy, O. & Weesie, J. (2009), 'Inequality and procedural justice in social dilemmas', *Journal of Mathematical Sociology* **33**(4), 303–322.

Aksoy, O. & Weesie, J. (2013), 'Social motives and expectations in one-shot asymmetric Prisoner's Dilemmas', *Journal of Mathematical Sociology* **37**(1), 24–58.

Aksoy, O. & Weesie, J. (2014), 'Hierarchical Bayesian analysis of outcome-and process-based social preferences and beliefs in Dictator Games and sequential Prisoners Dilemmas', *Social Science Research* **45**, 98–116.

Fox, J.-P. (2010), *Bayesian item response modeling: theory and applications*, New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vektari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis, Third Edition*, FL: CRC press.

Gelman, A. & Hill, J. (2007), *Data analysis using regression and multilevel and hierarchical models*, New York: Cambridge University Press.

Spiegelhalter, D. L. D. D., Thomas, A. & Best, N. (2009), 'The BUGS project: Evolution, critique and future directions (with discussion)', *Statistics in Medicine* **28**, 3049–3082.

Won, S., Morris, N., Lu, Q. & Elston, R. C. (2009), 'Choosing an optimal method to combine p-values', *Statistics in Medicine* **28**(11), 1537–1553.