Measuring the usefulness of statistical models [1]

David Azriel[a] and Yosef Rinott[b]

February 29, 2016

[a] Faculty of Industrial Engineering and Management, The Technion.
[b] The Federmann Center for the Study of Rationality, The Hebrew University, and LUISS, Rome.

Abstract

In this paper we propose a new measure, to be called GENO, of the usefulness of statistical models in terms of their relative ability to predict new data. It combines ideas from Erev, Roth, Slonim, and Barron (2007), from the well-known AIC criterion for model selection, and from cross-validation. GENO depends on the nature of the data and the given sample size, reflecting the fact that the usefulness of a model depends on both the adequacy of the model as a description the data generation process, and on our ability to estimate the model's parameters. Our research was motivated by the study of modeling decision making processes based on data from experimental economics, and and we also provide a detailed biological example.

1 Introduction

This paper presents a measure of usefulness of models in terms of their predictive power. Our work was motivated by the study of models in experimental economics, an area that has witnessed a surge of activity in the past several years. Game-theoretic experiments are conducted in numerous labs, typically producing large amounts of data in each experiment. The goal is to model learning processes and strategy choices in relatively simple games, and to assess the predictions of game theory. Classical hypotheses testing usually results in rejecting any model due to the large size of the data sets, demonstrating the first part of G. Box's saying "*All models are wrong, but some are useful.*"

When all models are wrong, model selection can be based on the classical Akaike information criterion AIC (Akaike, 1974) and its variations. Akaike (1983) and others proposed various functions of the AIC value of a given model relative to the best model as measures of the quality of the given model. These measures include Akaike differences and Akaike weights; see, e.g., Burnham and Anderson (2002) and Claeskens and Hjort (2009). Such measures appear rarely in application, whereas the AIC for model selection has been applied in numerous studies. This is probably due to certain difficulties in interpreting these measures, to be discussed later.

Clearly, the quality of models depend on the sample size with which they are to be used. For example, with a small sample, a complex model may be inappropriate due to overfitting, and Box's saying should be modified to "All models are wrong, but some models are useful

[1] This is part of a research project in experimental game theory with Shmuel Zamir and Irit Nowik supported by the Israel Science Foundation (grant no. 1474/10).

for certain sample sizes, and other models for other sample sizes". Indeed, the new measure we propose, GENO, is a function of the sample size with which the model will be used. For example, if model k stands for polynomial regression of a certain degree, or a M-step Markov chain, GENO(n, k) provides information on the quality of the model in question as a function of the sample size n, the model's dimension, and its fit to the data. GENO is a relative measure, and its value depends on the list of candidate models being considered. The fit to the data is measured in terms of a penalized likelihood of the data under the model with parameters estimated by maximum likelihood, akin to the AIC criterion. As shown later, this commonly accepted criterion can sometimes make very reasonable models look very bad. Like other AIC-related measures, GENO is a function of suitable AIC (Akaike) differences.

Our initial motivation in this work comes from experimental economics data and in particular from Erev, Roth, Slonim, and Barron (2007), henceforth ERSB, where models are compared according to a measure they called ENO (Equivalent Number of Observations), to be discussed later. We use our new measure GENO to study and compare models for the data of ERSB. A full description of the data set and the models we analyze is given in Section 5, and the required formulation of GENO for decision processes is given in Section 4. Figure 1 plots 500 sequential choices of actions of one pair of players in a certain game. The choices are binary and are smoothed by a moving average of window size 11. The smoothed data are compared to smoothed decision probabilities made according to three of the decision models considered: Nash equilibrium, Reinforcement Learning and a Markov model.

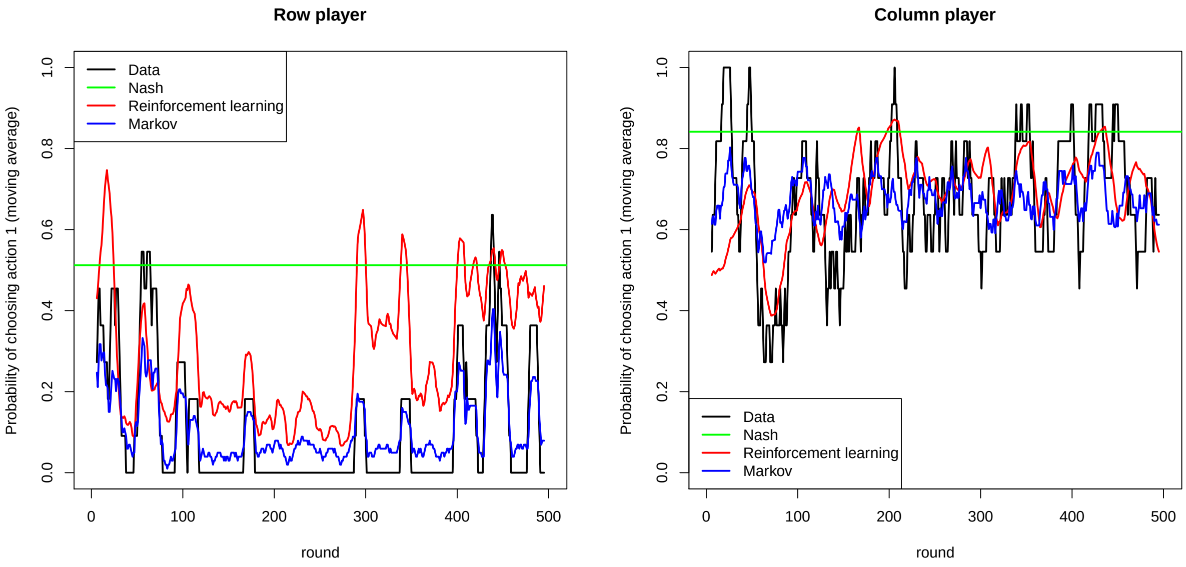


Figure 1: Plots of a moving average (of window size 11) of 500 sequential binary choices of a pair of players compared to three models: Nash equilibrium, reinforcement learning and a Markov model.

The Nash equilibrium assumes a constant known probability of choosing Action 1 out of two possible actions during the repeated game and in the two other models the probability changes based on previous actions and rewards. The Nash equilibrium does not provide good predictions and the other models are better. For the row player, the Markov model seems to fit the observed smoothed decidions better than reinforcement learning and for the column player, they seem to preform equally well. However, the Markov model has more parameters, and may not be a good model due to overfitting. Our goal is to quantify such statements in a meaningful way, and provide estimates and confidence intervals for this quantification.

As noted by ERSB, the Nash equilibrium model performs poorly in this data set, and this is evident also by our new measure GENO. As this model has no parameters that should be estimated, its GENO does not depend on the number of observations, that is, it is constant. A variation on the model CAB-1 of Plonsky, Teodorescu and Erev (2015), which is also a model with no parameters, performs better, and as a result the GENO value of Nash is zero. If CAB-1 is removed from consideration, then the estimated GENO of Nash is equal to 3, which means that after only 3 observations, there is a model in our list of candidate models that outperforms it. We found also that the Markov models have larger GENO than the reinforcement learning models of ERSB and hence they are preferred according to our criterion. For example, our estimates imply that a Markov model with only 32 observations can obtain the same expected likelihood of future data as the best reinforcement model (RLS) with 50 observations. Therefore, in this data set, the Markov models can obtain the same expected likelihood as the reinforcement learning models with fewer observations, and hence are preferred according to GENO. Another variation on CAB-1, called CAB-W, which is a one-parameter model, performs with only 14 observations as well as the RLS model with 50 observations. Therefore GENO of RLS with 50 observations is 14. As mentioned above, this is in tune with the AIC criterion, but the learning models may turn out better by other criteria. For further details see Section 5 and in particular Table 3.

ERSB's measure ENO of the value or usefulness of models for strategies in repeated games can be roughly described as follows. Consider two ways of predicting the proportion of playing a certain strategy in a given experiment. The first is to model players' behavior as a stochastic process, estimate its parameters, and use the estimated model for prediction. The second way consists of just using the empirical or sample proportion of playing each strategy. Assuming the models considered are only approximations, their predictions are generally inconsistent, and therefore the empirical estimator, being consistent under some mild assumptions, will be more accurate for a sufficiently large sample of players. On the other hand, a simple parametric model may provide a better predictor given a small sample. Consider a sample size m of players that are used to determine the empirical proportions of playing certain strategies. A model's ENO is an estimate of the sample size m for which the empirical proportions yield equally accurate predictions as the model. ENO does not depend on the sample size used for estimating the model's parameters, which is assumed to be large, and therefore it can be seen as quantifying the value of a model if its parameters are known or estimated with a very large sample.

We briefly discuss AIC differences. The AIC value of a model for a given data set from some source, is an estimate of the expected log-likelihood multiplied by -2 (a factor which we will not use in our definitions in Section 2 and later) of new (future) independent data from the same source, under the model with parameters estimated from the given data set. For a good model, the likelihood of such new data should be large, and hence the AIC as defined above should be small. Given a list of K candidate models, denote the AIC value of the kth model by AIC_k, and let AIC_{min} be the AIC of the model with the smallest AIC, that is, the model selected by the AIC criterion. The AIC differences Δ_k and weights w_k are defined as follows:

$$\Delta_k = \text{AIC}_k - \text{AIC}_{min} \quad \text{and} \quad w_k = \frac{e^{-\frac{1}{2}\Delta_k}}{\sum_{i=1}^{K} e^{-\frac{1}{2}\Delta_i}} \,. \tag{1}$$

It is hard to interpret the "raw" AIC differences. The AIC weights are sometimes interpreted as probabilities of models conditioned on the data. With a uniform prior on the set of models (which makes little sense) this interpretation could be meaningful if we believe that one of the models is true. Otherwise, the weights are still informative, but their interpretation is less clear. In our data, with models of different qualities and dimensions, some of the Δ_is take large values, making the weights numerically sensitive.

Our new measure of usefulness of a model is partly inspired by ENO and the developments of AIC. We take the liberty of calling it GENO (Generalized ENO). The generalization goes in several directions. ENO quantifies the value of a model for predicting proportions relative to empirical proportions. In the spirit of AIC differences, we compare models relative to the best model according to the AIC criterion, rather than just to empirical frequencies. Furthermore, we generalize from just predicting proportions to prediction of the whole process. Our estimation procedure is very different, and much simpler that that of ERSB. It is important to note that unlike ENO, our measure depends on the sample size. This is natural since the predictive value of a model depends not only on the underlying process generating the data, but also on the size of the sample used in estimating the model's parameters.

Our goal in defining GENO is to provide meaningful comparisons of different models with different sample sizes. Also, we use it to quantify the value of a model using data from different sources with different values of the model's parameters. These properties and others are demonstrated by our applications.

We next describe GENO informally. Formal definitions will be given in Section 2 for independent variables, and in Section 4 for Markov decision processes. We start with a comparison of two models. Consider two parametric models, indexed by k and ℓ with parameters $\theta^{(k)} \in \Theta^{(k)} \subseteq \mathbb{R}^{d_k}$ and $\theta^{(\ell)} \in \Theta^{(\ell)} \subseteq \mathbb{R}^{d_\ell}$. For any n, the expected log-likelihood (henceforth we may just say likelihood) of future data under a given model, if its parameters were to be estimated on the basis of past data (sample) of size n, is set as our criterion of the quality of the model for sample size n. We estimate it for any n on the basis of a given sample of size N, where N need not equal n. We proceed as follows: **(a)** For a given value of n we use our sample of size N to estimate the expected likelihood of future data for model k if its parameters were estimated using a sample of size n. **(b)** We compute a value m such that the expected likelihood under future data for model ℓ, if its parameters were estimated using a sample of size m, equals the likelihood estimated in part (a). The resulting value of m is denoted by GENO(n, k, ℓ). Estimation here means maximum likelihood estimation.

A comparison in terms of the sample size required by one model (or test, or estimator) to be as good as another with a given sample size, is closely akin to the notion of Pitman efficiency, see, e.g., Zacks (1975).

Note that if GENO(n, k, ℓ)> n, say, then model ℓ requires more observations than n in order to have equal quality as model k with n observations, and therefore for sample size n model k is preferable. This may happen, for example, if $d_k < d_\ell$, that is, model k has fewer parameters. When n is small, model k may be better, however, when n is large, model ℓ may become the better model, since a larger sample size allows estimation of more parameters without overfitting.

Given a list of K candidate models indexed by $1, \ldots, K$, we define

$$\mathrm{GENO}(n, k) = \min_{\ell \in \{1,\ldots,K\}} \mathrm{GENO}(n, k, \ell).$$

GENO(n, k) represents the sample size needed for the best competitor model to obtain the same expected likelihood as model k with n observations. Clearly GENO$(n, k, k) = n$, implying

$\text{GENO}(n,k) \le n$.

We believe that given a data set of size N, there is interest in comparing models not only under the assumption that they will be implemented with this N. A statement such as: model 1 is better than model 2 for sample size ≤ 100, say, and then model 2 is better, may be generally informative to someone having a sample of $N = 200$, say, and in particular further experiments of the same kind with different sample sizes are planned.

In general, given data of size N generated by some process, our goal is to quantify and compare the predictive quality of different models to be used with different sample sizes which do not necessarily coincide with N. For a list of K candidate parametric models we define our measure of the quality, $\text{GENO}(n,k)$, of the kth model as a function of the sample size n and discuss its estimation on the basis of the N observations. We first present GENO in the case of iid samples, and apply it to study of usefulness of Hardy-Weinberg type models to DNA data. We then extend it to Markov decision processes, which we apply to the analysis of data from experimental game theory.

2 GENO for iid observations

2.1 Definitions

In this section we define GENO and its estimation formally. For simplicity we start with samples of iid observations. Let $Y_1, Y_2, \ldots$ be iid random variables from a distribution having an unknown density g to which we refer as the 'true' density. Throughout the paper we write expressions like $g(y)dy$ although the distribution need not be continuous, and g can be a density with respect to any suitable measure. A list of parametric candidate models or families of densities for Y_i are considered: $\{f_k\} = \{f_k(y, \theta^{(k)})\}_{\theta^{(k)} \in \Theta^{(k)}}$, where $\Theta^{(k)} \subseteq \mathbb{R}^{d_k}$, $k = 1, \ldots, K$. We do not assume that g must be in any of these families, however, g and $\{f_k\}$ are assumed to be densities with respect to a common measure. Given a sample $Y_1, \ldots, Y_n$, let

$$\widehat{\theta}_n^{(k)} := \arg \max_{\theta^{(k)} \in \Theta^{(k)}} \sum_{i=1}^{n} \log f_k(Y_i, \theta^{(k)})$$

be the maximum likelihood estimator (MLE) for the kth model based on n observations. Henceforth we consider only models and sample sizes for which MLEs exist uniquely. Moreover, the AIC type approximations below require standard conditions on the models' likelihood function and their derivatives which, in particular, entail that the MLEs are asymptotically normally distributed. See Conditions A1 - A6 in White (1982).

In general we are interested in studying 'good' models, and we assume that our candidate models are adequate models, that is, they are reasonably close in the Kullback-Leibler sense to the true g. The approximations in the following sections are valid under this assumption, which is standard in the AIC literature.

Consider the models $\{f_k\}$ at the MLE $\widehat{\theta}_n^{(k)}$ based on a sample $Y_1, \ldots, Y_n$ from g, that is, $f_k(\cdot, \widehat{\theta}_n^{(k)})$. In the spirit of Akaike's AIC and cross validation, we imagine that a new independent sample $Y_1^*, \ldots, Y_{n^*}^*$ from g is observed. The expected average log-likelihood of the new data, (which in the normal case coincides with a sum of squares of deviations), is given by

$$\frac{1}{n^*} E \sum_{i=1}^{n^*} \log f_k(Y_i^*, \widehat{\theta}_n^{(k)}) = E \int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy; \tag{2}$$

the expectation on the left is with respect to both the Y_i^*'s whose density is g and with respect to the MLE $\widehat{\theta}_n^{(k)}$, and the expectation on the right is only with respect to the latter. In view of (2), the size, n^*, of the Y^*'s sample, does not play any role and can be taken to be one. The expression in (2) quantifies the quality of model k given n observations.

For two models f_k and f_ℓ in our list of candidates, we define $\mathrm{GENO}(n,k,\ell)$ in view of (2) as the value m such that

$$\mathrm{GENO}(n,k,\ell) = \left\{ \sup_{m \geq r} \, : \, E\int_{-\infty}^{\infty} g(y) \log f_\ell(y, \widehat{\theta}_m^{(\ell)}) dy \leq E\int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy \right\}, \tag{3}$$

where r is the minimum number of observations for the MLE $\widehat{\theta}_m^{(\ell)}$ to exists, and sup over the empty set is taken to be r. In words, $\mathrm{GENO}(n,k,\ell) = m$ means that model k with n observations is equivalent, in terms of expected log-likelihood, to model ℓ with m observations. Note that the larger $\mathrm{GENO}(n,k,\ell)$, the better model k is relative to model ℓ. The inequality in (3) could hold for all $m \geq r$, in which case $\mathrm{GENO}(n,k,\ell)$ is infinity, indicating that model k with n observations is better than model ℓ with any number of observations.

We next discuss estimation. We estimate the quantities on the right and left-hand side in (3) and solve for m that satisfies equality rather than inequality. Therefore our estimate of GENO is generally not an integer; however, this is a natural way to go, rather than having to deal with integer parts. We do this without further mention in other definitions, such as that of $\widehat{T}$ in section 2.4.

Suppose we have N iid observations $Y_1, \ldots, Y_N$ from the true g. Our goal is to estimate $\mathrm{GENO}(n,k,\ell)$ for any n. In the standard AIC approach one takes $N = n$ in order to select the best model for the sample size at hand. Thus, the standard approach in the AIC literature might be to estimate only $\mathrm{GENO}(N,k,\ell)$, that is, GENO for the given sample size. Our approach is to compare models for all sample sizes, since $\mathrm{GENO}(n,k,\ell)$ for other values of n can be informative, telling us not only which model is better for N observations, but also providing the range of sample sizes for which this happens, and quantifying the relative quality of the models in a meaningful way, akin to Pitman efficiency.

We first estimate the quantity in (2) with N observations. A simple variation on standard AIC type large sample approximations, and calculations as in Claeskens and Hjort (2009) yield the estimator

$$AIC(n,k) = \frac{1}{N}\sum_{i=1}^{N} \log\left[f_k(Y_i, \widehat{\theta}_N^{(k)}\right] - d_k(\frac{1}{2n} + \frac{1}{2N}), \tag{4}$$

which coincides with the usual AIC when $N = n$ (if multiplied by the constant $-2N$, see e.g., Burnham and Anderson (2002) page 61). Applying this estimator to the quantities in (3) we obtain

$$\widehat{\mathrm{GENO}}(n,k,\ell) := \frac{d_\ell}{\frac{2}{N}\sum_{i=1}^{N} \log\left[f_\ell(Y_i, \widehat{\theta}_N^{(\ell)}) \Big/ f_k(Y_i, \widehat{\theta}_N^{(k)})\right] - \frac{d_\ell - d_k}{N} + \frac{d_k}{n}}, \tag{5}$$

where if the denominator is not positive, then $\widehat{\mathrm{GENO}}(n,k,\ell) = \infty$. When $d_\ell = 0$ we define $\widehat{\mathrm{GENO}}(n,k,\ell) = \infty, n$, or 0 when the denominator in (5) is negative, zero, or positive, respectively.

For each model k and sample size n, we define

$$\ell(n,k) = \arg\min_{\ell \in \{1,\ldots,K\}} \mathrm{GENO}(n,k,\ell) \quad \text{and} \quad \mathrm{GENO}(n,k) = \mathrm{GENO}(n,k,\ell(n,k)). \tag{6}$$

If the minimizer above is not unique then arg min above is a set, and any ℓ in it can be chosen. Similarly, we define $\widehat{\ell}(n,k) = \arg\min_{k\in\{1,\dots,K\}} \widehat{\mathrm{GENO}}(n,k,\ell)$ and $\widehat{\mathrm{GENO}}(n,k) = \widehat{\mathrm{GENO}}(n,k,\widehat{\ell}(n,k))$ to be the estimators. Thus,

$$\widehat{\mathrm{GENO}}(n,k) := \frac{d_{\widehat{\ell}(n,k)}}{\frac{2}{N}\sum_{i=1}^{N} \log\left[f_{k(n)}(Y_i, \widehat{\theta}_N^{(\widehat{\ell}(n,k))}) \Big/ f_k(Y_i, \widehat{\theta}_N^{(k)})\right] - \frac{d_{\widehat{\ell}(n,k)} - d_k}{N} + \frac{d_k}{n}}, \tag{7}$$

estimates the minimal number of observations required by the best competing model to achieve the same expected likelihood as model k with sample size n.

It is easy to calculate that

$$\widehat{\mathrm{GENO}}(n,k,\ell) = \left\{ \frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k)] + \frac{1}{n} \right\}^{-1}, \tag{8}$$

showing the relation between GENO and AIC differences. In Appendix A we show that (8) implies that $\widehat{\mathrm{GENO}}(n,k)$ agrees with the AIC ranking, that is,

Proposition 2.1. $\widehat{\mathrm{GENO}}(n,k) > \widehat{\mathrm{GENO}}(n,k') \quad \Leftrightarrow \quad AIC(n,k) > AIC(n,k')$.

2.2 GENO for many experiments with a common model

In the ERSB's experimental games data we analyzed, there is a sample of 180 players, each playing repeatedly one of 10 games all having the same nature, with varying payoffs chosen at random. When studying the way subjects play such games, ERSB assumed a common model with the same parameters for all players; however, since the games and players vary, we assume that different players may have different parameters. A similar remark applies to the DNA data. In this case, we model a large number of SNPs using the same models for all of them, allowing different parameters. We want to understand the quality of the HW model using GENO, assuming a common model for all SNPs. Deviations from HW are generally due to common reasons for all SNPs, such as recent migration, lack of random mating, and strong selection. Other examples may arise when one wants to construct a common regression model for different data sets, such as economic data from different countries, in order to compare the coefficients between different countries, when one wants to compare the influence of covariates on survival in different hospitals, etc.

We consider a collection of J experiments or data sets, possibly of different sizes N_j, that are to be analyzed together with a common model, allowing distinct parameters. Note that it is possible to compute $\widehat{\mathrm{GENO}}(n,k)$ by (7) on each experiment separately, and choose a model for each experiment, but here the emphasis is on choosing a common model for all of them.

Let $Y_{1,j},\dots,Y_{N_j,j}$ be N_j iid observations having density g_j in the jth experiment, $j = 1,\dots,J$. For the jth experiment, the density at the MLE of the kth model is $f_k(y,\widehat{\theta}_{N_j,j}^{(k)})$ and we assume a common k for all experiments. Similar to (2) the expected average log-likelihood for model k with n observations of new data is

$$E\sum_{j=1}^{J}\int_{-\infty}^{\infty} g_j(y)\log f_k(y,\widehat{\theta}_{n,j}^{(k)})dy. \tag{9}$$

We define in analogy to (3)

$$\mathrm{GENO}(n,k,\ell)=\left\{\sup_{m\geq r}\,:\,E\sum_{j=1}^{J}\int_{-\infty}^{\infty}g_j(y)\log f_\ell(y,\widehat{\theta}^{(\ell)}_{m,j})dy\leq E\sum_{j=1}^{J}\int_{-\infty}^{\infty}g_j(y)\log f_k(y,\widehat{\theta}^{(k)}_{n,j})dy\right\}. \tag{10}$$

The expectation (9) is estimated in analogy to (4) by

$$AIC(n,k)=\sum_{j=1}^{J}\frac{1}{N_j}\sum_{i=1}^{N_j}\log\left[f_k(Y_{i,j},\widehat{\theta}^{(k)}_{N_j,j})\right]-\sum_{j=1}^{J}d_k(\frac{1}{2n}+\frac{1}{2N_j}), \tag{11}$$

which leads as in (5) to

$$\widehat{\mathrm{GENO}}(n,k,\ell):=\frac{d_\ell}{\frac{1}{J}\sum_{j=1}^{J}\frac{2}{N_j}\sum_{i=1}^{N_j}\log\left[f_\ell(Y_{i,j},\widehat{\theta}^{(\ell)}_{N_j,j})\Big/f_k(Y_{i,j},\widehat{\theta}^{(k)}_{N_j,j})\right]-\frac{1}{J}\sum_{j=1}^{J}\frac{d_\ell-d_k}{N_j}+\frac{d_k}{n}}, \tag{12}$$

and in analogy to (7),

$$\widehat{\mathrm{GENO}}(n,k):=\frac{d_{\widehat{\ell}(n,k)}}{\frac{1}{J}\sum_{j=1}^{J}\frac{2}{N_j}\sum_{i=1}^{N_j}\log\left[f_{\widehat{\ell}(n,k)}(Y_{i,j},\widehat{\theta}^{(\widehat{\ell}(n,k))}_{N_j,j})\Big/f_k(Y_{i,j},\widehat{\theta}^{(k)}_{N_j,j})\right]-\frac{1}{J}\sum_{j=1}^{J}\frac{d_{\widehat{\ell}(n,k)}-d_k}{N_j}+\frac{d_k}{n}}. \tag{13}$$

2.3 A bootstrap confidence interval for GENO

We now discuss the construction of confidence intervals for GENO. It would possible to estimate the variance of the sum in the numerator of (7) using standard jackknife or bootstrap methods if the Y_i's were iid observations. One could then use the delta method to compute the variance of GENO, and use the asymptotic normality of the sum for a confidence interval. The same could be applied for each of the terms $Z_j:=\frac{2}{N_j}\sum_{i=1}^{N_j}\log\left[f_{\widehat{\ell}(n,k)}(Y_{i,j},\widehat{\theta}^{(\widehat{\ell}(n,k))}_{N_j,j})\Big/f_k(Y_{i,j},\widehat{\theta}^{(k)}_{N_j,j})\right]$ in (13) in order to construct a confidence interval under the assumption that $Y_{1,j},\ldots,Y_{N_j,j}$ are iid for each j.

In our motivating example one cannot assume that actions in repeated games are chosen independently. Therefore, we are interested in the case of many experiments for non-independent data, and we consider another approach. Assume that the different experiments are independent. Then the above Z_j's are independent but not identically distributed. The theory of bootstrap in this case was developed in Liu (1998), who showed that if the Z_j's have asymptotically a common mean then the distribution of $\frac{1}{J}\sum_{j=1}^{J}Z_j$ and its bootstrap distribution are asymptotically (in J) the same. Therefore, an asymptotic, $(1-\alpha)\%$ confidence interval for GENO(n,k), under suitable homogeneity assumptions, is

$$\left(\frac{d_{\widehat{\ell}(n,k)}}{\tilde{F}_J^{-1}(\alpha/2)-\frac{1}{J}\sum_{j=1}^{J}\frac{d_{\widehat{\ell}(n,k)}-d_k}{N_j}+\frac{d_k}{n}},\ \frac{d_{\widehat{\ell}(n,k)}}{\tilde{F}_J^{-1}(1-\alpha/2)-\frac{1}{J}\sum_{j=1}^{J}\frac{d_{\widehat{\ell}(n,k)}-d_k}{N_j}+\frac{d_k}{n}}\right), \tag{14}$$

where $\tilde{F}_J$ is the bootstrap distribution of $\frac{1}{J}\sum_{j=1}^{J}Z_j$. Such a confidence intervals informs us of the variability or potential range of values of $\widehat{\mathrm{GENO}}(n,k)$ around the observed value. This confidence interval neglects the variability that comes from the estimation of $\ell(n,k)$. A confidence interval for $\mathrm{GENO}(n,k,\ell)$ can be constructed in the same way.

This approach is tested in Section B.3. The results indicate that this method works well for large J, even if the means of the Z_j's are not exactly the same, and in this case the confidence intervals are slightly conservative. Indeed, a careful examination of the proof of Theorem 1 in Liu (1998) shows that when the J summands in the denominator of (13) do not have the same means, that is, the μ_i's are different (in Liu's notation), the confidence intervals are conservative.

2.4 A tie of two models

Let $\widehat{T} = \widehat{T}(k,\ell)$ (T for tie) denote that value of n such that $\widehat{\mathrm{GENO}}(n,k,\ell) = n$, that is, the sample size for which the two models are considered equally good. Assuming $d_\ell > d_k$ is it easy to obtain from (13) that

$$\widehat{T}(n,k,\ell) := \frac{d_\ell - d_k}{\frac{1}{J}\sum_{j=1}^{J}\frac{2}{N_j}\sum_{i=1}^{N_j}\log\left[f_\ell(Y_{i,j},\widehat{\theta}^{(\ell)}_{N_j,j})\Big/ f_k(Y_{i,j},\widehat{\theta}^{(k)}_{N_j,j})\right] - \frac{1}{J}\sum_{j=1}^{J}\frac{d_\ell - d_k}{N_j}}. \tag{15}$$

A bootstrap confidence interval for $\widehat{T}$ can be constructed as in (14), mutatis mutandis.

3 The multinomial distribution and Hardy-Weinberg model

Before getting to our motivating application in Sections 4 and 5, we provide a simpler application, which we think is potentially useful. In this section we discuss GENO for multinomial data and the Hardy-Weinberg model, which is of great importance in population biology. This example can easily be extended to any models for multinomial data, and to model selection using goodness of fit statistics.

Let $X_1,\dots,X_n$ be observations taking L possible values, say, $a_1,\dots,a_L$ with $P(X_i = a_\ell) = p_\ell$, and set $Y_\ell = \#\{i : X_i = a_\ell\}$, $\ell = 1,\dots,L$. We assume here that the true model g is multinomial with a given parameter $\mathbf{p} = (p_1,\dots,p_L)$ so that $Y = (Y_1,\dots,Y_L) \sim \mathrm{Multinomial}(n,\mathbf{p})$. In fact, the true model g is multinomial if the X_i's are independent and if the p_ℓ's are fixed throughout the experiment, an assumption that is often made, at least approximately. We compare different models $\mathbf{p} = \mathbf{p}(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$.

3.1 Hardy–Weinberg model

We focus on a classical model that plays a prominent role in genetics, the Hardy–Weinberg (HW) model. For a single diploid locus with two possible alleles A, G say, let (Y_1, Y_2, Y_3) denote the frequencies of the three genotypes, AA, AG, GG. Under the multinomial model the likelihood of (Y_1,Y_2,Y_3) is proportional to $\prod_{\ell=1}^{3} p_\ell^{Y_\ell}$. Denoting the probabilities of A and G by θ, $1-\theta$, the HW model specifies $\mathbf{p}_{HW}(\theta) = \Big(\theta^2, 2\theta(1-\theta), (1-\theta)^2\Big)$, and the MLE is $\widehat{\theta}_n^{(HW)} = \frac{2Y_1+Y_2}{2n}$. Higher-dimensional HW type models are discussed in Appendix B. GENO for HW models for SNPs in human DNA data are given in 3.3.

3.2 A numerical example

As a simple (artificial) first example, consider a specific true multinomial distribution and a HW model whose probabilities are presented in Table 1. The Kullback-Leibler projection of $\mathbf{p}$ on the HW model is $\mathbf{p}_{HW}(\theta_0)$, where θ_0 is computed similarly to the MLE with Y_ℓ/n replaced by p_ℓ of the true multinomial model, that is, $\theta_0 = 2p_1 + p_2$.

Table 1: Probabilities of the true and HW models.

Genotype	AA	AG	GG
True model $\mathbf{p}$	0.185	0.455	0.36
Nearest $\mathbf{p}_{HW}(\theta_0)$	0.1701	0.4847	0.3452

We consider only two models, the HW and the full multinomial models, referring to the latter as Full, and denote their probability functions by f_{HW} and f, respectively. Figure 2 presents the expected log-likelihood of (2) in this case. Here the true model is assumed known, and the expectations and GENO values are computed from the true model by simulation. For example, for $n = 200$ the expected likelihood is $E\left\{\log f_{HW}\left(Y^*, \widehat{\theta}_n^{(HW)}\right)\right\} = -1.043$. The value of m, for which $E\left\{\log f\left(Y^*, \widehat{\mathbf{p}}_m\right)\right\} = -1.043$ is 233, where $\widehat{\mathbf{p}}_m$ is the MLE of $\mathbf{p}$ under the multinomial model f with m observations, which is the vector of sample proportions. Therefore, $\text{GENO}(200, HW, Full) = 233$. Since for $n = 233$ the HW has the higher log-likelihood, then $\text{GENO}(233, Full) = 200$ and $\text{GENO}(233, HW) = 233$. Figure 2 also shows the expected log-likelihood under the two models as a function of n. When $n < 255$, the expected log-likelihood of HW is larger, and otherwise, that of the full multinomial model is larger. Therefore $T = 255$ (see Section 2.4). The HW model is better for samples smaller than 255, and with more data the full multinomial is better. An extensive simulation study of different HW models is given in Appendix B, where we investigate the performance of the estimates of GENO (13) and the confidence interval (14).

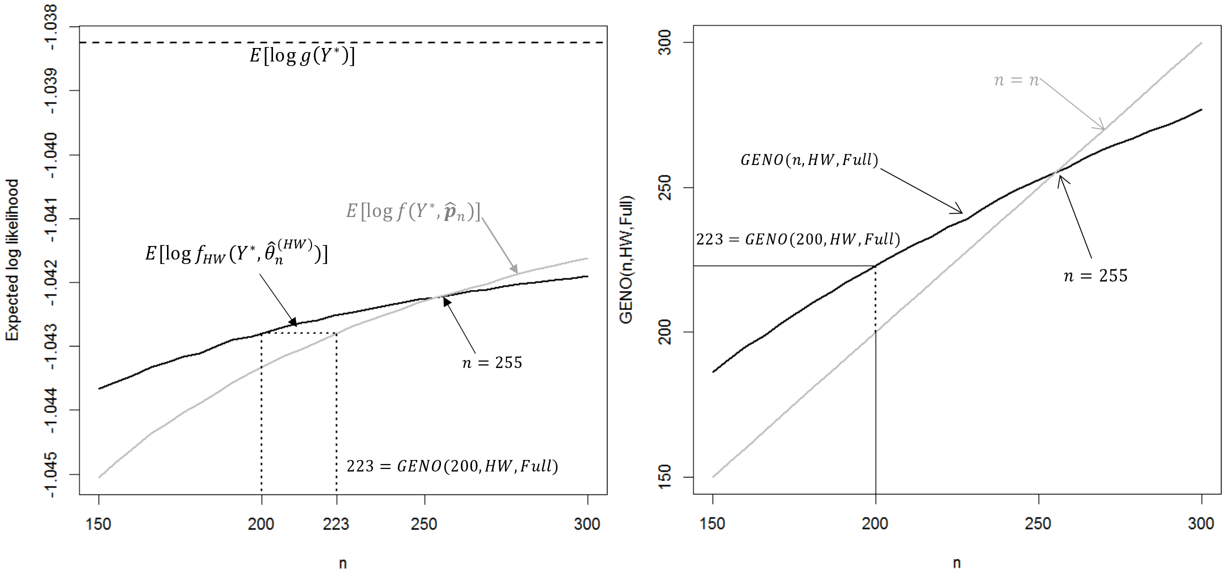


Figure 2: The computation of $\text{GENO}(n, HW, Full)$ based on the expected log-likelihoods $E\{\log f_{HW}(Y^*, \widehat{\theta}_n^{(HW)})\}$ and $E\{\log f(Y^*, \widehat{\mathbf{p}}_n)\}$.

3.3 Analysis of DNA

We now compute GENO of the HW model with a data set from the international HapMap project (Gibbs et al., 2003). We consider data on SNPs in two genetically distinct populations. In our context, a SNP is a diploid DNA site which can exhibit one of three versions, as in Section 3.1. In the HapMap data, SNPs that are far from HW equilibrium were excluded since a significant deviation from HW is typically attributed to errors. Given a sample of J SNPs, an estimate of GENO is

$$\widehat{\mathrm{GENO}}(n, HW, Full) = \frac{2}{\frac{1}{J}\sum_{j=1}^{J} 2\sum_{\ell=1}^{3} \widehat{p}_{j,\ell} \log\left[\widehat{p}_{j,\ell}/\{p_{HW}(\widehat{\theta}_j)\}_\ell\right] - \frac{1}{J}\sum_{j=1}^{J}\frac{1}{N_j} + \frac{1}{n}},$$

where for each SNP j, the sample size is N_j, the multinomial parameters are estimated by the jth sample proportions of the genotypes $(\widehat{p}_{j,1}, \widehat{p}_{j,2}, \widehat{p}_{j,3})$, and $\widehat{\theta}_j = 2\widehat{p}_{j,1} + \widehat{p}_{j,2}$.

The first data set we consider is a sample of 53 individuals, taken from a population with African ancestry in Southwest USA (ASW), and the second is a sample of 113 Utah residents with northern and western European ancestry (CEU). The ASW (CEW, respectively) data set contains information on about million (two million, respectively) SNPs where the minor allele frequency is at least 0.1. GENO for chromosome X was computed separately (see Table 6), however, for reasons explained below it is very different from the other chromosomes, and therefore it was excluded from the calculations of GENO for the two populations. We sampled $J = 56,694$ SNPs from ASE, and $J = 98,081$ SNPs from CEU, which are 5% of the total number of SNPs, at inter-SNP distances that allow us to consider the SNPs as independent. A plot of $\widehat{\mathrm{GENO}}(n, HW, Full)$ is given in Figure 3; a 95% bootstrap confidence intervals is also computed. Notice that although the sample size for each SNP is relatively small, we can infer GENO for large n's. This is due to the large number of SNPs whose information is used in estimating the combined GENO.

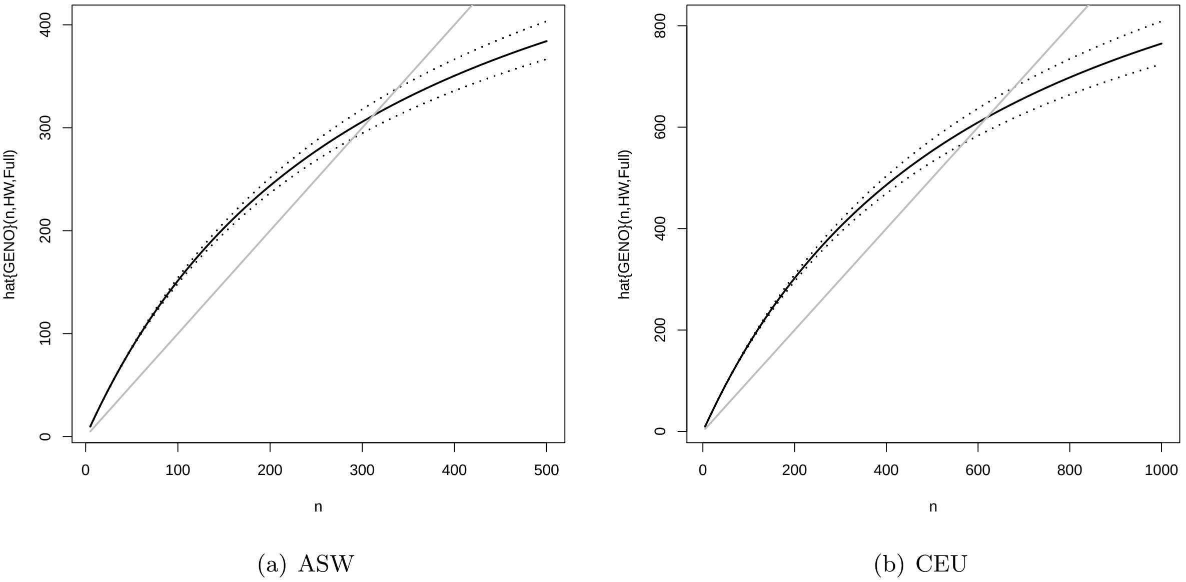
Let $\widehat{T} = \widehat{T}(HW, Full)$ defined in (15) denote the value of n for which the two models are equally good. Our calculations yield $\widehat{T}$=332.4 for ASW and $\widehat{T}$=591.4 for CEU; a 95% bootstrap confidence interval is (307.7,362.8) and (544.6,647.2), respectively. We calculated $\widehat{T}$ for each chromosome based on a sample of size $\mathcal{N}_c/20$, where $\mathcal{N}_c$ is the number of SNPs in chromosome c. We did not perform this calculation for chromosome Y, since there are only a few hundreds such SNPs. The results are given in Table 6 of Appendix C. Chromosome X is clearly different since the estimated $\widehat{T}$s are about 8 and 5 for the two populations, and for the other chromosomes it is a few hundreds.

While the HW model does not apply to Chromosome X, it can be used to explain its frequencies as follows. Males have a single chromosome X and outside of the pseudoautosomal region they are hemizygous; that is, in the above example, their genotype is either A or G but not AG. This requires a modification of HW for non-pseudoautosomal X chromosome SNPs (Hartwig, 2014). In order to assess the effect of this on GENO, consider a certain SNP where in women the proportion of the three genotypes follow HW $\big(\theta^2, 2\theta(1-\theta), (1-\theta)^2\big)$ and in men the proportion is $(\theta, 0, 1-\theta)$. Assuming that half of the population are women, the proportion in the population is

$$\mathbf{p} := \big((\theta^2+\theta)/2, \theta(1-\theta), \{(1-\theta)^2 + 1 - \theta\}/2\big)\,.$$

With $\mathbf{p}$ as above, we have $\theta_0 = 2p_1 + p_2 = \theta$ and $\mathbf{p}_{HW}(\theta_0) = \big(\theta^2, 2\theta(1-\theta), (1-\theta)^2\big)$. In this case, by (15), $\widehat{T} = 1/2\sum_{\ell=1}^{3} \widehat{p}_\ell \log\left[\widehat{p}_\ell/\{p_{HW}(\widehat{\theta})\}_\ell\right]$. Here we consider $\widehat{\theta} \in [0.1, 0.9]$ for which

$\widehat{T}$ varies between 4 and 6.5, close to the estimated $\widehat{T}$ of chromosome X in the data. All other chromosomes, that is, the autosomal chromosomes, the Hardy-Weinberg is a good model and the full multinomial is better only when the sample size consists of more than several hundreds individuals. The fact that $\widehat{T}$ is consistently larger for the CEU population (Table 6) means that the HW models fits this population better, and may be explained by the fact that this population was less subject to recent migration, and may be more homogeneous than the ASW population.



(a) ASW

(b) CEU

Figure 3: Plots of $\widehat{\mathrm{GENO}}(n, HW, Full)$ for ASW and CEU. A 95% bootstrap confidence interval is plotted in the dotted lines. The line $n = n$ is also drawn.

4 GENO for decision processes

The initial impetus for this paper comes from work on data analysis in experimental game theory, and from Erev, Roth, Slonim, and Barron (2007), whose data, to be described in detail in Section 5, is reanalyzed below. It involves actions in repeated games that are not independent, and we next adapt GENO to this situation. We start with more general decision processes, and later specialize to our motivating problem.

4.1 The decision process set-up

For a given decision process, or game for short, let $Z_1, \dots, Z_n$ be actions taken at times $1, \dots, n$ with values in some finite space $\mathcal{Z}$, the decision space, and let $V_1, \dots, V_n$ denote the corresponding rewards. At stage t of the process the decision maker (player) bases the current decision on the information in

$$\mathcal{D}_{t-1} := (Z_1, \dots, Z_{t-1}, V_1, \dots, V_{t-1}). \tag{16}$$

Decision are determined by a mixed strategy having probability $p_{\mathcal{D}_{t-1}}(z_t) = P(Z_t = z_t \mid \mathcal{D}_{t-1})$ of making the decision z_t at time t on the basis of $\mathcal{D}_{t-1}$ for $t = 1, 2, \ldots$. This decision model, where actions are based on past actions and rewards, is often called 'reinforcement learning'. We assume that the (random) reward V_t at time t depends only on the action Z_t, through a known conditional probability (or density) function $p(v_t \mid z_t)$ that depends on the game. Under these assumptions, a simple calculation shows that the likelihood of the player's sequence of actions $z_1, \ldots, z_n$ and rewards $v_1, \ldots, v_n$ is

$$L(z_1, \ldots, z_n; v_1, \ldots, v_n) = \prod_{t=1}^{n} [p(v_t \mid z_t) p_{\mathcal{D}_{t-1}}(z_t)].$$

As $\prod_{t=1}^{n} p(v_t \mid z_t)$ does not depend on the decision model and its parameters, it can be regarded as a constant and ignored. Therefore we now write the likelihood as

$$L(z_1, \ldots, z_n; v_1, \ldots, v_n) = \prod_{t=1}^{n} p_{\mathcal{D}_{t-1}}(z_t). \tag{17}$$

A given player has a true strategy g, that is, $p_{\mathcal{D}_{t-1}}(z_t) = g(z_t \mid \mathcal{D}_{t-1})$. Having experimental data, the goal is to approximate this unknown strategy by different models, and we use GENO to quantify the predictive value of such models.

Consider K candidate models for such a mixed strategy, where for $k = 1, \ldots, K$, $p_{\mathcal{D}_{t-1}}(z_t)$ is modeled by a function $f_{k,t}(z_t \mid \mathcal{D}_{t-1}, \theta^{(k)})$, with $\theta^{(k)} \in \Theta^{(k)} \subseteq \mathbb{R}^{d_k}$. As usual, we do not assume that players really play according to any of these models, but we do assume that with suitable values of the parameters for different players, these models can be useful for analysis and prediction of players' behavior. We shall focus on stationary Markov decision processes or games, where the distribution of the action of Z_t at time t depends only on S_{t-1}, the state of the process at time $t-1$, and not on the time t, and therefore $f_{k,t}$ will be replaced by f_k. If, for example, the kth model assumes that the decision is based on the last M_k actions and rewards, then $S_{t-1}^{(k)} = (Z_{t-M_k}, \ldots, Z_{t-1}, V_{t-M_k}, \ldots, V_{t-1})$. For ERSB's learning models discussed below we have $S_{t-1} = (Q_{t-1}, V_{t-1}, Z_{t-1})$, where the so-called propensity Q_t is updated according to a formula of the type $Q_t = \kappa(Q_{t-1}, V_{t-1}, Z_{t-1})$ for a suitable function κ to be discussed in Section 5.1 equation (22), and the distribution of Z_t at time t depends only on Q_t. Such models are Markov with respect to the state space as defined. We assume that the true strategy g and all our candidate models are M-step Markov decision processes for some $M \geq 1$. In particular, under the kth model the mixed strategy is given by $f_k(z_t \mid S_{t-1}^{(k)}, \theta^{(k)})$.

Under the Markov assumption on the true process, the process $(Z_t, S_{t-1}^{(k)})$ possesses a stationary distribution under well-known ergodicity conditions that are assumed. We denote this stationary distribution by $q_k(z, s)$, with s in a suitable space where $S_{t-1}^{(k)}$ takes values. Integrals ds should be interpreted with s in the latter space.

By (17), the log-likelihood under the kth model for a given player becomes

$$\sum_{t=1}^{n} \log f_k(Z_t \mid S_{t-1}^{(k)}, \theta^{(k)}). \tag{18}$$

Let $\widehat{\theta}_n^{(k)}$ be the MLE, i.e., the maximizer of (18), based on n observations, and let the projection parameter $\theta_0^{(k)}$ is defined as

$$\theta_0^{(k)} := \arg \max_{\theta \in \Theta^{(k)}} \sum_{z \in \mathcal{Z}} \int q_k(z, s) \log f_k\left(z \mid s, \theta^{(k)}\right) ds. \tag{19}$$

Extending the AIC type expansions requires that $\sqrt{n}(\widehat{\theta}_n^{(k)} - \theta_0^{(k)})$ converges to normal in distribution. This holds for finite Markov chains under simple ergodicity conditions. Results of this type for (more general) Markov chains can be found in Billingsley (1961a,b), and Roussas (1968). There is a large body of literature on related results for more general stationary ergodic processes which is beyond the scope of this paper.

In the spirit of GENO described above, imagine that a new player is playing the same game n^* times, with the same strategy g; let $(Z_t^*, S_t^{*(k)})$ be the new decision, and state according to the kth model at time t for $t = 1, \dots, n^*$. As before, we want to consider the expected log-likelihood of the new hypothetical data under the model, if the MLE $\widehat{\theta}_n^{(k)}$ is based on the given data with sample size n. The expected log-likelihood is

$$\frac{1}{n^*} E \sum_{t=1}^{n^*} \log f_k(Z_t^* \mid S_{t-1}^{*(k)}, \widehat{\theta}_n^{(k)}) = E \sum_{z \in \mathcal{Z}} \int q_k(z,s) \log f_k \left(z \mid s, \hat{\theta}_n^{(k)} \right) ds,$$

where the expectation on the left is with respect to all the starred variables under the true model g and with respect to the MLE $\widehat{\theta}_n^{(k)}$, and the expectation on the right is only with respect to the latter.

4.2 GENO for decision processes

Similar to the iid case, we define

$$\begin{aligned} \mathrm{GENO}(n,k,\ell) := \Big\{ \max_{m \geq r} \ : \ & \sum_{z \in \mathcal{Z}} E \int q(z,s) \log f_\ell \left(z \mid s, \widehat{\theta}_m^{(\ell)} \right) ds \\ & \leq \sum_{z \in \mathcal{Z}} E \int q_k(z,s) \log f_k \left(z \mid s, \widehat{\theta}_n^{(k)} \right) ds \Big\}, \end{aligned} \tag{20}$$

where r is the minimum number of observations required for $\widehat{\theta}_m^{(\ell)}$ to exist. As in the iid case, AIC type approximations and some calculations for Markov chains, as in Ogata (1980) and Tong (1975), lead to estimation of the quantities in (20), and in analogy with (5) we obtain the estimator

$$\widehat{\mathrm{GENO}}(n,k,\ell) := \frac{d_\ell}{\frac{2}{N} \sum_{t=1}^N \log \left\{ f_\ell(Z_t | S_{t-1}^{(\ell)}, \widehat{\theta}_N^{(\ell)}) \Big/ f_k(Z_t | S_{t-1}^{(k)}, \widehat{\theta}_N^{(k)}) \right\} - \frac{d_\ell - d_k}{N} + \frac{d_k}{n}}. \tag{21}$$

The above represents GENO for a single player. When the data come from J players, this is generalized as in Section 2.2, equations (12) where a sum over the J experiment is added. The notion of GENO(n,k) and its estimator are defined as in (6) and (13). We will not repeat the details.

5 Game theory experiments: analysis of the motivating data

We now apply our approach to part of the data of ERSB. In their experiment 180 subjects are arranged in 90 fixed pairs. There are 10 different games, and every game is played by 9 of these pairs, 500 times each. In every two-player game, each player chooses one action out of two; these choices determine the probabilities of winning a fixed amount or zero. The winning probabilities of the two players for each profile of actions add up to one, so in expectation this is a fixed-sum game.

5.1 The models

Following ERSB we consider models of strategies having the same parametric form for all players; however, we allow different values of parameters for different players. Therefore, it is enough to describe the modeled strategy for a single player. The rewards depend on both players in a pair, however, from the point of view of the first player, his reward is a random function of his action as in Section 4.1.

For the first three models, $k = 1, 2, 3$, we have $Q_t = \Big(Q_t(0), Q_t(1)\Big)$, where $Q_t(i)$ is referred to as the propensity to select action i. The propensities are updated at each stage according to

$$Q_t(i) = \left\{ \begin{array}{cl} (1-\alpha)Q_{t-1}(i) + \alpha V_{t-1} & \text{if } \ Z_{t-1} = i \\ Q_{t-1}(i) & \text{if } \ Z_{t-1} \neq i \end{array} \right. , \quad i = 0, 1, \tag{22}$$

where $0 < \alpha < 1$ is a parameter of the model. Initially $Q_1(0)$ and $Q_1(1)$ are equal to the player's expected payoff when both players choose each strategy with equal probability. The following models are considered:

1. Reinforcement learning (RL): action 1 at round t is chosen with probability

$$p_{\mathcal{D}_{t-1}}(1) = \frac{Q_t(1)}{Q_t(1) + Q_t(0)}.$$

2. Reinforcement learning lambda (RLL): action 1 at round t is chosen with probability

$$p_{\mathcal{D}_{t-1}}(1) = \frac{\lambda + Q_t(1)}{2\lambda + Q_t(1) + Q_t(0)},$$

   where $\lambda > 0$ is an unknown parameter. When λ is large, $p_{\mathcal{D}_{t-1}}(1) \approx \frac{1}{2}$, and the propensities are weighted down.

3. Reinforcement learning stickiness (RLS): action 1 at round t is chosen with probability

$$p_{\mathcal{D}_{t-1}}(1) = (1-\xi)\frac{Q_t(1)}{Q_t(1) + Q_t(0)} + \xi Z_{t-1},$$

   $0 < \xi < 1$ is a "stickiness" parameter; when ξ is close to 1 the player repeats his choice with high probability.

4. Toss: at each round, action 1 is chosen with probability p independently of previous rounds.

5. Nash: at each round, action 1 is chosen with probability predicted by Nash equilibrium. This model has no free parameters.

6. M-step Markov: the probability of choosing action 1 at stage t is based on the last M actions, i.e., $Z_{t-M}, \ldots, Z_{t-1}$ and the last reward V_{t-1}. There are 2^{M+1} possible sequences of M past decisions and the last reward. Therefore the model has 2^{M+1} parameters for each player, consisting of the probability of choosing action 1 for each such sequence. We consider $M = 1, 2, 3$ and denote them by 1-M, 2-M, 3-M. We also consider a two-actions two-rewards (denoted by 2a2r) model, with action at stage t is based on Z_{t-2}, Z_{t-1} and V_{t-2}, V_{t-1}.

7. CAB-WM For $t > K$, let $x_t := (Z_{t-K}, \ldots, Z_{t-1}, V_{t-K}, \ldots, V_{t-1})$. For each $x \in \{0,1\}^{2^K}$ and $t \leq K$ define $N_1(x;t) = 0, N_0(x;t) = 0$ and for $t > K$ define recursively
$$N_{Z_t}(x;t) = \begin{cases} N_{Z_t}(x;t-1) + V_t & x = x_t \\ N_{Z_t}(x;t-1) & x \neq x_t \end{cases}, N_{1-Z_t}(x;t) = \begin{cases} N_{1-Z_t}(x;t-1) + \alpha(1 - V_t) & x = x_t \\ N_{1-Z_t}(x;t-1) & x \neq x_t \end{cases}.$$
where $\alpha \geq 0$ is a parameter. For small α, the strategy puts more weight on a gain than on a loss. Set
$$p_{\mathcal{D}_{t-1}}(1) = \frac{N_1(x_t, t-1)^\beta + 1/2}{N_1(x_t, t-1)^\beta + N_0(x_t, t-1)^\beta + 1}.$$
for $\beta \geq 0$, a parameter. When β is large then $p_{\mathcal{D}_{t-1}}(1)$ is close to 1 or 0, depending whether $N_1(x_t, t-1)$ is larger than $N_0(x_t, t-1)$ or not.

8. CAB-W is CAB-WM with $\beta = 1$, and CAB-M is CAB-WM with $\alpha = 1$. CAB-K is CAB-WM with $\alpha = \beta = 1$. We consider only the case K=1.

Models 1–3 are variations on the reinforcement model (Erev and Roth, 1998), 4 and 5 are standard. Models 3 and 6 have not been studied previously in this context, to the best of our knowledge. The MLE in models 1–3 and 7–8 is computed by numerical maximization of the log-likelihood, and estimation in models 4 and 6 is straightforward. The CAB models are variations on a the model CAB-K of Plonsky, Teodorescu and Erev (2015). This model decides by a majority rule without randomization, and works very well on the data in the latter article. When applying a likelihood criterion as we do, non-randomized strategies are ruled out because a single deviation in the data from the deterministic rule makes the likelihood vanish. This may happen to good models, with data that deviate only rarely from the model's strategy. Our version is randomized. Note that a large β in these models make them closer to being deterministic. Our estimates in the ERSB data show that most α's are close to 0, and almost all are between 0 and 0.4, showing that gains weigh more than losses in the decision. The estimates of β, when we set $\alpha = 0$, are close to 1 (80% are between 0.5 and 1.5), indicating that players usually do not decide in a deterministic fashion.

We first computed AIC(N_j, k) for each of the $J = 90$ pairs and the models, as defined in (11), where $N_j = 500$, the number of games played by each pair. Under the proposed models, considering a pair of players as a single player, with $J = 90$, or as two players, with $J = 180$ amounts to the same AIC and GENO. The averages and standard deviations are given in Table 2. These numbers should be adjusted by the common additive (negative) value which was neglected as in (17).

Table 2 shows AIC(n, k) for different models where the CAB model have K=1, since larger values of K did not yield improved models. CAB-W is the best when $n < 55$; for $56 \leq n \leq 61$ CAB-WM is preferred and for larger n's the Markov models have the largest values of AIC and therefore are preferred according to the AIC criterion for the given sample sizes. For n smaller than 162, 1-step Markov is the best model and for larger n, smaller than 511, 2-step Markov is preferred. For larger n, 3-step Markov is the best among our candidate models. It is interesting to note that 3-step Markov has a somewhat larger AIC value than 2-actions 2-rewards for all n, and the same number of parameters, so 3-M will be preferred over 2a2r. This suggests that the third previous action is somewhat more informative (in AIC sense) than the second previous reward.

The computations we performed differ from a recent similar calculation in Marchiori and Warglien (2008) in several ways: we use the MLE estimates for each model, rather than first

Table 2: Average (SD) over the 90 pairs of AIC(n, k) for different n's and k's. Models with the largest AIC(n, k) are in bold face.

Model k	AIC$(50, k)$	AIC$(125, k)$	AIC$(250, k)$
RL	-1.25 (0.136)	-1.238 (0.136)	-1.235 (0.136)
RLL	-1.218 (0.147)	-1.194 (0.147)	-1.188 (0.147)
RLS	-1.051 (0.268)	-1.027 (0.268)	-1.021 (0.268)
Toss	-1.156 (0.243)	-1.144 (0.243)	-1.141 (0.243)
Nash	-1.443 (0.369)	-1.443 (0.369)	-1.443 (0.369)
1-M	-1.007 (0.253)	**-0.959 (0.253)**	-0.947 (0.253)
2-M	-1.063 (0.253)	-0.967 (0.253)	**-0.943 (0.253)**
3-M	-1.207 (0.249)	-1.015 (0.249)	-0.967 (0.249)
2a2r	-1.209 (0.248)	-1.017 (0.248)	-0.969 (0.248)
CAB-1	-1.359 (0.143)	-1.359 (0.143)	-1.359 (0.143)
CAB-M	-1.221 (0.193)	-1.209 (0.193)	-1.206 (0.193)
CAB-W	**-0.998 (0.249)**	-0.986 (0.249)	-0.983 (0.249)
CAB-WM	-1 (0.25)	-0.976 (0.25)	-0.97 (0.25)

Model k	AIC$(300, k)$	AIC$(500, k)$	AIC$(700, k)$
RL	-1.233 (0.136)	-1.232 (0.136)	-1.231 (0.136)
RLL	-1.185 (0.147)	-1.182 (0.147)	-1.181 (0.147)
RLS	-1.017 (0.268)	-1.015 (0.268)	-1.013 (0.268)
Toss	-1.139 (0.243)	-1.138 (0.243)	-1.137 (0.243)
Nash	-1.443 (0.369)	-1.443 (0.369)	-1.443 (0.369)
1-M	-0.941 (0.253)	-0.935 (0.253)	-0.933 (0.253)
2-M	**-0.929 (0.253)**	**-0.919 (0.253)**	-0.914 (0.253)
3-M	-0.94 (0.249)	**-0.919 (0.249)**	**-0.91 (0.249)**
2a2r	-0.943 (0.248)	-0.921 (0.248)	-0.912 (0.248)
CAB-1	-1.359 (0.143)	-1.359 (0.143)	-1.359 (0.143)
CAB-M	-1.204 (0.193)	-1.203 (0.193)	-1.202 (0.193)
CAB-W	-0.981 (0.249)	-0.98 (0.249)	-0.979 (0.249)
CAB-WM	-0.967 (0.25)	-0.964 (0.25)	-0.963 (0.25)

moments which in the presence of dependence are not sufficient statistics, we consider the likelihood function itself and not just the prediction of the model on the average choice, and unlike Marchiori and Warglien (2008) and ERSB, we do not assume that all players have common parameters. We found that allowing individual parameters leads to smaller AIC numbers and therefore are preferred. For example, if we consider a common parameter for all players the average AIC$(500, k)$, where k is the RL model, is -1.312, whereas the corresponding number when individual parameters are allowed is -1.232.

5.2 GENO: results

Tables 3 and Figure 4 show $\widehat{\text{GENO}}(n, k)$ for different n's and for the models mentioned in the previous section along with confidence intervals at the 95% level based on (14). The models' $\hat{\ell}(n, k)$ is also given. The model in boldface is the best for the given n. For example, for $n = 200$, the 2-M model is best, $\widehat{\text{GENO}}$(200,1-M)=179 and $\widehat{\ell}(n$,1-M)=2-M. This means that using 1-M rather than 2-M with $n = 200$, amounts to a loss of $200 - 179$, or about 20 observations. Having quantified the loss, the user can now decide between 1-M and 2-M according to considerations such as simplicity of the model, or prior preferences. The model CAB-W is best for $n = 50$ observations. However when $n = 200$ its performance is comparable to 1-M with only 72 observations, meaning that for $n = 200$ one can do much better than CAB-W which incurs a loss of about 130 observations.

Since players within a pair cannot be considered independent as requited for the Bootstrap confidence intervals, we considered each pair of players as a single player, making a pair of decisions simultaneously, when constructing the confidence intervals. For $n = 500$, the sample size of each experiment in the data, the best model is 2-step Markov, and 3-step Markov is a close second, and becomes best for n larger than approximately 510.

By our criterion, learning models 1-3 do not perform well. One should keep in mind that ERSB used them for a different goal: predicting proportions of actions played, and not for prediction of the process or the whole likelihood. Nash's model has no free parameters that need to be estimated and, therefore, its GENO does not depend on n and is 0 since CAB-1 with no parameters is better. Thus, we find that the learning models are more useful than the Nash model, as did ERSB, using their measure ENO. The GENO of the Markov models are higher than the learning models and therefore, by our measure, the Markov models are more useful.

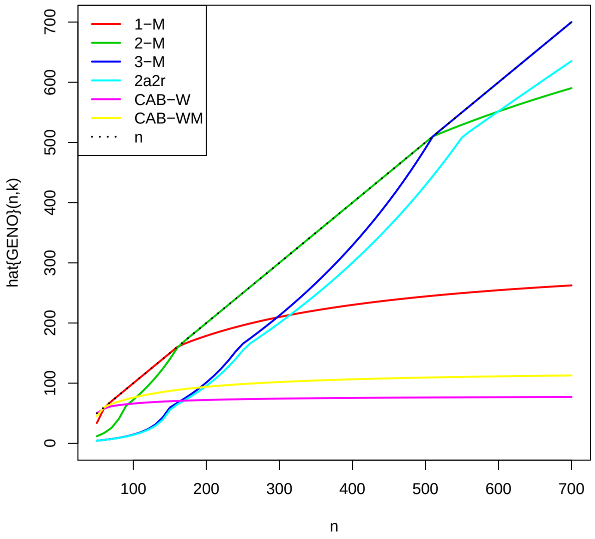


Figure 4: Plot of $\widehat{\text{GENO}}(n, k)$, where k is one of the models mentioned in Section 5.1 (only models with large GENO are plotted) and $n = 50, \ldots, 700$.

Table 3: Estimates (95% confidence intervals) of GENO(n, k) and $\hat{\ell}(n, k)$ for different n's and for chosen model k's. The best models are in boldface.

Model k	$\widehat{\text{GENO}}(50, k)$	$\hat{\ell}(50, k)$	$\widehat{\text{GENO}}(125, k)$	$\hat{\ell}(125, k)$	$\widehat{\text{GENO}}(200, k)$	$\hat{\ell}(200, k)$
RL	4 (3 , 4)	CAB-W	4 (3 , 5)	CAB-W	4 (3 , 5)	CAB-W
RLL	4 (4 , 5)	CAB-W	5 (4 , 6)	CAB-W	5 (4 , 6)	CAB-W
RLS	14 (11 , 18)	CAB-W	21 (15 , 31)	CAB-W	23 (17 , 39)	CAB-W
Toss	6 (5 , 7)	CAB-W	6 (5 , 8)	CAB-W	6 (5 , 8)	CAB-W
Nash	0 (0 , 0)	CAB-1	0 (0 , 0)	CAB-1	0 (0 , 0)	CAB-1
1-M	34 (26 , 51)	CAB-W	**125 (125 , 125)**	1-M	179 (156 , 205)	2-M
2-M	12 (11 , 13)	CAB-W	102 (90 , 122)	1-M	**200 (200 , 200)**	2-M
3-M	4 (4 , 5)	CAB-W	27 (21 , 38)	CAB-W	101 (84 , 130)	1-M
2a2r	4 (4 , 5)	CAB-W	25 (21 , 34)	CAB-W	96 (83 , 115)	1-M
CAB-1	3 (2 , 3)	CAB-W	3 (2 , 3)	CAB-W	3 (2 , 3)	CAB-W
CAB-M	4 (4 , 5)	CAB-W	4 (4 , 5)	CAB-W	4 (4 , 5)	CAB-W
CAB-W	**50 (50 , 50)**	CAB-W	68 (59 , 80)	1-M	72 (61 , 85)	1-M
CAB-WM	45 (35 , 66)	CAB-W	82 (71 , 96)	1-M	94 (80 , 112)	1-M

Model k	$\widehat{\text{GENO}}(300, k)$	$\hat{\ell}(300, k)$	$\widehat{\text{GENO}}(500, k)$	$\hat{\ell}(500, k)$	$\widehat{\text{GENO}}(700, k)$	$\hat{\ell}(700, k)$
RL	4 (3 , 5)	CAB-W	4 (3 , 5)	CAB-W	4 (3 , 5)	CAB-W
RLL	5 (4 , 6)	CAB-W	5 (4 , 6)	CAB-W	5 (4 , 6)	CAB-W
RLS	25 (18 , 44)	CAB-W	27 (18 , 49)	CAB-W	28 (19 , 53)	CAB-W
Toss	6 (5 , 8)	CAB-W	6 (5 , 8)	CAB-W	6 (5 , 8)	CAB-W
Nash	0 (0 , 0)	CAB-1	0 (0 , 0)	CAB-1	0 (0 , 0)	CAB-1
1-M	210 (180 , 245)	2-M	244 (204 , 296)	2-M	262 (216 , 323)	2-M
2-M	**300 (300 , 300)**	2-M	**500 (500 , 500)**	2-M	590 (499 , 696)	3-M
3-M	212 (191 , 244)	2-M	490 (391 , 710)	2-M	**700 (700 , 700)**	3-M
2a2r	200 (187 , 216)	2-M	429 (372 , 509)	2-M	635 (520 , 798)	3-M
CAB-1	3 (2 , 3)	CAB-W	3 (2 , 3)	CAB-W	3 (2 , 3)	CAB-W
CAB-M	4 (4 , 5)	CAB-W	4 (4 , 5)	CAB-W	4 (4 , 5)	CAB-W
CAB-W	74 (63 , 88)	1-M	76 (64 , 91)	1-M	77 (65 , 92)	1-M
CAB-WM	102 (86 , 123)	1-M	109 (90 , 134)	1-M	113 (93 , 140)	1-M

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723.

Akaike, H. (1983). Information measures and model selection. *International Statistical Institute* **44** 277–291.

Billingsley, P. (1961a). *Statistical Inference for Martov Processes.* The University of Chicago Press, Chicago.

Billingsley, P. (1961b). Statistical methods in Markov chains. *Annals of Mathematical Statistics* **32** 12–40.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference.* New York: Springer.

Claeskens, G. and Hjort, N. L. (2009). *Model Selection and Model Averaging.* Cambridge University Press: Cambridge.

Erev, I. and Roth, A.E. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed Strategy equilibria. *The American Economic Review*, **88** 848–881.

Erev, I., Roth, A.E., Slonim, R.L., Barron, G. (2007). Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory* **33** 29–51.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., ... & Zhang, H. (2003). The international HapMap project. *Nature*, **426**, 789–796.

Hartwig, F.P. (2014) Considerations to Calculate Expected Genotypic Frequencies and Formal Statistical Testing of Hardy-Weinberg Assumptions for non-pseudoautosomal X chromosome SNPs. *Genetic Syndromes and Gene Therapy*, **5** : 231.

Liu,R.Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, **16**, 1696–1708.

Marchiori, D. and Warglien M. (2008). Predicting human interactive learning by regret-driven neural networks. *Science* **319** 1111–1113.

Ogata, Y. (1980). Maximum Likelihood Estimates of Incorrect Markov Models for Time Series and the Derivation of AIC . *Journal of Applied Probability* **17** 59–72.

Plonsky, D., Teodorescu, K. and Erev, I. (2015). Reliance on Small Samples, the Wavy Recency Effect, and Similarity-Based Learning. *Psychological Review* **4** 621–647.

Roussas, G. G. (1968). Asymptotic normality of the maximum likelihood estimate in Markov processes. *Metrika* **14** 62–70.

Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability* **12** 488–497.

van der Vaart A.W. (1998). *Asymptotic Statistics.* New York: Cambridge University Press

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50** 1–25.

Zacks, S. (1985). Pitman efficiency. *In Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson (eds). New York: Wiley & Sons.

Appendices

A Proof of Proposition 2.1

We show that

$$\widehat{\text{GENO}}(n,k) > \widehat{\text{GENO}}(n,k') \quad \Leftrightarrow \quad AIC(n,k) > AIC(n,k').$$

Proof: If $AIC(n,k) > AIC(n,k')$ then for all $\ell \in \{1,\ldots,K\}$

$$\frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k)] < \frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k')]$$

and hence using (8) we obtain

$$\begin{aligned}\widehat{\text{GENO}}(n,k) = &\min_{\ell\in\{1,\ldots,K\}} \left\{\frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k)] + \frac{1}{n}\right\}^{-1} \\ &> \min_{\ell\in\{1,\ldots,K\}} \left\{\frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k')] + \frac{1}{n}\right\}^{-1} = \widehat{\text{GENO}}(n,k').\end{aligned}$$

Conversely, if $\widehat{\text{GENO}}(n,k) > \widehat{\text{GENO}}(n,k')$ then,

$$\max_{\ell\in\{1,\ldots,K\}} \frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k)] < \max_{\ell\in\{1,\ldots,K\}} \frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k')].$$

The maximum of the right hand-side is obtained at $\hat{\ell}(n,k')$. Then,

$$\max_{\ell\in\{1,\ldots,K\}} \frac{1}{2d_\ell}[AIC(n,\ell) - AIC(n,k)] < \frac{1}{2d_{\hat{\ell}(n,k')}}[AIC(n,\hat{\ell}(n,k')) - AIC(n,k')].$$

The last inequality holds for all $\ell \in \{1,\ldots,K\}$ and in particular for $\hat{\ell}(n,k')$. Hence,

$$\frac{1}{2d_{\hat{\ell}(n,k')}}[AIC(n,\hat{\ell}(n,k')) - AIC(n,k)] < \frac{1}{2d_{\hat{\ell}(n,k')}}[AIC(n,\hat{\ell}(n,k')) - AIC(n,k')],$$

implying $AIC(n,k) > AIC(n,k')$. □

B A numerical study of GENO for extended Hardy-Weinberg models

In this section we present an example of a more complex HW type model, and explain the computations and estimation for a particular case in Sections B.1 - B.2. In Section B.3 we demonstrate the results by a simulation study.

Consider Hardy-Weinberg models for a single diploid locus with three (rather than two as in previous sections) possible alleles a,b,c, appearing with probabilities θ_1, θ_2, $\theta_3 = 1 - (\theta_1 + \theta_2)$, respectively. This leads to a two-dimensional model. Alternatively, we consider a one-dimensional model with $\theta_1 = \theta_2 = \eta$. The model with η is called Model 1 and the bigger model with $\theta = (\theta_1, \theta_2)$ is called Model 2. The probabilities of the different genotypes, according to Models 1 and 2, are presented in Table 4.

Table 4: The probabilities according to the model.

Genotype	aa	ab	bb	bc	ac	cc
Probability - Model 1	η^2	$2\eta^2$	η^2	$2\eta(1-2\eta)$	$2\eta(1-2\eta)$	$(1-2\eta)^2$
Probability - Model 2	θ_1^2	$2\theta_1\theta_2$	θ_2^2	$2\theta_2\theta_3$	$2\theta_1\theta_3$	θ_3^2

We have

$$\widehat{\theta}_1 = \frac{2Y_1 + Y_2 + Y_5}{2n}, \ \widehat{\theta}_2 = \frac{Y_2 + 2Y_3 + Y_4}{2n}; \ \widehat{\eta} = \frac{2(Y_1 + Y_2 + Y_3) + Y_4 + Y_5}{4n}. \tag{23}$$

The likelihood of $(Y_1, \ldots, Y_6)$ is proportional to $\prod_{\ell=1}^{6} p_\ell^{Y_\ell}$. The p_ℓ's are the components of $\mathbf{p}^{(1)} = \mathbf{p}^{(1)}(\eta) = \left(\eta^2, 2\eta^2, \eta^2, 2\eta(1-2\eta), 2\eta(1-2\eta), (1-2\eta)^2\right)$ or $\mathbf{p}^{(2)} = \mathbf{p}^{(2)}(\theta)$ $= \left(\theta_1^2, 2\theta_1\theta_2, \theta_2^2, 2\theta_2\theta_3, 2\theta_1\theta_3, \theta_3^2\right)$ under Model 1 or Model 2, respectively.

B.1 A numerical example

We consider a specific multinomial distribution. The probability vector $\mathbf{p}$ under the full multinomial model with 6 cells (denoted by Full, with dimension = 5), and the resulting probabilities under Models i are presented in Table 5, where η_0 and θ_0 minimize the Kullback-Leibler divergence $D(Full||Model\ i)$, $i = 1, 2$. Here $\mathbf{p}$ was chosen only to provide a numerical example in which the two candidate models are reasonable but not close to perfect, so that the resulting GENOs are not trivial.

Table 5: The probabilities under Models 1 and 2 and under the full model.

	1	2	3	4	5	6
Full model $\mathbf{p}$	0.0700	0.2120	0.0824	0.2632	0.2080	0.1644
$\mathbf{p}^{(1)}(\eta_0)$	0.09	0.18	0.09	0.24	0.24	0.16
$\mathbf{p}^{(2)}(\theta_0)$	0.0784	0.1792	0.1024	0.2560	0.2240	0.1600

In this example we assume that the Full Model is also the true model. In this case, the components $p_\ell^{(1)}(\eta_0)$ and $p_\ell^{(2)}(\theta_0)$ of $\mathbf{p}^{(1)}(\eta_0)$ and $\mathbf{p}^{(2)}(\theta_0)$, respectively, are computed similarly to (23), with Y_ℓ/n replaced by the multinomial cell probabilities p_ℓ.

Figure 5 compares the function GENO$(n, k, Full)$ for $k = 1, 2$, computed according to (3) by numerical calculation of the expectations involved. For small n the small Model 1 has the largest GENO and, hence, it is preferred. For example, GENO$(70, 1, Full) \approx 160 \approx$ GENO$(90, 2, Full)$, that is, Model 1 with 70 observations is equivalent to the full multinomial model with 160 observation, and also equivalent to Model 2 with 90 observations. When n is larger than about 170 and smaller than about 250, Model 2 is better, while for larger n's the full multinomial model is preferred.

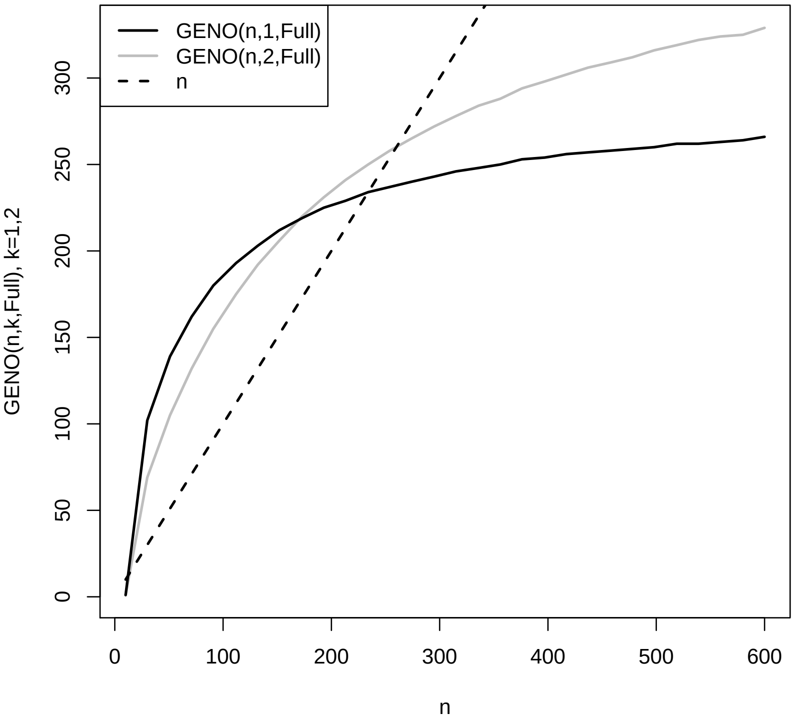


Figure 5: Plots of GENO$(n, k, Full)$ for $k = 1, 2$ and for $n = 1, \ldots, 600$.

B.2 Estimation

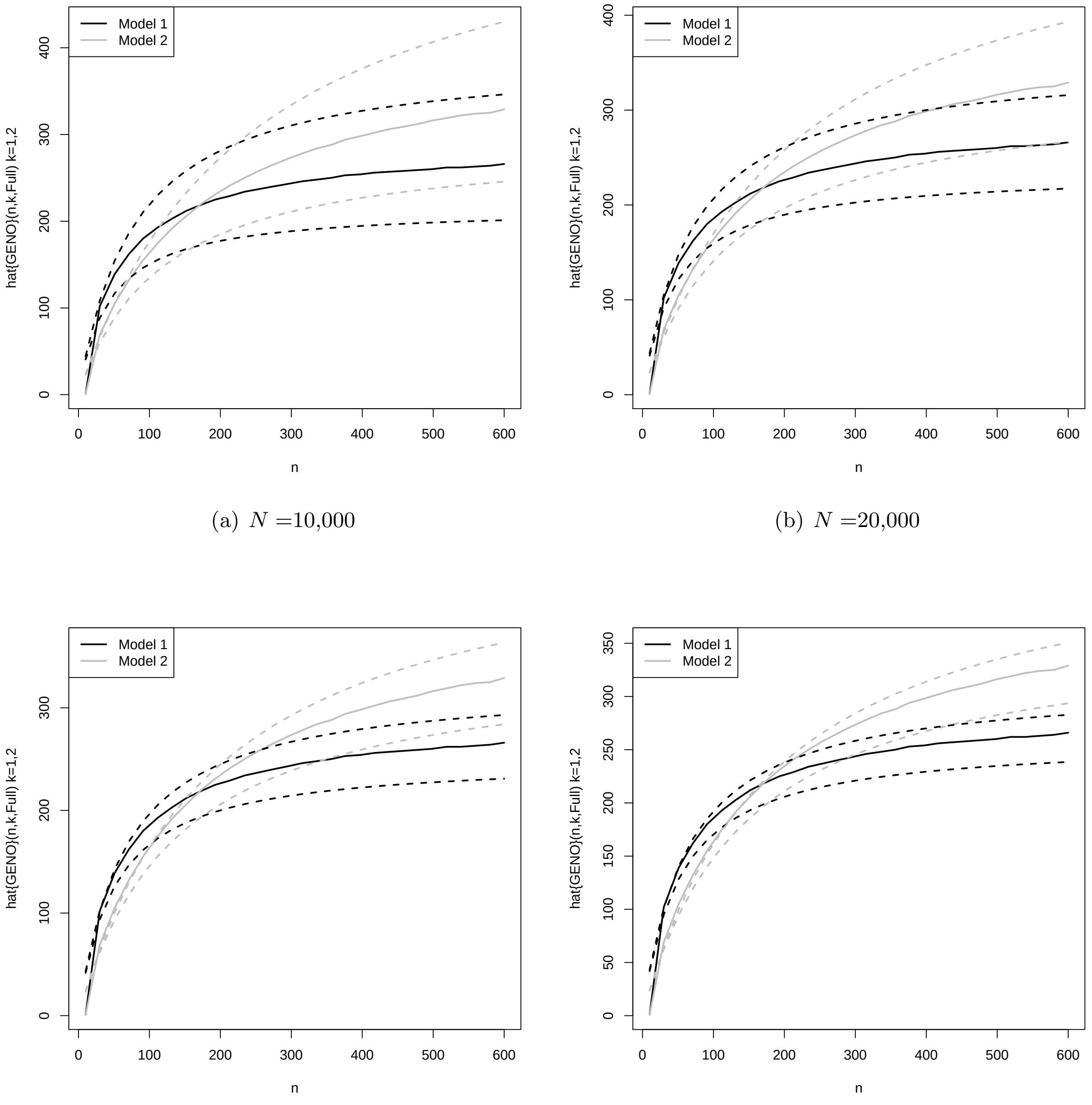
We now consider the estimator (5) and study its behavior for different values of N. For a sample $(Y_1, \cdots, Y_6) \sim \text{Multinomial}(N, p)$, (5) reads as

$$\widehat{\text{GENO}}(n, 1, Full) = \frac{5/2}{\sum_{\ell=1}^{6} \widehat{p}_\ell \log[\widehat{p}_\ell / p_\ell^{(1)}(\widehat{\eta})] - \frac{5-1}{2N} + \frac{1}{2n}}, \quad \text{and}$$

$$\widehat{\text{GENO}}(n, 2, Full) = \frac{5/2}{\sum_{\ell=1}^{6} \widehat{p}_\ell \log[\widehat{p}_\ell / p_\ell^{(2)}(\widehat{\theta})] - \frac{5-2}{2N} + \frac{2}{2n}},$$

where the MLE estimators are given by (23) and $\widehat{p}$ is the empirical mean.

Figure 6 plots the region where 95% of the estimates $\widehat{\text{GENO}}(n, 1, Full)$, $\widehat{\text{GENO}}(n, 2, Full)$ fall, based on the 0.025,0.975 quantiles of 10000 simulations of $\sum_{\ell=1}^{6} \widehat{p}_\ell \log[\widehat{p}_\ell / \{p^{(1)}(\widehat{\eta})\}_\ell]$ for Model 1 and $\sum_{\ell=1}^{6} \widehat{p}_\ell \log[\widehat{p}_\ell / \{p^{(2)}(\widehat{\theta})\}_\ell]$ for Model 2. For small N, the confidence intervals of GENO$(n, 1, Full)$ and GENO$(n, 2, Full)$ are quite wide and they overlap. For example, for n=300, GENO$(300, 1, Full) = 244$, while the confidence intervals are (189,312), (202,286), (214,267), (221,259) for N =10,000, 20,000, 50,000, 100,000, respectively. Thus, in this example N needs to be quite large in order to obtain good estimates.



(a) N =10,000

(b) N =20,000

(c) N =50,000

(d) N =100,000

Figure 6: Plots of estimates of GENO$(n, 1, \mathit{Full})$ (black) and GENO$(n, 2, \mathit{Full})$ (gray). The dashed lines are bounds based on the 0.025,0.975 quantiles of the 10000 simulations of $\sum_{j=1}^{6} \widehat{p}_j \log[\widehat{p}_j/p_j^{(1)}(\widehat{\eta})]$ for Model 1 or $\sum_{j=1}^{6} \widehat{p}_j \log[\widehat{p}_j/p_j^{(2)}(\widehat{\boldsymbol{\theta}})]$ for Model 2.

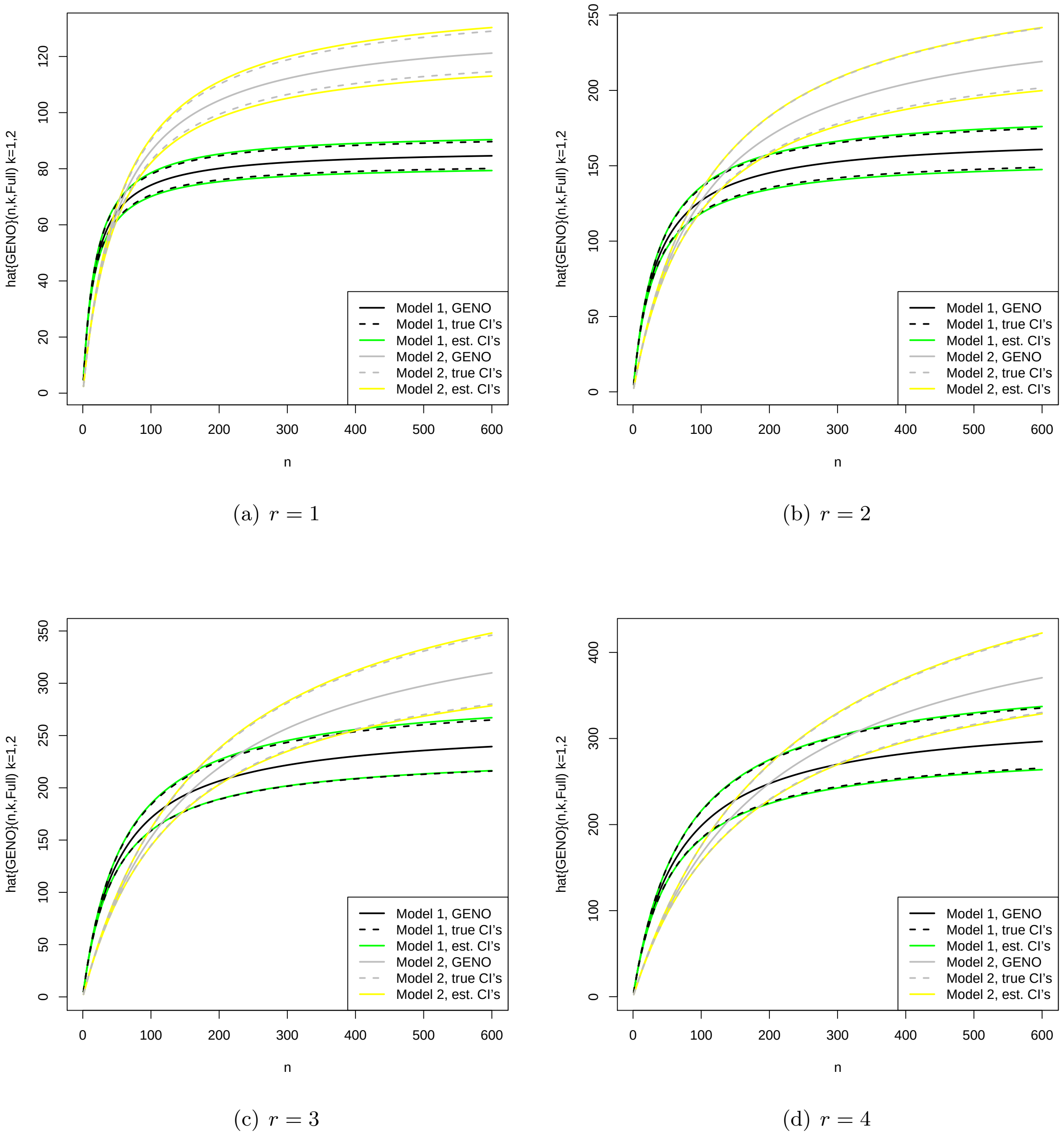
B.3 Many experiments

In this Section we perform simulations to assess the estimates of GENO under the scenario of Sections 2.2 and 3.3, namely, that there are many experiments, each of which has a different $\mathbf{p}$, with models as in Section B.1. Based on these experiments we would like to estimate the common GENO as in (13) and to compare it to the true value. We consider the case where all N_j's are equal to some N. This scenario is quite standard when we have a sample of DNA of N organisms of some species and we consider allele frequencies in different loci, which can be assumed independent; that is, there is no linkage disequilibrium.

We conducted simulations of a scenario having the above flavor. The description of the simulation is somewhat involved. We performed 1000 simulations of $J = 500$ experiments, each of size $N = 200$. In terms of the DNA example of Section 3.3, this corresponds to analyzing 500 SNPs on the basis of a sample of 200 DNA sequences. These numbers are somewhat in between the values of the DNA data of Section 3.3 (where $N \approx 50$ and $J \approx 50,000$) and the experimental economics data of Section 5 (where $N = 500$ and $J = 180$). We start by fixing a limiting value of GENO for each $j = 1, \ldots, J$ and then computing corresponding probability vectors. More specifically, for each of the J experiments we first chose a value for $\lim_{n\to\infty} \mathrm{GENO}(n, 1, \mathit{Full})$. These values, denoted by $G^{(j)}$, $j = 1, \ldots, 200$, were chosen by sampling from the Normal distribution $\mathcal{N}\left(100, 25^2\right)$, so that we have 200 GENOs that are roughly of the same order. We then chose, by solving a non-linear equation, $\mathbf{p}^{(j)}$, using the approximation $\lim_{n\to\infty} \mathrm{GENO}(n, 1, \mathit{Full}) \approx \frac{5/2}{\sum_{\ell=1}^{6} p_\ell \log[p_\ell / p_\ell^{(1)}(\eta_0)]}$, such that $\lim_{n\to\infty} \mathrm{GENO}(n, 1, \mathit{Full}) = G^{(j)}$ and $\lim_{n\to\infty} \mathrm{GENO}(n, 2, \mathit{Full}) = 1.5G^{(j)}$. In other words, in the j-th experiment, the first model with infinitely many observations (that is, a large number) is equivalent to the full model with $G^{(j)}$ observations, and the second is equivalent to the full model with $1.5G^{(j)}$ observations. The factor 1.5 was chosen since it is close to the numbers 407.1/283.8=1.43 of Section B.1.

We computed 95% bootstrap confidence intervals by (14) and compared them to the true distribution. The way the above GENO's were generated is, of course, arbitrary, and we repeated the whole experiment four times, generating GENOs from the $\mathcal{N}\left(100r, (25r)^2\right)$ distribution with $r = 1$ (the case above) and also $r = 2, 3, 4$.

Figure 7 plots $E\{\widehat{\mathrm{GENO}}(n, k, \mathit{Full})\}$ (estimated by simulations), and the mean bounds of the bootstrap confidence intervals compared to the true bounds computed from the simulation quantiles. The bootstrap confidence intervals are (very) slightly conservative, as expected.



(a) $r = 1$

(b) $r = 2$

(c) $r = 3$

(d) $r = 4$

Figure 7: Plots of estimates of $E\{\widehat{\text{GENO}}(n, k, Full)\}$, $k = 1, 2$ for many experiments. The green and yellow lines are the mean bounds of the bootstrap 95% confidence interval of $\widehat{\text{GENO}}(n, k, Full)$, $k = 1, 2$ respectively, where the mean is over the 1000 simulations, and the dashed lines are bounds based on quantiles of the 1000 repetitions of the GENO estimates.

C $\widehat{T}$ for the DNA data

Table 6: $\widehat{T} = \widehat{T}(HW, Full)$ for the different chromosomes for the populations ASW, CEU.

	ASW				CEU			
Chromosome	$\widehat{T}$	CI	$\mathcal{N}_c$	$J = \frac{\mathcal{N}_c}{20}$	$\widehat{T}$	CI	$\mathcal{N}_c$	$J = \frac{\mathcal{N}_c}{20}$
Chromosome 1	282.7	(226.4,365)	92929	4646	947.5	(643.8,1713.1)	149620	7481
Chromosome 2	273.5	(221.9,355.4)	95916	4796	683.5	(516.6,1000.9)	169373	8469
Chromosome 3	420.5	(299.5,683)	79350	3968	436.4	(350.6,576.4)	138785	6939
Chromosome 4	272.9	(213.8,380.7)	72049	3602	487.8	(382.1,678.8)	125677	6284
Chromosome 5	406.1	(289.8,658.9)	71898	3595	582.2	(442.7,835.6)	130933	6547
Chromosome 6	278.7	(216.8,381.4)	73943	3697	445.9	(359.1,587.5)	139229	6961
Chromosome 7	233.5	(184.5,314.2)	62170	3108	443.2	(344.8,617.3)	111532	5577
Chromosome 8	388.6	(275.1,659.3)	62755	3138	969.3	(629.1,2060.9)	113614	5681
Chromosome 9	249.9	(191.7,352.3)	52507	2625	949.2	(598.1,2168.5)	94208	4710
Chromosome 10	326.4	(241.9,492.4)	59495	2975	388.3	(313.7,502.9)	103347	5167
Chromosome 11	303.8	(228.7,448.3)	57809	2890	783.9	(527.9,1478.2)	100706	5035
Chromosome 12	392.1	(265.6,711.3)	55002	2750	646.5	(453.9,1109.4)	94978	4749
Chromosome 13	337.2	(227.9,622.1)	42386	2119	463.3	(344.3,686.4)	77526	3876
Chromosome 14	246.6	(183,368.5)	37043	1852	774.6	(481.9,1800.4)	63702	3185
Chromosome 15	460.3	(287.9,1132.5)	34474	1724	505.2	(355.9,866.5)	55102	2755
Chromosome 16	589.6	(325.6,2761.8)	36406	1820	523.4	(372.8,874.3)	55874	2794
Chromosome 17	350.8	(232.3,700.2)	30687	1534	426.8	(312.6,684.1)	46989	2349
Chromosome 18	438.3	(268.9,1111.6)	34357	1718	664.5	(421.1,1473.4)	58296	2915
Chromosome 19	184.6	(130.6,312.4)	20858	1043	613.8	(368.7,1728)	31015	1551
Chromosome 20	212	(157.3,321)	29107	1455	997.3	(538.5,5362.6)	47875	2394
Chromosome 21	388.6	(231.8,1136.5)	16410	820	3433.6	(767.1,-1475.7)	27037	1352
Chromosome 22	186.9	(133.3,302.6)	16012	801	590.1	(355.6,1648.6)	25761	1288
Chromosome X	8.3	(8,8.6)	42207	2110	4.6	(4.6,4.7)	54889	2744