Measuring the usefulness of statistical models [1]

David Azriel[a] and Yosef Rinott[b]

September 21, 2015

[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 of the usefulness of a statistical model in terms of its relative ability to predict new data. The new measure, to be called GENO, 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 initially motivated by the study of data from experimental economics, and we also provide a detailed biological example.

1 Introduction

In the context of model selection, the classical Akaike information criterion AIC, (Akaike, 1974) of a given model and sample can be seen as an estimate of the model's expected log-likelihood of new hypothetical data, with parameters determined by the given sample. This expected log-likelihood represents the predictive quality of the model. The AIC is based on a penalized maximized log-likelihood function, whose value, however, does not easily lead to an understandable measure of the relative value of the chosen model. Akaike (1983) and others proposed to consider various functions of the AIC value of a given model relative to the best model according to AIC, as a measure of the quality of the given model. These measures include AIC differences, likelihood of models, and Akaike weights; See, e.g., Burnham and Anderson (2002).

Given a sample, our main goal is to provide a meaningful measure of the predictive quality of models. The measure we propose, to be called GENO, allows us to compare different models with different sample sizes and quantify their predictive value in a way that takes into account both the nature of the sample, and its size. Thus, to G. Box's saying "*all models are wrong, but some are useful*," we would add that "some models are useful for certain sample sizes, and other models for other sample sizes," highlighting the well-known fact that the usefulness of a model depends not only on the nature of the data, but also on the available sample size. As in much of the model selection literature, we search for useful rather than 'true' models, and usefulness is measured in terms of the predictive power of the model.

[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).

Our original motivation and main example is in the context of data analysis of repeated games and players' behavior in experimental game theory. A meaningful measure of the value or usefulness of models for strategies in game-theoretic experiments, called ENO (Equivalent Number of Observations), was defined and discussed in Erev, Roth, Slonim, and Barron (2007) (henceforth ERSB); see also the references therein for an earlier version of ENO.

Roughly speaking, consider two ways of predicting the proportion of playing a certain strategy in a given experiment. The first way consists of just using the empirical sample proportion. The second is to model players' behavior as a stochastic process, estimate its parameters, and use the estimated model for prediction. 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.

Our new measure of usefulness of a model is 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. We generalize and compare models relative to a properly chosen reference model, rather than just to empirical frequencies. Furthermore, we generalize from predicting proportions to more general prediction, and we quantify prediction differently, using Akaike's AIC approach. 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 estimation of GENO is connected to the AIC approach in a natural way, and is simpler than that of ENO. It differs from other AIC related measures as those mentioned above in many ways. In particular GENO provides meaningful comparisons of different models with different sample sizes. Also, it can 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 in terms of a simple example. Consider data modeled either by the multinomial distribution with three cells having probabilities p_1, p_2, p_3 summing to one and thus belonging to a two-dimensional simplex, or by the Hardy–Weinberg (HW) model that assumes that (p_1, p_2, p_3) lie on a one-dimensional path given by $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$, $p_3 = (1-\theta)^2$, for some parameter θ. If the HW model fits the data reasonably well, one can compare estimation of the p_i's on the basis of this model, to simple relative frequency estimators of the multinomial parameters p_i. It is to be expected that for a small data set, the smaller HW model will be preferred, whereas with more data, one can estimate the p_i's by empirical frequencies and make predictions without using such models. GENO of the HW model with n observations relative to the multinomial model, is the value of m, the multinomial sample size, for which the two models are equally accurate in a sense to be defined. Note again that GENO is a function of the model and a given sample size n with which the model is to be used. It may seem that the empirical frequency estimators of p_i are model-free, but this is not so. The multinomial model is indeed a model. It may serve as a natural reference model, and, in general, GENO compares models in terms of a natural reference model, to be discussed below.

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, GENO$(n;k)$, of the kth model with 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. 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 Definition

In this section we define GENO. Its estimation is discussed in Section 2.3. 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. Several 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 MLE for the kth model based on n observations, and

$$\theta_0^{(k)} := \arg \max_{\theta^{(k)} \in \Theta^{(k)}} \int g(y) \log f_k(y, \theta^{(k)}) dy = \arg \min_{\theta^{(k)} \in \Theta^{(k)}} \int g(y) \log \frac{g(y)}{f_k(y, \theta^{(k)})} dy. \tag{1}$$

The latter integral is the Kullback–Leibler (henceforth KL) divergence between g and $f_k(y, \theta^{(k)})$ and therefore $\theta_0^{(k)}$, or more precisely $f_k(y, \theta_0^{(k)})$, is the projection in terms of the KL divergence of g on the family $\{f_k\}$.

In addition to these models we consider a "*reference model*" which forms a high-dimensional family, $\{f\} = \{f(y, \theta)\}_{\theta \in \Theta}$, where $\Theta \subseteq \mathbb{R}^d$, and $d_k \leq d$ for $k = 1, \ldots, K$. Set

$$\widehat{\theta}_n := \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(Y_i, \theta) \text{ and } \theta_0 := \arg \max_{\theta \in \Theta} \int g(y) \log f(y, \theta) dy$$

to be the MLE of the family $\{f\}$ and the corresponding projection, respectively. In general, we assume that the reference model family $\{f\}$ is closer to the true model g than each of the K models in the KL divergence sense, that is,

$$\int_{-\infty}^{\infty} g(y) \log f(y, \theta_0) dy \geq \int_{-\infty}^{\infty} g(y) \log f_k(y, \theta_0^{(k)}) dy, \tag{2}$$

for $k = 1, \ldots, K$, where $f(y, \theta_0)$ is the KL projection of g on the family $\{f\}$. This is of course true if $\{f\}$ contains all K candidate models. For example, it may be their mixture, so that $f(y, \theta) = \sum_k \alpha_k f_k(y, \theta^{(k)})$, with θ comprising the α_k's and $\theta^{(k)}$'s. Other examples include the full multinomial model, described in Section 1 and discussed again below, which contains models

of the Hardy–Weinberg type, or a regression model where the reference model contains a set of observed covariates, and the other K models contain different subsets of these variables. The main application, using m-step Markov models, is discussed in Section 5. The choice of the reference model will be further discussed in Section 2.4.

In general we are interested in studying 'good' models, and we assume that our candidate models and the reference model are adequate models, that is, they are reasonably close in the KL sense to the true g. The approximations in the following sections are valid under this assumption.

We now introduce GENO (Generalized Equivalent Number of Observations), a measure of usefulness of a model for a given sample size. Given a sample of n observations, consider the models $\{f_k\}$ at the MLE $\widehat{\theta}_n^{(k)}$ based on the sample, 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 goal is to select the model k that maximizes the expected log-likelihood of the new data, which in the normal case coincides with a sum of squares of deviations,

$$\frac{1}{n^*} E \sum_{i=1}^{n^*} \log f_k(Y_i^*, \widehat{\theta}_n^{(k)}) = E \int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy; \tag{3}$$

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 (3) the size of the Y^*'s sample does not play any role. In the spirit of ERSB we now define, GENO$(n;k)$, a measure that quantifies the usefulness of the kth model with n observations relative to the reference model $\{f\}$. This is done by invoking a sequence of approximate quantities GENO$_1$-GENO$_3$ before getting to the final definition of GENO$(n;k)$ in (9). We define GENO$_1$ as the value m that satisfies $E \int_{-\infty}^{\infty} g(y) \log f(y, \widehat{\theta}_m) dy \approx E \int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy$. In view of (3), GENO$_1$ is the value of m for which the k-th model with parameters estimated from a sample of n observation produces the same expected log-likelihood for new observations as the reference model $\{f\}$ with m observations. More formally:

$$\text{GENO}_1(n;k) := \left\{ \max_m \ : \ E \int_{-\infty}^{\infty} g(y) \log f(y, \widehat{\theta}_m) dy \le E \int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy \right\}; \tag{4}$$

that is, GENO$_1$ is the largest value of m for which the k-th model with n observations is still better than the high-dimensional reference model $\{f\}$ with m observations. In view of (2) one can expect, as will be shown below, that with large enough m the model $\{f\}$ will be better than $\{f_k\}$ with n observations, while the smaller model $\{f_k\}$, having fewer parameters, may be better if m is small. More specifically, by (2), the expression $E \int_{-\infty}^{\infty} g(y) \log f(y, \widehat{\theta}_m) dy$ converges to a larger value than the any term on the right-hand side of the inequality in (4).

GENO$_1(n;k) < n$ means that the reference model with n observations is better than the kth model, whereas GENO$_1(n;k) > n$ suggests that the kth model is better. For a given sample size n, it now makes sense to choose the model with the largest GENO$_1$. Clearly, GENO$_1$ depends on the choice of the reference model, but when we compare GENO$_1$ for two models, the reference model cancels. Using GENO$_1$, one can compare different models with different sample sizes: a relation such as GENO$_1(50,1)$=GENO$_1(100,2)$, which by (4) does not depend on the reference model, means that model 1 with $n = 50$ and model 2 with $n = 100$ are equally useful. In the same way, equality of two GENOs as defined by (9) below, or two GENO estimates defined by (10), does not depend on the choice of the reference model.

2.2 Approximations from GENO_1 to GENO

In order to extract the value of m for which the two integrals in (4) are equal, and later estimate it, we approximate GENO_1. The precise required conditions can be found in detail in White (1982). These conditions include the existence and domination of second order derivatives appearing below, and require that parameters of KL projections $\theta_0^{(k)}$ are interior points of their parameter space. In addition, for estimates as in (5) below, we need to assume the existence of third order dominated derivatives. To start, we have

$$\int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy = \int_{-\infty}^{\infty} g(y) \log f_k(y, \theta_0^{(k)}) dy + \int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy - \int_{-\infty}^{\infty} g(y) \log f_k(y, \theta_0^{(k)}) dy.$$

We perform standard AIC-type computations similar to those appearing in the study of the asymptotic distribution of MLE's as in Burnham and Anderson (2002) Chapter 7.2 with different notations. By (1) and the assumption that $\theta_0^{(k)}$ is interior, we have $\int_{-\infty}^{\infty} g(y) \frac{\partial}{\partial \theta} \log f_k(y, \theta_0^{(k)}) dy = 0$, and therefore a second order Taylor expansion leads to

$$E\left[\int_{-\infty}^{\infty} g(y) \log f_k(y, \widehat{\theta}_n^{(k)}) dy - \int_{-\infty}^{\infty} g(y) \log f_k(y, \theta_0^{(k)}) dy\right] = -\frac{Tr_k}{2n} + O(1/n^{3/2}), \qquad (5)$$

where Tr_k denotes the trace of the matrix H defined below as the product two $d_k \times d_k$ matrices

$$H := \left\{-E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_k(Y_1, \theta)\Big|_{\theta=\theta_0^{(k)}}\right]\right\}^{-1} E\left[\frac{\partial}{\partial \theta_i} \log f_k(Y_1, \theta)\Big|_{\theta=\theta_0^{(k)}} \frac{\partial}{\partial \theta_j} \log f_k(Y_1, \theta)\Big|_{\theta=\theta_0^{(k)}}\right],$$

and $\theta = (\theta_1, \ldots, \theta_{d_k})$, $\theta_0^{(k)} \in \mathbb{R}^{d_k}$; above we use the notation $[A_{ij}]$ for a matrix whose entries are A_{ij}. Later, when f_k is replaced by f in H, we denote its trace by Tr. The trace is quite difficult to compute or to estimate, and hence it is often replaced by d_k, the dimension. The justification is that if g is contained in the family $\{f_k\}$ so that $g(y) = f_k(y, \theta_0^{(k)})$, then by standard definitions both matrices appearing in H (the first matrix before taking inverse) express Fisher's information matrix, and therefore $H = I_{d_k}$, the identity matrix of order d_k, and $Tr_k = d_k$. We do not make the assumption that g is contained in the family $\{f_k\}$, and thus this step is an approximation, which is good for 'good' models in the sense of being close to the true g. This approach is akin to the notion of Pitman's efficiency, which compares sample sizes of contiguous alternatives (see, e.g., van der Vaart, 1998, Chapter 14).

Combining the result of (5) and a similar calculation applied to the family $\{f\}$ instead of $\{f_k\}$ with d replacing d_k, we obtain

$$\mathrm{GENO}_1(n;k) = \left\{\max_m : \frac{Tr}{2m} + O(\frac{1}{m^{3/2}}) \geq \int_{-\infty}^{\infty} g(y) \log\left[f(y, \theta_0) \Big/ f_k(y, \theta_0^{(k)})\right] dy + \frac{Tr_k}{2n} + O(\frac{1}{n^{3/2}})\right\}. \qquad (6)$$

For good models GENO is not small, and if also n is not very small, we can drop the smaller order terms in m and n in (6), and approximate GENO_1 by

$$\mathrm{GENO}_2(n;k) := \left\{\max_m \ : \ \frac{Tr}{2m} \geq \int_{-\infty}^{\infty} g(y) \log\left[f(y, \theta_0) \Big/ f_k(y, \theta_0^{(k)})\right] dy + \frac{Tr_k}{2n}\right\}. \qquad (7)$$

In view of the above discussion we replace the traces by the corresponding dimensions, and obtain

$$\mathrm{GENO}_3(n;k) := \left\{ \max_m \ : \ \frac{d}{2m} \geq \int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy + \frac{d_k}{2n} \right\}. \tag{8}$$

The right-hand side of the inequality in (8) is a positive constant while the left-hand side is decreasing to zero with m so the maximum is finite. We aim at the critical m for which equality holds approximately, and thus we finally define

$$\mathrm{GENO}(n;k) := \frac{d/2}{\int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy + \frac{d_k}{2n}}. \tag{9}$$

We have that

$$\frac{\mathrm{GENO}_2(n;k)}{\mathrm{GENO}_3(n;k)} = \frac{1 + \frac{Tr-d}{d}}{1 + \frac{Tr_k - d_k}{2n} \Big/ \left\{ \int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0)/f_k(y,\theta_0^{(k)}) \right] dy + \frac{d_k}{2n} \right\}}.$$

For fixed n, this ratio is close to 1 if $Tr - d$ and $Tr_k - d_k$ are close to zero. This means that the approximation $\mathrm{GENO}_2(n;k) \approx \mathrm{GENO}_3(n;k)$ is valid for relatively good models, as discussed above. In the sequel we will work with GENO as given by (9), which is easier to compute and to estimate, and consider it as the definition of GENO.

2.3 Estimation of GENO

Recall that $\mathrm{GENO}(n;k)$ measures the predictive value of model k when its parameters are estimated by a sample of n observations. For the estimation of GENO we assume we have a training sample $Y_1, \ldots, Y_N$ of size N from g, where N may equal n, but in general we have in mind the case where $N > n$. For example, suppose that N observations arise from pooling data from many subjects in order to select a model. We are going to apply the selected model to a particular subject, estimating the subject's parameters on the basis of his n observations. The measure $\mathrm{GENO}(n;k)$, estimated on the basis of the N observations, quantifies the predictive value of the model for such a subject. Here we assume for simplicity that the data come from the same g (at least to a reasonable approximation). This assumption is relaxed and the example is elaborated in Section 2.6.

In order to estimate $\mathrm{GENO}(n;k)$ on the basis of a sample $Y_1, \ldots, Y_N$ from g note first that

$$\int_{-\infty}^{\infty} g(y) \log f_k(y,\theta_0^{(k)}) dy$$
$$= \frac{1}{N} \sum_{i=1}^{N} \log f_k(Y_i, \widehat{\theta}_N^{(k)}) - \left[\frac{1}{N} \sum_{i=1}^{N} \log f_k(Y_i, \widehat{\theta}_N^{(k)}) - \int_{-\infty}^{\infty} g(y) \log f_k(y,\theta_0^{(k)}) dy \right].$$

Again, by standard AIC-type calculations, the expectation of the square brackets is approximated by $\frac{d_k}{2N}$, and by a similar argument for the family $\{f\}$, an approximately unbiased estimate to the integral appearing in the denominator of (9), $\int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy$ is

$$\frac{1}{N} \sum_{i=1}^{N} \log \left[f(Y_i, \widehat{\theta}_N) \Big/ f_k(Y_i, \widehat{\theta}_N^{(k)}) \right] - \frac{d - d_k}{2N}.$$

Summing up, we define our estimator as

$$\widehat{\mathrm{GENO}}(n;k) := \frac{d/2}{\frac{1}{N}\sum_{i=1}^{N}\log\left[f(Y_i,\widehat{\theta}_N)\Big/ f_k(Y_i,\widehat{\theta}_N^{(k)})\right] - \frac{d-d_k}{2N} + \frac{d_k}{2n}}. \tag{10}$$

Our standard assumption is that the reference model f is close to the true g in the KL sense. The sum in the denominator of (10) is small when the model f_k is close to the reference model f, and in this case the bias corrections appearing in this denominator play a role. Behind the approximations leading to the bias corrections is the assumption that f_k is a 'good' models that is close to both f and the true g. For a 'poor' model f_k, the sum in the denominator of (10) will be large, the bias corrections will not play a role, but GENO will be small, indicating that the model is poor.

2.4 The choice of the reference model

In certain situations the reference model arises naturally. This may be the case when we consider data in cells or a contingency table. A multinomial model is natural as the reference model, and it can be compared to models like HW, or models expressing certain dependence structures in contingency tables, where the parameters lie in lower dimensional parameters spaces. When models are nested one may choose a maximal model as the reference, if such a model is natural. However, in the absence of a natural choice, selecting the reference model $\{f\}$ by the AIC criterion based on the available N observations, seems natural. This is the case in the application that motivated this study, given in Section 5.

2.5 Model selection with $\widehat{\mathrm{GENO}}(n;k)$

Consider a situation where we have at hand a training sample of size N, and we want to choose a model for the analysis of a sample of size n of similar data, which may be available now or in the future. These two samples may be distinct, or they may be one and the same sample that we wish to analyze, in which case $N = n$. We propose to choose the model having the largest value of $\widehat{\mathrm{GENO}}(n;k)$. If another model that is preferable for some technical or aesthetic reasons has a high GENO, it may be chosen, and GENO provides a measure of how much is lost relative to the best model. Examples are given in Section 3.

The estimator $\widehat{\mathrm{GENO}}(n;k)$ of (10) quantifies the value of the kth model with n observations, and allows comparisons of different models with different sample sizes. By (10), choosing the model with the largest GENO for a given sample size n amounts to choosing the model with k having the smallest value of $-\frac{1}{N}\sum_{i=1}^{N}\log\left[f_k(Y_i,\widehat{\theta}_N^{(k)})\right] + d_k(\frac{1}{2n}+\frac{1}{2N})$. Based on the training sample, we choose the best model for inference based on a sample of size n. When $N = n$, the above is equivalent to choosing the model k that minimizes Akaike's AIC criterion, where

$$AIC := -\sum_{i=1}^{n}\log\{f_k(Y_i,\widehat{\theta}_n^{(k)})\} + d_k. \tag{11}$$

Note that it is quite common to call AIC twice the above quantity. The AIC was derived in Akaike (1974) from the point of view of the Kullback–Leibler projection described above.

When two models, f_k and f_ℓ, are compared then the first will be chosen if

$$-\frac{1}{N}\sum_{i=1}^{N}\log\left[f_k(Y_i,\widehat{\theta}_N^{(k)}/f_\ell(Y_i,\widehat{\theta}_N^{\ell})\right] + (d_k - d_\ell)(\frac{1}{2n}+\frac{1}{2N}) < 0.$$

The correction term $(d_k - d_\ell)(\frac{1}{2n} + \frac{1}{2N})$ is relevant only if the two models are close so that in absolute value the first term above is small, and formally only if one converges to the other in the spirit of Pitman. This applies also to the classical AIC.

2.6 GENO for many experiments with a common model

The flexibility of the notion of GENO is demonstrated in this section in which we consider a collection of J experiments or data sets, possibly of different sizes N_j, that are to be analyzed together. In order to have a systematic approach that allows comparisons between the results of the analyses in the different data sets, it may be desirable to analyze all of them with a single common model, and estimate the parameters separately in each data set. Our motivation arises from game-theoretic experiments on different individuals that we wish to analyze and compare. Other examples may arise when one wants to construct a common multiple regression model for different data sets, say, economic data from different countries, in order to compare the coefficients in different countries, or when one considers genetic data from a set of populations to be analyzed with a common Hardy–Weinberg-type model. Note that it is possible to compute $\mathrm{GENO}(n;k)$ by (10) 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.

Suppose that there are J experiments, each with N_j iid observations $Y_{1,j}, \ldots, Y_{N_j,j}$ having density g_j in the jth experiment. For the jth experiment, the likelihood at the MLE of the kth model is $f_k(y, \widehat{\theta}^{(k)}_{N_j,j})$, the projection is $f_k(y, \theta^{(k)}_{0,j})$, and similar notations are used for the model f. We assume that all experiments should be analyzed by the same model, to be chosen by GENO below. In short, we have in mind experiments of a similar kind, to be analyzed by the same models, but possibly with different parameter values. $\mathrm{GENO}(n;k)$ is defined below as the value of m such that the reference model f with m observations in each experiment, is equivalent to having n observations in each experiment with f_k. Thus, the first definition of GENO is

$$\mathrm{GENO}_1(n;k) := \left\{ \max_m \ : \ E\sum_{j=1}^{J} \int_{-\infty}^{\infty} g_j(y) \log f(y, \widehat{\theta}_{m,j}) dy \le E\sum_{j=1}^{J} \int_{-\infty}^{\infty} g_j(y) \log f_k(y, \widehat{\theta}^{(k)}_{n,j}) dy \right\}.$$

The above quantity represents the expected likelihood of new data as in (3), where we have the same amount of new data from each g_j, or alternatively if one subject is chosen at random from one of $g_1, \ldots, g_J$. It is redefined, as in (9), by

$$\mathrm{GENO}(n;k) := \frac{d/2}{\frac{1}{J}\int_{-\infty}^{\infty} \sum_{j=1}^{J} g_j(y) \log\left[f(y,\theta_{0,j}) \Big/ f_k(y, \theta^{(k)}_{0,j}) \right] dy + \frac{d_k}{2n}}, \tag{12}$$

and is estimated by

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

Note that as above, the denominator of this GENO is similar to the AIC: it is the empirical likelihood corrected by the dimension of the models, and maximizing $\widehat{\mathrm{GENO}}(n;k)$ with respect to k amounts to maximizing the bias-corrected empirical likelihood over the k models.

2.7 The notion of GLU

GENO$(n;k)$ measures the predictive value of model k when its parameters are estimated by a sample of n observations and, hence, for each model k it is a function of n. It seems useful also to provide for each model the sample size for which its predictive value equals that of the reference model. Let GLU(k) (Upper Limit of sample size by GENO) be the number of observations n for which GENO$(n;k) = n$. When comparing the reference model and model k, if the number of observations n that we have is smaller than GLU(k) then model k is better, and otherwise the reference model is better. This suggest that model k would not be selected when n is larger than GLU(k) if the reference model is considered as one of the candidate models. We compute GLU(k) only when $d_k < d$; if $d_k = d$ and (2) holds, then GENO$(n;k) < n$ for all n, and GLU$(k) = \infty$.

More precisely, we first define

$$\mathrm{GLU}_1(k) := \left\{ \max_n \ : \ \mathrm{GENO}_1(n;k) \geq n \right\}.$$

As in (9), it is redefined by

$$\mathrm{GLU}(k) := \frac{d - d_k}{2\int_{-\infty}^{\infty} g(y) \log\left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy},$$

and is estimated by

$$\widehat{\mathrm{GLU}}(k) := \frac{(d-d_k)/2}{\frac{1}{N}\sum_{i=1}^{N} \log\left[f(Y_i,\widehat{\theta}_N) \Big/ f_k(Y_i,\widehat{\theta}_N^{(k)}) \right] - \frac{d-d_k}{2N}}.$$

For the case of many experiments as in Section 2.6 we have

$$\widehat{\mathrm{GLU}}(k) := \frac{(d-d_k)/2}{\frac{1}{J}\sum_{j=1}^{J} \frac{1}{N_j} \sum_{i=1}^{N_j} \log\left[f(Y_{i,j},\widehat{\theta}_{N_j,j}) \Big/ f_k(Y_{i,j},\widehat{\theta}_{N_j,j}^{(k)}) \right] - \frac{1}{J}\sum_{j=1}^{J} \frac{d-d_k}{2N_j}}.$$

2.8 A bootstrap confidence interval for GENO

We now discuss the construction of confidence intervals for GENO. It is possible to estimate the variance of the sum in the numerator of (10) using standard jackknife or bootstrap methods for iid observations. One can 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 $Z_j := \frac{1}{N_j}\sum_{i=1}^{N_j} \log\left[f(Y_{i,j},\widehat{\theta}_{N_j,j}) \Big/ f_k(Y_{i,j},\widehat{\theta}_{N_j,j}^{(k)}) \right]$ in (13) in order to construct a confidence interval under the assumption that the experiments are independent.

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 shows 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 the bootstrap distribution are asymptotically the same. Therefore, an asymptotic in J, $(1-\alpha)\%$ confidence interval for GENO$(n;k)$, under suitable homogeneity assumptions, is

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

where $\tilde{F}_J$ is the bootstrap distribution of $\frac{1}{J}\sum_{j=1}^{J} Z_j$. Similarly, a confidence interval for GLU(k) is

$$\left(\frac{(d-d_k)/2}{\tilde{F}_J^{-1}(\alpha/2) - \frac{1}{J}\sum_{j=1}^{J} \frac{d-d_k}{2N_j}}, \ \frac{(d-d_k)/2}{\tilde{F}_J^{-1}(1-\alpha/2) - \frac{1}{J}\sum_{j=1}^{J} \frac{d-d_k}{2N_j}} \right). \tag{15}$$

This approach is tested in Section A.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.

3 The multinomial distribution and Hardy-Weinberg model

The multinomial distribution often describes how data is generated. It appears also in connection with goodness-of-fit tests. We next discuss the application of GENO to multinomial models.

Let $X_1, \ldots, X_n$ be observations taking L possible values, say, $a_1, \ldots, a_L$ with $P(X_i = a_\ell) = p_\ell$, and set $Y_\ell = \#\{i : X_i = a_\ell\},\ \ell = 1, \ldots, L$. We assume here that the true model g is multinomial with a given parameter $\mathbf{p} = (p_1, \ldots, p_L)$ so that $Y = (Y_1, \ldots, Y_L) \sim \text{Multinomial}(n, \mathbf{p})$, which we also take as the reference model $\{f(y, \mathbf{p})\}_{\mathbf{p}\in\Delta}$, where Δ denotes the unit simplex. 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 (henceforth HW) model. For a single diploid locus with two possible alleles A, G say, that appear with probabilities $\theta, 1-\theta$, the frequencies of the three genotypes, AA, AG, GG, according to HW are $\theta^2, 2\theta(1-\theta), (1-\theta)^2$. The likelihood of (Y_1, Y_2, Y_3) is proportional to $\prod_{\ell=1}^{3} p_\ell^{Y_\ell}$, where the p_ℓ's are the components of $\mathbf{p}_{HW}(\theta) := \Big(\theta^2, 2\theta(1-\theta), (1-\theta)^2\Big)$, and the MLE is $\widehat{\theta}_n^{(HW)} = \frac{2Y_1+Y_2}{2n}$. Higher-dimensional HW type models are discussed in Appendix A.

3.2 A numerical example

As a simple example, consider a specific multinomial distribution; the probabilities under the true and the HW models are presented in Table 1. In this example we assume that the reference model is also the true model. The 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: The probabilities the true and HW models.

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

Figure 1 demonstrates the computation of $\mathrm{GENO}_1(n;HW)$ in this case. 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 reference model, that is the multinomial model with m observations, which is the vector of sample proportions. Therefore, $\mathrm{GENO}_1(200;HW)=233$. Figure 1 shows also $\mathrm{GLU}_1(HW)=255$. When $n<255$, the expected log likelihood of HW is larger, and otherwise that of the reference model is larger.

For the reference model (multinomial) we have $d=2$ and the HW model has dimension=1. By (9), since $\sum_{\ell=1}^{3}p_\ell\log[p_\ell/p_{HW}(\theta_0)_\ell]=0.00187$ we obtain

$$\mathrm{GENO}(n;HW)=\frac{1}{0.00187+1/2n}=229 \text{ for } n{=}200,\ \mathrm{GLU}(HW)=\frac{1/2}{0.00187}=267.$$

Recall that GENO and GLU are approximations to GENO_1 and GLU_1. We obtained $\mathrm{GENO}_1(200;HW){=}233$ $\mathrm{GLU}_1(HW)=255$ showing a good approximation.

Figure 2 compares the function $\mathrm{GENO}_1(n;HW)$, given by (4), and $\mathrm{GENO}(n;HW)$, as defined in (9), and it is demonstrated that the two are close; the difference is approximately 2% of $\mathrm{GENO}_1(n;HW)$.

An extensive simulation study of different HW models is given in Appendix A, where we investigate the performance of the estimates of GENO (10) and (13) and the confidence interval (14).

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 several genetically homogeneous 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

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

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

We considered data from two populations, the first is from 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 equal or greater than 0.1. GENO for chromosome X was computed separately (see Table 7), however, for reasons explained below it is very different from the other chromosomes, and therefore is was excluded from the calculations of GENO and GLU for the two populations. We sampled $J=56,694$ SNPs from ASE, and $J=98,081$ SNPs from CEU, which are one twentieth of the total number of SNPs and can be considered independent. A plot of the estimated GENO is given in Figure 3; a 95% bootstrap confidence intervals, based on (14), is also computed. Notice that although the sample size 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.

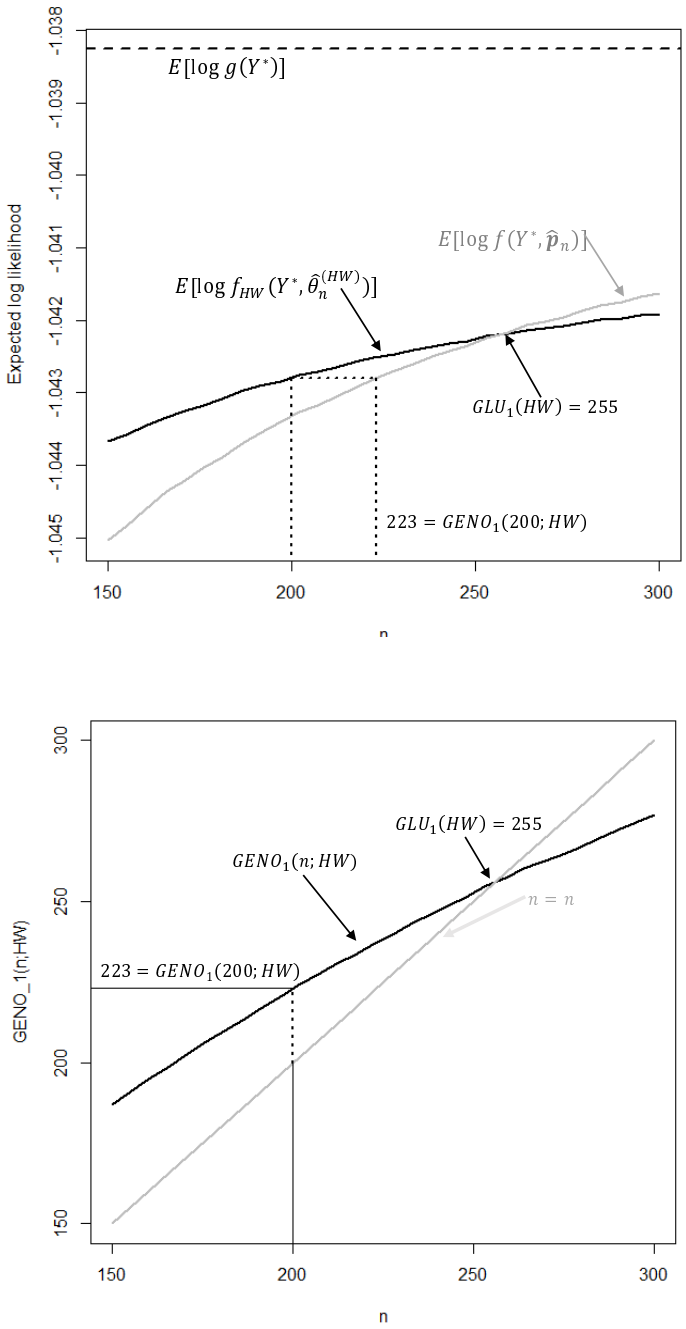


Figure 1: The computation of $\mathrm{GENO}_1(n; HW)$ based on the expected log-likelihoods $E\{\log f_{HW}(Y^*, \hat{\theta}_n^{(HW)})\}$ and $E\{\log f(Y^*, \hat{\mathbf{p}}_n)\}$. Also shown is GENO of the multinomial reference model, which is equal to n, and $\mathrm{GLU}_1(HW) = 255$.

Our calculations yield $\widehat{\mathrm{GLU}}$(HW)=332.4 for ASW and $\widehat{\mathrm{GLU}}$(HW)=591.4 for CEU; a 95% confidence interval computed by (15) is (307.7,362.8) and (544.6,647.2), respectively . We further calculated GLU(HW) 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 with minor allele frequency that is equal or greater than 0.1.

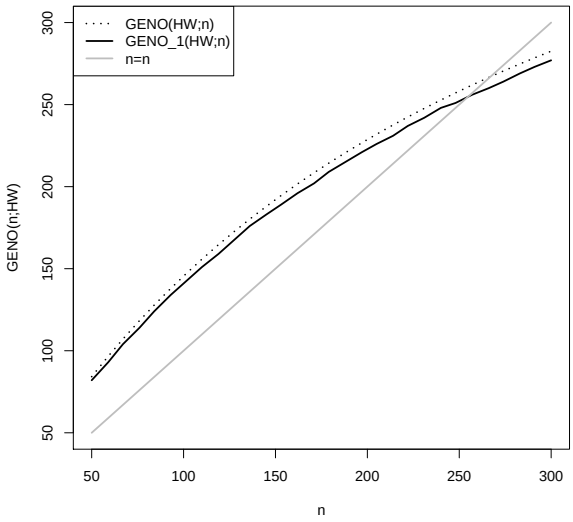


Figure 2: Plot of GENO$(n; HW)$, GENO$_1(n; HW)$ and GENO of the reference model which is equal to n.

We did not perform this calculation for chromosome Y, since there are only a few hundreds such SNPs. The results are given in Table 7 of Appendix B. Chromosome X is clearly different since the estimated GLU(HW)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 $\left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right)$ and in men the proportion is $(\theta, 0, 1-\theta)$. Assuming that half of the population are women, the proportion in the population is

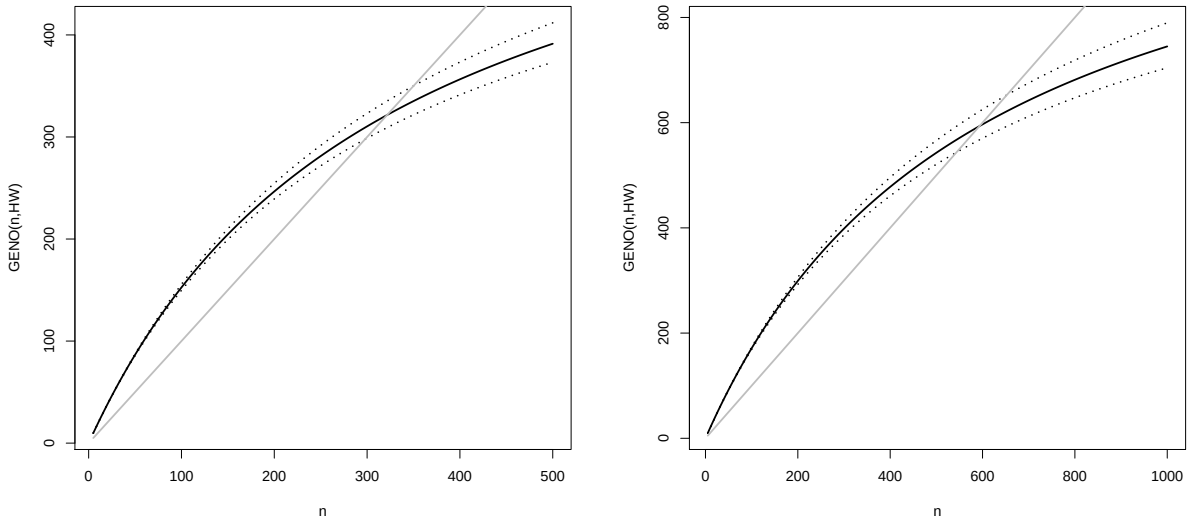
$$\mathbf{p} := \left((\theta^2+\theta)/2, \theta(1-\theta), \{(1-\theta)^2+1-\theta\}/2\right).$$

With $\mathbf{p}$ as above, we have $\theta_0 = 2p_1 + p_2 = \theta$ and $\mathbf{p}_{HW}(\theta_0) = \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right)$. GLU in this case is $1/2 \sum_{\ell=1}^{3} p_\ell \log\left[p_\ell / \{p_{HW}(\theta)\}_\ell\right]$. Here we consider $\theta \in [0.1, 0.9]$ for which GLU varies between 4 and 6.5, close to the estimated GLU of chromosome X in the data.

The bottom line is that 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 exact number seems to be different among different populations.

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



(a) ASW (b) CEU

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

5, is reanalyzed below. It involves actions in repeated games that are not independent, and we next adapt GENO to this situation. In the next section we start with more general decision processes, and later specialize to our motivating problem.

4.1 The decision process set-up

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

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

The decision process is 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 type of play, which is based on one's past actions and rewards is often called 'reinforcement learning'. We assume that the (random) reward V_i at time i depends only on the action Z_i, through a conditional probability (or density) function $p(v_i \mid z_i)$ 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

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

As $\prod_{t=1}^{n} p(v_i \mid z_i)$ 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 $f_{k,t}(z_t \mid \mathcal{D}_{t-1}, \theta^{(k)})$, where $f_{k,t}$ are known functions, 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 action at time t depends only on $U_{t-1} = (Z_{t-1}, V_{t-1}, S_{t-1})$; here S_{t-1} represents a state of the process at time $t-1$ which together with the action Z_{t-1} and the resulting reward V_{t-1} determine the distribution of the action Z_t at any time t. If, for example, the kth model assumes that the decision is based on the last M actions and rewards, then $S_{t-1}^{(k)}$ is equal to the vector $(Z_{t-M}, \ldots, Z_{t-2}, V_{t-M}, \ldots, V_{t-2})$. We assume that the true strategy g, the reference model f, and our candidate models are decision processes as defined above. In particular, Under the kth model we can model a strategy as $f_k(z_t \mid Z_{t-1}, V_{t-1}, S_{t-1}^{(k)}, \theta^{(k)}) = f_k(z_t \mid U_{t-1}^{(k)}, \theta^{(k)})$, with $U_{t-1}^{(k)} = (Z_{t-1}, V_{t-1}, S_{t-1}^{(k)})$, $\theta^{(k)} = (\alpha^{(k)}, \beta^{(k)})$, and for some functions s_k we have $S_t = S_t^{(k)} = s_k(Z_{t-1}, V_{t-1}, S_{t-1}^{(k)}, \alpha^{(k)})$. The index k is omitted under the reference model f. Under the true strategy, the process $(Z_t, U_{t-1}, S_{t-1}^{(k)})$ is Markov having a stationary distribution under well-known ergodicity conditions that are assumed. Therefore, the existence of a stationary distribution of the process $(Z_t, U_{t-1}^{(k)})$ is guaranteed, and we denote it by $q_k(z, u)$, with u in a suitable space where $U_{t-1}^{(k)}$ takes values. Integrals du should be interpreted with u in the latter space.

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

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

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

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

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 $U_t^{*(k)} = (Z_t^*, V_t^*, S_t^{*(k)})$ be the new decision, reward,

and state according to the kth model at time t. As before, we want to consider the expected log-likelihood of the new hypothetical data under the model, where the MLE $\widehat{\theta}_n^{(k)}$ is based on the given data with sample size n. The expected log-likelihood is

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

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 Definition, approximation, and estimation of GENO

Similar to the iid case, GENO_1 of the kth model is defined by

$$\begin{aligned}\mathrm{GENO}_1(n;k) := \Big\{ \max_m \ : \ &\sum_{z\in\mathcal{Z}} E\int q(z,u)\log f\left(z \mid u, \widehat{\theta}_m\right)du \\ &\le \sum_{z\in\mathcal{Z}} E\int q_k(z,u)\log f_k\left(z \mid u, \widehat{\theta}_n^{(k)}\right)du\Big\},\end{aligned}$$

where f, $\widehat{\theta}_m$ and $q(z,u)$ pertain to the reference model. As in the iid case, our assumptions on the process yield the approximation

$$E\left[\sum_{z\in\mathcal{Z}}\int q_k(z,u)\log f_k\left(z \mid u, \widehat{\theta}_n^{(k)}\right)du - \sum_{z\in\mathcal{Z}}\int q_k(z,u)\log f_k\left(z \mid u, \theta_0^{(k)}\right)du\right] \approx -\frac{d_k}{2n}, \quad (20)$$

where $\theta_0^{(k)}$ is defined in (19).

As in (5), unless the model f_k is the true model, we get a trace rather than the dimension d_k in (20). The approximation is justified when we assume that the models considered are good enough so that the trace can be replaced by the dimension, as above; for details in the Markov case see Ogata (1980); Tong (1975). Similarly,

$$E\Big[\frac{1}{N}\sum_{t=1}^{N}\log f_k(Z_t \mid U_{t-1}^{(k)}, \widehat{\theta}_N^{(k)}) - \sum_{z\in\mathcal{Z}}\int q_k(z,u)\log f_k\left(z \mid u, \theta_0^{(k)}\right)du\Big] \approx \frac{d_k}{2N}. \quad (21)$$

Applying (20) and (21) to the reference model we obtain the same approximation without the index k and in particular with d replacing d_k . Under (20), GENO can be redefined by

$$\mathrm{GENO}(n;k) := \frac{d/2}{\sum_{z\in\mathcal{Z}}\int q(z,u)\log f(z|u,\theta_0)du - \sum_{z\in\mathcal{Z}}\int q_k(z,u)\log f_k(z|u,\theta_0^{(k)})du + \frac{d_k}{2n}}, \quad (22)$$

and (21) implies that having N observations, we can estimate GENO by

$$\widehat{\mathrm{GENO}}(n;k) := \frac{d/2}{\frac{1}{N}\sum_{t=1}^{N}\log\left\{f(Z_t|U_{t-1},\widehat{\theta}_N)\Big/ f_k(Z_t|U_{t-1}^{(k)},\widehat{\theta}_N^{(k)})\right\} - \frac{d-d_k}{2N} + \frac{d_k}{2n}}. \quad (23)$$

The above represents GENO for a single player. When the data come from J players, this is generalized as in Section 2.6, equations (12) and (13), that is, we sum also over j, and the index j appears in suitable places. We will not repeat the details.

5 Game theory experiments: analysis of the motivating data

We now apply our approach to 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 an action; these choices determine the probabilities of winning a fixed amount or zero.

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 $S_t^{(k)} = \Big(S_t(0), S_t(1)\Big)$, where $S_t(i)$ is referred to as the propensity to select action i. The propensities are updated at each stage according to

$$S_t(i) = \begin{cases} (1-\alpha)S_{t-1}(i) + \alpha V_{t-1} & \text{if } Z_{t-1} = i \\ S_{t-1}(i) & \text{if } Z_{t-1} \neq i \end{cases}, \qquad i = 0, 1,$$

where $0 < \alpha < 1$ is a parameter of the model. Initially $S_1(0) = S_1(1)$ are equal to the 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

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

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

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

   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

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

   $0 < \xi < 1$ 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 $M = 1, 2, 3$. We also consider a two-actions two-rewards model, with action at stage t is based on Z_{t-2}, Z_{t-1} and V_{t-2}, V_{t-1}.

Models 1–3 are variations on the reinforcement model (Erev and Roth, 1998), 4 and 5 are standard, and models 3 and 6 have not been studied previously in this context, to the best of our knowledge. The MLE in models 1–3 is computed by numerical maximization of the log-likelihood, and estimation in models 4 and 6 is straightforward.

We first computed the AIC number of each of the models, as define in (11). The results are given in Table 2. We find that the M-step Markov models are preferred according to the AIC criterion. It is interesting to note that 3-step Markov is better than 2-actions 2-rewards. Both models have the same number of parameters, but 3-step Markov has larger log-likelihood. This means that the third previous action is more informative than the second previous reward. The $M = 2, 3$ Markov models have almost the same AIC number; therefore, for the definition of GENO, we will consider the Markov model with $M = 3$ as our reference model.

Table 2: The mean (standard deviation) of the log-likelihood and the AIC number over the 180 players.

Model	d_k	log-likelihood	AIC
RL	1	-306.9 (34.0)	307.9 (34.0)
RLL	2	-293.6 (36.8)	295.6 (36.8)
RLS	2	-251.6 (66.9)	253.6 (66.9)
Toss	1	-283.5 (60.6)	284.5 (60.6)
Nash	0	-360.8 (92.2)	360.8 (92.2)
1-step Markov	4	-229.9 (63.3)	233.9 (63.3)
2-step Markov	8	-221.7 (63.2)	229.7 (63.2)
3-step Markov	16	-213.8 (62.4)	229.8 (62.4)
2-actions 2-rewards	16	-214.3 (62.1)	230.3 (62.1)

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 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 a common parameter. 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 for the RL model is 656.1, whereas the corresponding number when individual parameters are allowed is 615.8.

5.2 GENO: results

Tables 3 and 4 show $\widehat{\mathrm{GENO}}(n, k)$ and $\widehat{\mathrm{GLU}}(k)$ for different n's and for the models mentioned in the previous section. Confidence intervals at the 95% level based on (14) are also provided.

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.

Among learning models 1–3, GENO of the RLS model varies from 95 to 125 while GENO of the other learning models is approximately 50. Nash's model has no free parameters that need to be estimated and, therefore, its GENO does not depend on n. GENO of the Nash model is about 30, while GENO of the Toss model is about 60. 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 ranges from 100 to 240 for small n and are approximately n when n is $200-600$. Hence, by our measure, the Markov models are more useful than the learning models.

For n smaller than approximately 150, 1-step Markov is the best model and for larger n, smaller than approximately 510, 2-step Markov is preferred. For larger n, 3-step Markov, which is defined as the reference model, is the best among our candidate models.

Table 3: Estimates (95% confidence intervals) of GENO(n,k) for different n's and k's. Models with the largest GENO are in bold face.

Model k	GENO$(50,k)$	GENO$(100,k)$	GENO$(150,k)$
RL	44.1 (39.3,50)	45.4 (40.3,51.6)	45.8 (40.7,52.1)
RLL	48.3 (42.6,55.3)	51.4 (45,59.4)	52.5 (45.8,60.9)
RLS	97.8 (89,108.2)	111.5 (100.2,125.1)	116.9 (104.5,132)
Toss	59.5 (51.5,69.3)	61.8 (53.2,72.4)	62.6 (53.8,73.5)
Nash	28.8 (25.1,33.2)	28.8 (25.1,33.2)	28.8 (25.1,33.2)
1-step Markov	**132.8 (123.8,142.6)**	**198.9 (179.2,221.5)**	**238.4 (210.7,271.7)**
2-step Markov	91.1 (88.6,93.2)	167.2 (159.1,174.7)	231.8 (216.4,246.4)
2-actions 2-rewards	49.6 (48.8,50.4)	98.6 (95.1,101.8)	146.8 (139.3,154)
3-step Markov	50	100	150
Model k	GENO$(200,k)$	GENO$(400,k)$	GENO$(600,k)$
RL	46 (40.9,52.4)	46.4 (41.1,52.8)	46.5 (41.2,53)
RLL	53.1 (46.2,61.7)	54 (46.9,62.9)	54.3 (47.2,63.3)
RLS	119.8 (106.8,135.7)	124.5 (110.5,141.7)	126.1 (111.8,143.9)
Toss	63 (54.1,74.1)	63.6 (54.6,74.9)	63.8 (54.7,75.2)
Nash	28.8 (25.1,33.2)	28.8 (25.1,33.2)	28.8 (25.1,33.2)
1-step Markov	264.7 (231,306.4)	317.2 (270,379)	339.6 (286.1,411.5)
2-step Markov	**287.4 (264,310)**	**448.4 (394.1,506.2)**	551.5 (471.6,641.5)
2-actions 2-rewards	194.3 (181.4,207.2)	378 (332,430.1)	551.8 (459.1,670.3)
3-step Markov	200	400	**600**

Table 4: Estimates (95% confidence intervals) of GLU(k) for different k's.

Model k	GLU(k)
RL	43.8 (38.7,50)
RLL	47.7 (41.4,55.7)
RLS	111.9 (98.8,128.4)
Toss	60.2 (51.6,71.2)
Nash	28.8 (25.1,33.2)
1-step Markov	296.7 (244,369.3)
2-step Markov	510.2 (388.6,689.1)

Acknowledgement: We are grateful to Ido Erev for the data used in Section 5.

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

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.

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.

Appendices

A First appendix: 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 A.1 - A.2. In Section A.3 we demonstrate the results by a simulation study. Consider Hardy-Weinberg models for a single diploid locus with three possible alleles a,b,c that appear 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 5.

Table 5: 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

$$\widehat{\theta}_1 = \frac{2Y_1 + Y_2 + Y_5}{2n},\ \widehat{\theta}_2 = \frac{Y_2 + 2Y_3 + Y_4}{2n};\ \widehat{\eta} = \frac{2(Y_1 + Y_2 + Y_3) + Y_4 + Y_5}{4n}. \tag{24}$$

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.

A.1 A numerical example

We consider a specific multinomial distribution. The probability vector $\mathbf{p}$ under the multinomial model with 6 cells which is the reference model, and the resulting probabilities under Models 1, 2 are presented in Table 6. 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 6: The probabilities under Models 1 and 2 and under the reference model.

	1	2	3	4	5	6
Reference 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

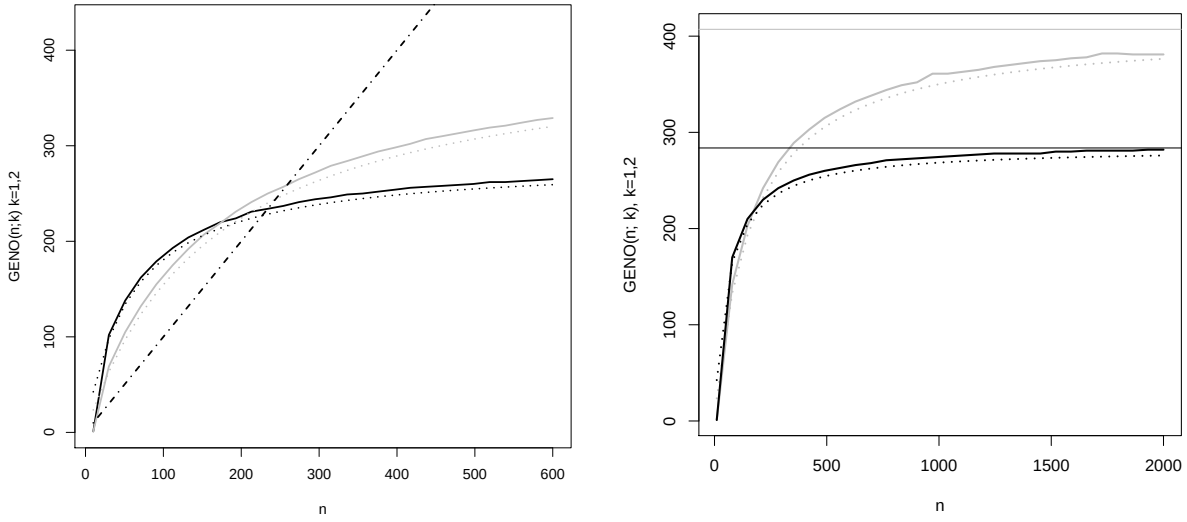
In this example we assume that the reference model is also the true model. In this case, the components $p_\ell^{(1)}(\eta_0)$ and $p_\ell^{(2)}(\theta_0)$ of $\mathbf{p}^{(1)}(\eta_0)$ and $\mathbf{p}^{(2)}(\theta_0)$, respectively, are computed similarly to (24), with Y_ℓ/n replaced by p_ℓ from the reference model.

For the reference model (multinomial) we have $d = 5$, and according to (9)

$$\mathrm{GENO}(n;1) = \frac{5/2}{\sum_{\ell=1}^6 p_\ell \log[p_\ell/p_\ell^{(1)}(\eta_0)] + 1/2n}, \;\; \mathrm{GENO}(n;2) = \frac{5/2}{\sum_{\ell=1}^6 p_\ell \log[p_\ell/p_\ell^{(2)}(\theta_0)] + 2/2n},$$

$\lim_{n\to\infty} \mathrm{GENO}(n;1) = 283.8$, $\lim_{n\to\infty} \mathrm{GENO}(n;2) = 407.1$, and GLU(1)=227.03, GLU(2)= 244.26.

Figure 4 compares the function $\mathrm{GENO}_1(n;k)$, given by (4), and $\mathrm{GENO}(n,k)$, as defined in (9), and it is demonstrated that the two are close. For small n the small Model 1 has the largest GENO and, hence, it is preferred. For example, $\mathrm{GENO}(70,1) \approx 160 \approx \mathrm{GENO}(90,2)$, which indicates that Model 1 with 70 observations is 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 reference model is preferred.



(a) (b)

Figure 4: Plots of $\mathrm{GENO}(n;1)$ (black), $\mathrm{GENO}(n;2)$ (gray) for different scales of n. The solid (respectively, dotted) line is GENO according to definition (4) (respectively, (9)). The horizontal lines represent the limit as n goes to infinity. The dot-dashed line in Plot (a) is $\mathrm{GENO}(n;)$ of the reference model, which is equal to n. Plot (a) (respectively, (b)) shows $\mathrm{GENO}(n;)$ for $n = 1, \ldots, 600$ (respectively, $n = 1, \ldots, 2000$).

A.2 Estimation

We now consider the estimator (10) and study its behavior under p of the previous section for different values of N. For a sample $(Y_1, \cdots, Y_6) \sim \mathrm{Multinomial}(N, p)$, (10) reads as

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

where the MLE estimators are given by (24) and $\widehat{p}$ is the empirical mean.

Figure 5 plots the region where 95% of the estimates $\widehat{\text{GENO}}(n;1)$, $\widehat{\text{GENO}}(n;2)$ 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$(n;1)$ and GENO$(n;2)$ are quite wide and they overlap. For example, for n=300, GENO(300,1)=238.6, 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.

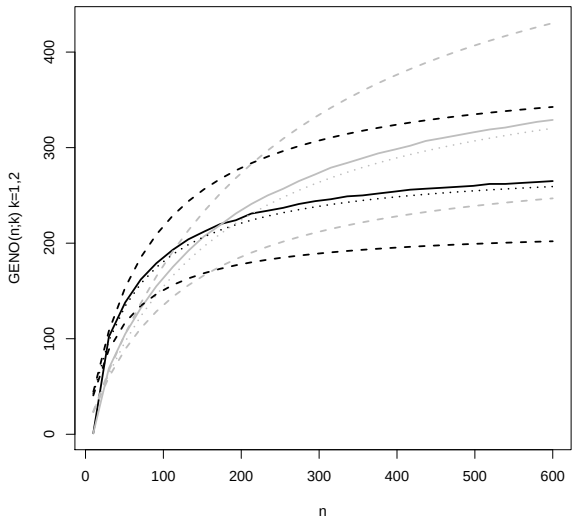
A.3 Many experiments

In this Section we perform simulations to assess the estimates of GENO under the scenario of Section 2.6, namely, that there are many experiments, each of which has a different $\mathbf{p}$, with models as in Section A.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.

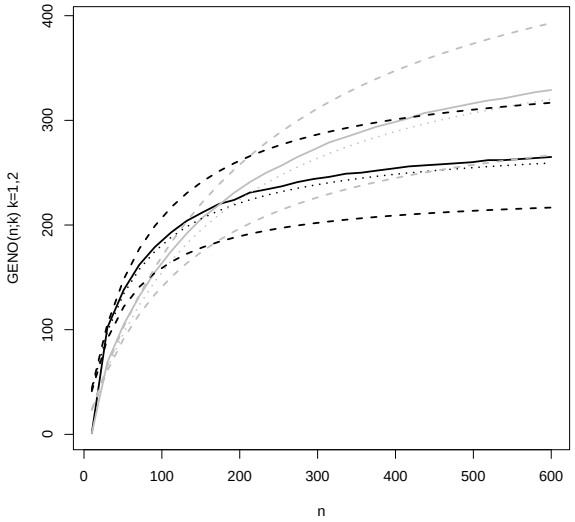
We conducted simulations of a scenario having the above flavor. The description of the simulation is somewhat involved. We performed 1000 simulations of $J = 500$ experiments, each of size $N = 200$. In terms of the DNA example of Section 3.3, this corresponds to analyzing 500 SNPs on the basis of a 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 $\lim_{n\to\infty} \text{GENO}(n;1)$. 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. Then we chose, by solving a non-linear equation, $\mathbf{p}^{(j)}$ such that $\lim_{n\to\infty} \text{GENO}(n;1) = G^{(j)}$ and $\lim_{n\to\infty} \text{GENO}(n;2) = 1.5G^{(j)}$, where these GENOs correspond to (9) for a single experiment. In other words, in the j-th experiment, the first model with infinitely many observations (that is, a large number) is equivalent to the reference model with $G^{(j)}$ observations, and the second is equivalent to the reference 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 A.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 $r = 2, 3, 4$.

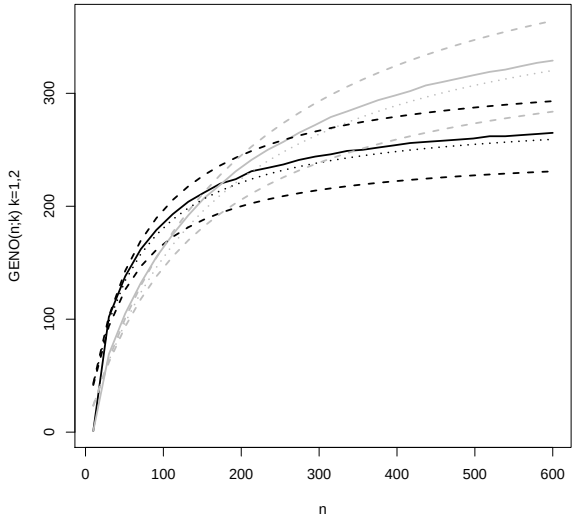
Figure 6 plots GENO$(n;k)$, and the mean bounds of the bootstrap confidence intervals and the true bounds computed from the simulation quantiles. The bootstrap confidence intervals are slightly conservative, as expected.



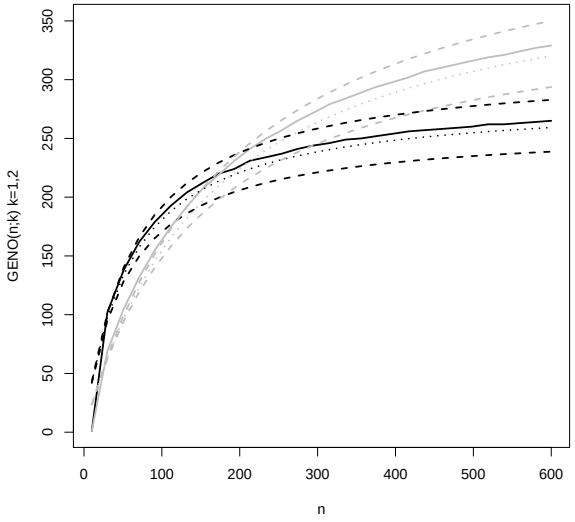
(a) $N = 10,000$



(b) $N = 20,000$

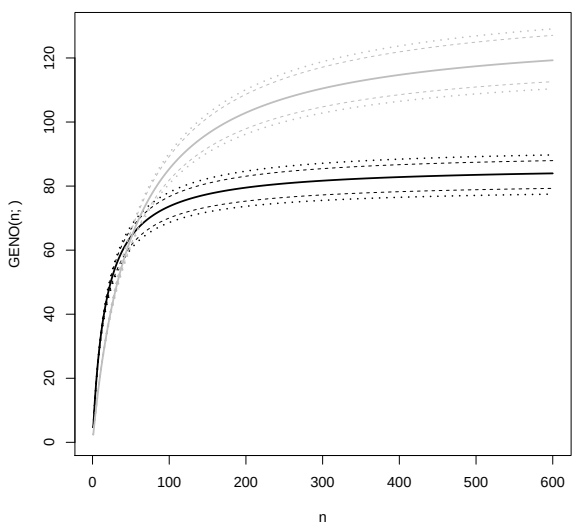


(c) $N = 50,000$

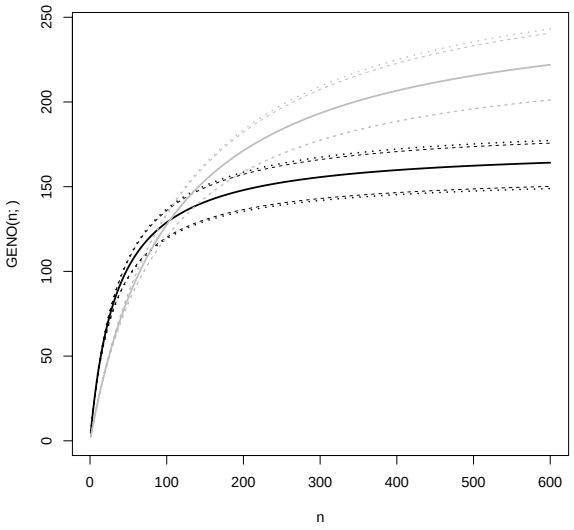


(d) $N = 100,000$

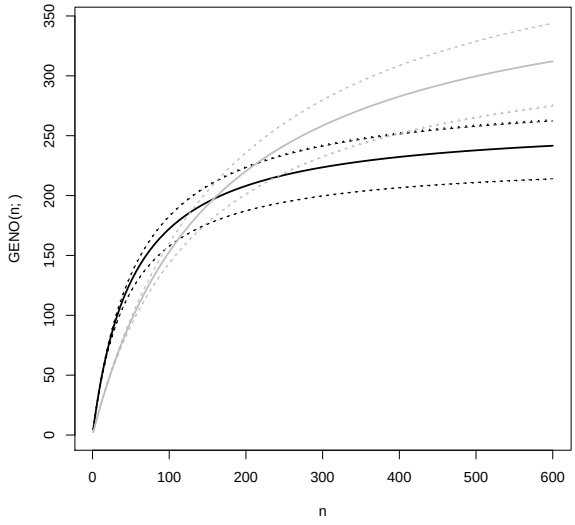
Figure 5: Plots of estimates of GENO$(n;1)$ (black) and GENO$(n;2)$ (gray). The dashed lines are bounds based on the 0.025,0.975 quantiles of the 10000 simulations of $\sum_{j=1}^{6} \widehat{p}_j \log[\widehat{p}_j / p_j^{(1)}(\widehat{\eta})]$ for Model 1 or $\sum_{j=1}^{6} \widehat{p}_j \log[\widehat{p}_j / p_j^{(2)}(\widehat{\boldsymbol{\theta}})]$ for Model 2. The solid line is GENO according to definition (4) and the dotted line is the approximation (9).



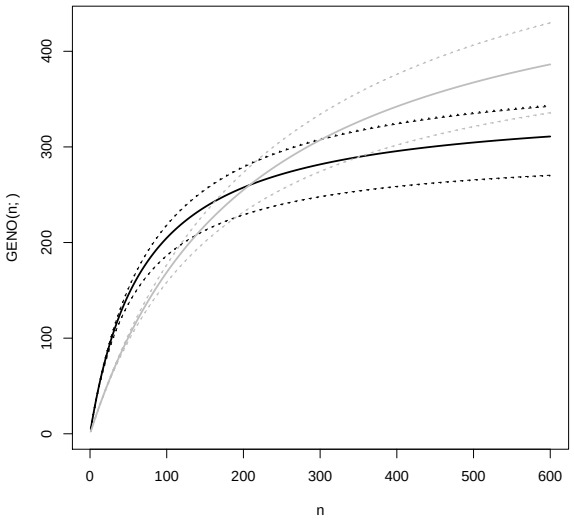
(a) $r = 1$



(b) $r = 2$



(c) $r = 3$



(d) $r = 4$

Figure 6: Plots of estimates of $\mathrm{GENO}(\cdot; 1), \mathrm{GENO}(\cdot; 2)$ (black, gray, respectively) for many experiments. The dotted lines are the mean bounds of the bootstrap 95% confidence interval, 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. The solid line is GENO defined in (12).

B Second Appendix: $\widehat{\text{GLU}}$ for the DNA data

Table 7: GLU(HW) for the different chromosomes for the populations ASW, CEU.

	ASW				CEU			
Chromosome	$\widehat{\text{GLU}}$(HW)	CI	$\mathcal{N}_c$	$J = \frac{\mathcal{N}_c}{20}$	$\widehat{\text{GLU}}$(HW)	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