האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

ON MEASURING AND COMPARING USEFULNESS OF STATISTICAL MODELS

By

DAVID AZRIEL and YOSEF RINOTT

Discussion Paper # 669 (October 2014)

מרכז פדרמן לחקר הרציונליות

THE FEDERMANN CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

On measuring and comparing usefulness of statistical models [1]

David Azriel[a] and Yosef Rinott[b]

October 24, 2014

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

Abstract

Statistical models in econometrics, biology, and most other areas, are not expected to be correct, and often are not very accurate. The choice of a model for the analysis of data depends on the purpose of the analysis, the relation between the data and the model, and also on the sample or data size. Combining ideas from Erev, Roth, Slonim, and Barron (2007) and the well-known AIC criterion and cross-validation, we propose a variant of model selection approach as a function of the models and the data size, with quantification of the chosen model's relative value. Our research is motivated by data from experimental economics, and we also give a simple biological example.

1 Introduction

This paper deals with model selection and model comparisons, when different models with different sample sizes are considered. Starting from G. Box's well-known 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 fact that the usefulness of a model depends not

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

only on the nature of the data, but also on the available sample size. According to the Oxford Online Dictionary a model is a "*a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions.*" Clearly, terms like "simplified" or "idealized," which are often used in defining models, reflect the well-known fact that models should not be precisely correct and, in fact, one should often prefer a simple, understandable model to a more precise but too complex one, especially given limited amounts of data, and this is particularly true when the goal of choosing a model involves prediction of future data.

Classical model selection methods, e.g., the AIC (Akaike, 1974) and its variations, and the BIC (Schwarz, 1978), deal with the question of selecting the best model from a class of models, having a sample of a given size. The AIC is based on a maximized 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. See also Burnham and Anderson (2002).

A meaningful measure of the value or usefulness of a model, 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. It was proposed in the context of comparing models' predictive value of players' strategies in game-theoretic experiments.

Roughly speaking, we consider the problem of estimating the proportion of playing a certain strategy by either modeling players' behavior or just using empirical sample proportions. A model's ENO is an estimate of the required sample size for which the sample proportion of playing a certain strategy in a game is equally accurate to the predictions based on 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). Our generalization goes in several directions. ENO quantifies the value of a model for predicting means, relative to empirical means. We generalize and allow the comparison of a model relative to any reference model, not just empirical frequencies. Furthermore, we generalize from predicting means to more general prediction, and we quantify prediction differently, using Akaike's AIC approach. Unlike ENO, our measure depends on the sample size n, which 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. Finally, our estimation of GENO is connected to the AIC approach in a

natural way.

We next describe GENO in terms of a simple example. Consider data modeled by the multinomial distribution with three cells having probabilities p_1, p_2, p_3 summing to one, and the Hardy–Weinberg (HW) model which 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 theoretically compare estimation of the p_i's on the basis of this model with n observations, to simple relative frequency estimators of the multinomial parameters p_i based on m observations from the true multinomial model; 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 parameters and make predictions without using such models. Comparing the HW model with n observations to the full multinomial model with m observations, the GENO of the HW with n observations is the value of m 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 relative frequency estimators are model-free, but this is not so. The multinomial model is indeed a model. It may serve as a natural "full" reference model as explained below, and, in general, GENO compares models in terms of a full model that is hopefully a natural reference model.

Given data of size N, our goal is to quantify and compare the predictive quality of different models to be used with a sample of n possibly other observations, generated by the same process. More specifically, given a candidate list of parametric models $\{f_k\} = \{f_k(y, \theta^{(k)})\}_{\theta^{(k)} \in \Theta^{(k)}}$, $k = 1, \ldots, K$, we define our measure of the quality of the model, GENO$(n; k)$, for each model $\{f_k\}$ and sample size n and discuss its estimation on the basis of N observations. Examples and further motivation are given below.

Like ERSB and most users of the AIC, we do not assume that any of our candidate models is "true." Rather, for different sample sizes, different models should be used, and GENO quantifies the predictive value of a model as a function of the sample size n, and provides a tool for model selection.

In Section 2 we describe GENO for samples of iid observations, and for non iid data from different experiments with a common model. We also discuss estimation of GENO, and confidence intervals. In Section 3 we discuss the theory and provide simulations for a multinomial example, and compute GENO for some Hardy–Weinberg-type models, which are of great importance in

genetics. In Section 4 we generalize GENO to decision processes, and apply the results to experimental data, with the goal of selecting useful models, and quantifying the quality of models for prediction of players' strategic choices in certain games.

2 GENO for iid observations

For simplicity we start with samples of iid observations.

2.1 Definition

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 the density could be with respect to any suitable measure. Several parametric 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

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

be the MLE for the k-th model based on n observations, and

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

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 $\{f_k\}$, $k = 1, \ldots, K$. As a reference model we shall consider a "high-dimensional" family, to be called the *full model*, $\{f\} = \{f(y, \theta)\}_{\theta \in \Theta}$, where $\Theta \subseteq \mathbb{R}^d$, and $d_k \leq d$ for $k = 1, \ldots, K$. Set $\hat{\theta}_n := \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(Y_i, \theta)$ to be the MLE of the family $\{f\}$ and $\theta_0 := \arg\max_{\theta \in \Theta} \int g(y) \log f(y, \theta) dy$ the corresponding projection. We shall sometimes assume that the full model contains all K 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 above and discussed again below, which contains models of the Hardy–Weinberg type, or a regression model where the full model contains a set of observed covariates, and the other K models contain different subsets of these variables.

We assume that the full model family $\{f\}$ is better than each of the K models in the KL divergence sense, that is,

$$\int_{-\infty}^{\infty} g(y) \log f(y, \theta_0) dy \geq \int_{-\infty}^{\infty} g(y) \log f_k(y, \theta_0^{(k)}) dy, \tag{1}$$

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 other models.

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 $\hat{\theta}_n^{(k)}$ based on the sample, that is, $f_k(\cdot, \hat{\theta}_n^{(k)})$. In the spirit of the AIC, and also its cross validation-version, (see Stone, 1977), 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, which in the normal case coincides with a sum of squares of deviations,

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

the expectation on the left is with respect to both the Y_i^*'s whose density is g and with respect to the MLE $\hat{\theta}_n^{(k)}$, and the expectation on the right is only with respect to the latter. In view of (2) the size of the Y^*'s sample does not play any role. In the spirit of ERSB we want to define a new notion, GENO$(n; k)$, that quantifies the usefulness of the k-th model with n observations relative to the full 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 (8). We define GENO$_1$ as the value m that satisfies $E \int_{-\infty}^{\infty} g(y) \log f(y, \hat{\theta}_m) dy \approx E \int_{-\infty}^{\infty} g(y) \log f_k(y, \hat{\theta}_n^{(k)}) dy$. In view of (2), 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 full dimensional model $\{f\}$ with m observations. More formally:

$$\text{GENO}_1(n; k) := \left\{ \max_m \ : \ E \int_{-\infty}^{\infty} g(y) \log f(y, \hat{\theta}_m) dy \leq E \int_{-\infty}^{\infty} g(y) \log f_k(y, \hat{\theta}_n^{(k)}) dy \right\}; \tag{3}$$

that is, GENO$_1$ is the largest value of m for which the k-th model with n observations is still better than the full dimensional model $\{f\}$ with m observations. In view of (1) 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. GENO$_1(n; k) < n$ means that the full model with n observations is better than the k-th model,

which suggests that the latter should not be used with n observations or more. 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 full model, but when we compare GENO_1 for two models, the full model cancels. Using GENO_1, one can compare different models with different sample sizes: a relation such as $\text{GENO}_1(50,1)$=$\text{GENO}_1(100,2)$, which by (3) does not depend on the full model, means that model 1 with $n = 50$ and model 2 with $n = 100$ are equally useful. For later reference we comment that in the same way, equality of two GENOs as defined by (8) below, or two GENO estimates defined by (9), does not depend on the choice of the full model.

2.2 Approximations from GENO_1 to GENO

In order to extract the value of m for which the two integrals in (3) are equal, and later estimate it, we approximate GENO_1. To start, we have

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

Standard AIC-type computations that involve a Taylor expansion similar to that appearing in the study of the asymptotic distribution of MLE's as in Burnham and Anderson (2002), Eq. (7.24), lead to

$$E\left[\int_{-\infty}^{\infty} g(y)\log f_k(y,\hat{\theta}_n^{(k)})dy - \int_{-\infty}^{\infty} g(y)\log f_k(y,\theta_0^{(k)})dy\right] = -\frac{Tr_k}{2n} + O(1/n^{3/2}), \tag{4}$$

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}. Also, when f_k is replaced by f in H, we denote its trace by Tr. This calculation can be found in Burnham and Anderson (2002), page 367. 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 information relations both matrices appearing in H (the first matrix before taking inverse) express the 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, that is, models that get close to the true g. We adopt this approach, which is close in its spirit to the notion of Pitman efficiency (see, e.g., van der Vaart, 1998, Chapter 14). Combining the result of (4) and a similar calculation applied to the family $\{f\}$ instead of $\{f_k\}$ with d replacing d_k, we obtain

$$\mathrm{GENO}_1(n;k) = \left\{ \max_m : \frac{Tr}{2m} + O(\frac{1}{m^{3/2}}) \geq \int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy + \frac{Tr_k}{2n} + O(\frac{1}{n^{3/2}}) \right\}. \tag{5}$$

Dropping the smaller order terms we approximate GENO_1 by

$$\mathrm{GENO}_2(n;k) := \left\{ \max_m \ : \ \frac{Tr}{2m} \geq \int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy + \frac{Tr_k}{2n} \right\}. \tag{6}$$

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

$$\mathrm{GENO}_3(n;k) := \left\{ \max_m \ : \ \frac{d}{2m} \geq \int_{-\infty}^{\infty} g(y) \log \left[f(y,\theta_0) \Big/ f_k(y,\theta_0^{(k)}) \right] dy + \frac{d_k}{2n} \right\}. \tag{7}$$

For small m the left-hand side is larger than the positive constant on the right-hand side and as m goes to infinity the left-hand side vanishes. We aim at the critical m for which equality holds approximately, and thus we finally define

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

We have that

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

This ratio converges to 1 if $Tr - d$ and $Tr_k - d_k$ converge 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. The approximation $\mathrm{GENO}_1(n;k) \approx \mathrm{GENO}_2(n;k)$ is justified by similar arguments (for details, see the Appendix). In the sequel we will work with GENO as given by (8), 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 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, the N observations may arise from pooling data from many subjects and we then want to make predictions for a new subject on the basis of this subject's n observations, assuming that the choice of the model can be determined by the training sample. We then choose the model with the highest $\mathrm{GENO}(n;k)$ and estimate its parameters by the sample of size n for this subject. Here we assume for simplicity that the data in the training sample all comes from the same g (at least to a reasonable approximation). This assumption is relaxed in Section 2.5.

We now estimate $\mathrm{GENO}(n;k)$ on the basis of a sample $Y_1, \ldots, Y_N$ from g. First note that

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

Again, by standard AIC-type calculations, the expectation of the square brackets is approximated by $\frac{d_k}{2N}$ (Burnham and Anderson, 2002, eq. 7.29), and by a similar argument for the family $\{f\}$, an approximately unbiased estimate to the integral appearing in the denominator of (8), $\int_{-\infty}^{\infty} g(y) \log \left[f(y, \theta_0) \Big/ f_k(y, \theta_0^{(k)}) \right] dy$ is

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

Summing up, we define our estimator as

$$\widehat{\mathrm{GENO}}(n;k) := \frac{d/2}{\frac{1}{N} \sum_{i=1}^{N} \log \left[f(Y_i, \hat{\theta}_N) \Big/ f_k(Y_i, \hat{\theta}_N^{(k)}) \right] - \frac{d-d_k}{2N} + \frac{d_k}{2n}}. \tag{9}$$

We have in mind two scenarios. One is that the full model f and the model f_k are fixed. In the other, in the spirit of the notion of Pitman's efficiency we consider, for example, the case that as the sample size increases, f_k approaches f. Consider the first case. The above calculations, including the replacement of the traces by dimensions shows that the bias of the denominator of (9) as an estimator of the denominator of (8) is of order 1/N. A standard delta method (Taylor approximation) shows that the bias of $\widehat{\mathrm{GENO}}(n;k)$ is of order $1/N$. A similar approximation shows that $\mathrm{Var}(\widehat{\mathrm{GENO}}(n;k))$ is also of order $1/N$ and therefore so is the mean square error. Note, however, that this would be true even without the bias correction $\frac{d-d_k}{2N}$ which becomes relevant only if f_k is close to, or, more formally, approaches f, corresponding to the second (Pitman) scenario.

2.4 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 estimator of $\mathrm{GENO}(n;k)$. If another model that is preferable for some technical or aesthetical reasons has a GENO that is close to the largest, it may be chosen, and GENO provides a measure of how much is lost. Examples are given in Section 3.

The estimator $\widehat{\mathrm{GENO}}(n;k)$ of (9) quantifies the value of the k-th model with n observations, and allows comparisons of different models with different sample sizes. By (9), 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 $-\frac{1}{N}\sum_{i=1}^{N}\log\left[f_k(Y_i,\hat{\theta}_N^{(k)})\right]+d_k(\frac{1}{2n}+\frac{1}{2N})$. 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

$$AIC := -\sum_{i=1}^{n} f_k(Y_i,\hat{\theta}_n^{(k)}) + d_k. \tag{10}$$

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

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

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

2.5 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 and compared. 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 (9) 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 j-th experiment. For the j-th experiment, the likelihood at the MLE of the k-th model is $f_k(y,\hat\theta^{(k)}_{N_j,j})$, 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 that deserve 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 in the model f with m observations in each experiment, which is equivalent to having n observations in each experiment with f_k.

Thus, the first definition of GENO is

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

it is redefined, as in (8), by

$$\mathrm{GENO}(n;k) := \frac{d/2}{\frac{1}{J}\int_{-\infty}^{\infty}\sum_{j=1}^J g_j(y)\log\left[f(y,\theta_{0,j})\Big/ f_k(y,\theta^{(k)}_{0,j})\right]dy + \frac{d_k}{2n}}, \tag{11}$$

and is estimated by

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

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 model, and maximizing $\widehat{\mathrm{GENO}}(n;k)$ with respect to k amounts to maximizing the corrected empirical likelihood over the k models.

2.6 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 (9) using standard jackknife or bootstrap methods for iid observations. Then one can 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 $Z_j := \frac{1}{N_j}\sum_{i=1}^{N_j} \log\left[f(Y_{i,j}, \hat{\theta}_{N_j,j}) \Big/ f_k(Y_{i,j}, \hat{\theta}^{(k)}_{N_j,j})\right]$ from (12) in order to construct a confidence interval under the assumption that the experiments are independent.

Since we are later interested in the case of many experiments for non-iid data, 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 is 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

$$\left(\frac{d/2}{\tilde{F}_J^{-1}(\alpha/2) - \frac{1}{J}\sum_{j=1}^{J}\frac{d-d_k}{2N_j} + \frac{d_k}{2n}},\ \frac{d/2}{\tilde{F}_J^{-1}(1-\alpha/2) - \frac{1}{J}\sum_{j=1}^{J}\frac{d-d_k}{2N_j} + \frac{d_k}{2n}}\right), \tag{13}$$

where $\tilde{F}_J$ is the bootstrap distribution of $\frac{1}{J}\sum_{j=1}^{J} Z_j$. This approach is tested in Section 3.4. The results indicate that this method works well for large J, even if the means 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 (12) do not have the same means, that is, the μ_i's are different (in Liu's notation), the confidence intervals are conservative.

3 An example: the multinomial case

Goodness-of-fit tests are related to the multinomial distribution, with which we start to demonstrate the notion of GENO. 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 $Y_\ell = \#\{i : X_i = a_\ell\}$; $\ell = 1, \ldots, L$. For the full model $\{f\}$ with a parameter p we take $Y = (Y_1, \ldots, Y_L) \sim$ Multinomial(n, p), where $p = (p_1, \ldots, p_L)$. The true model g will also be multinomial and hence contained in the family $\{f\}$ 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 $p = p(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$.

3.1 Hardy–Weinberg models

We discuss in our context a classical model that plays a prominent role in genetics, the Hardy–Weinberg model, considered here 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 1.

Table 1: 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

$$\hat{\theta}_1 = \frac{2Y_1 + Y_2 + Y_5}{2n},\ \hat{\theta}_2 = \frac{Y_2 + 2Y_3 + Y_4}{2n};\ \hat{\eta} = \frac{2(Y_1 + Y_2 + Y_3) + Y_4 + Y_5}{4n}. \tag{14}$$

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 $p^{(1)} = p^{(1)}(\eta) = \left(\eta^2, 2\eta^2, \eta^2, 2\eta(1-2\eta), 2\eta(1-2\eta), (1-2\eta)^2\right)$ or $p^{(2)} = p^{(2)}(\theta) = \left(\theta_1^2, 2\theta_1\theta_2, \theta_2^2, 2\theta_2\theta_3, 2\theta_1\theta_3, \theta_3^2\right)$ under Model 1 or Model 2, respectively.

3.2 A numerical example

We consider a specific multinomial distribution; the probabilities p under Models 1, 2, and under the full model are presented in Table 2.

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

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

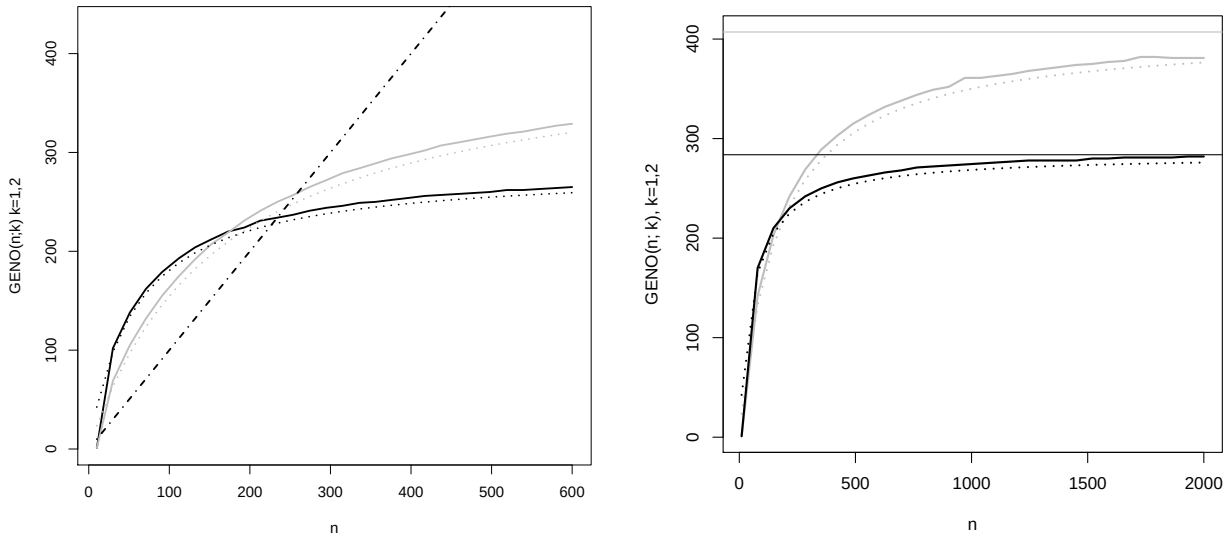
In this example we assume that the full model is also the true model. In this case, the components $p_j^{(1)}(\eta_0)$ and $p_j^{(2)}(\theta_0)$ of $p^{(1)}(\eta_0)$ and $p^{(2)}(\theta_0)$, respectively, are computed similarly to (14), with Y_ℓ/n replaced by p_ℓ from the full model.

For the full model (multinomial) we have $d = 5$, and according to (8)

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

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

Figure 1 compares the function $\mathrm{GENO}_1(n;k)$, given by (3), and $\mathrm{GENO}(n,k)$, as defined in (8), 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 full model is preferred.



(a) (b)

Figure 1: 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 (3) (respectively, (8)). The horizontal lines represent the limit as n goes to infinity. The dot-dashed line is $\mathrm{GENO}(n;)$ of the full model, which is equal to n.

3.3 Estimation

We now consider the estimator (9) and study its behavior under p of the previous section for different values of N. For a sample $(Y_1, \cdots, Y_6) \sim \text{Multinomial}(N, p)$, (9) reads as

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

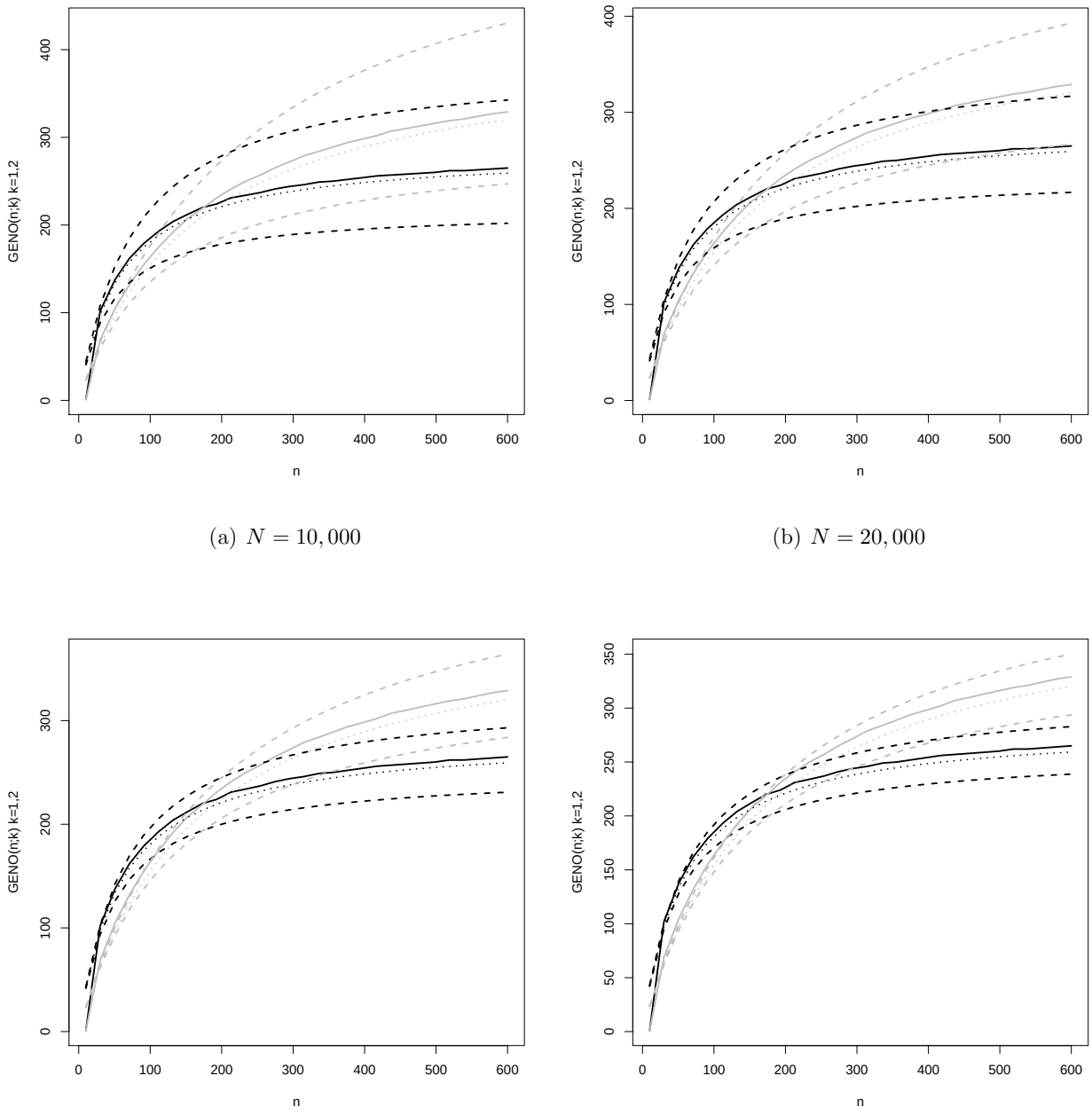
where the MLE estimators are given by (14) and $\hat{p}$ is the empirical mean.

Figure 2 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 $\sum_{\ell=1}^{6} \hat{p}_\ell \log[\hat{p}_\ell / \{p^{(1)}(\hat{\eta})\}_\ell]$ for Model 1 and $\sum_{\ell=1}^{6} \hat{p}_\ell \log[\hat{p}_\ell / \{p^{(2)}(\hat{\theta})\}_\ell]$ for Model 2. For small N, the confidence intervals of $\text{GENO}(n;1)$ and $\text{GENO}(n;2)$ 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.

3.4 Many experiments

We now consider the scenario of Section 2.5, namely, that there are many experiments, each of which has a different p. Based on these experiments we would like to estimate the common GENO as approximated by (11). We consider that 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 $J = 200$ experiments, each of size $N = 500$. 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 $N(100, 25^2)$ distribution, so that we have 200 GENOs that are roughly of the same order. Then we chose, by solving a non-linear equation, $p^{(j)}$ 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 (8) 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 full model with $G^{(j)}$ observations, and the second is equivalent



(a) $N = 10,000$ (b) $N = 20,000$

(c) $N = 50,000$ (d) $N = 100,000$

Figure 2: 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 $\sum_{j=1}^{6} \hat{p}_j \log[\hat{p}_j/p_j^{(1)}(\hat{\eta})]$ for Model 1 or $\sum_{j=1}^{6} \hat{p}_j \log[\hat{p}_j/p_j^{(2)}(\hat{\boldsymbol{\theta}})]$ for Model 2. The solid line is GENO according to definition (3) and the dotted line is the approximation (8).

to the full model with $1.5G^{(j)}$ observations. The factor 1.5 was chosen since it is close to the above 407.1/283.8=1.43.

We computed 95% bootstrap confidence intervals by (13) 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 $N(100\ell, (25\ell)^2)$ distribution with $\ell = 1$ (the case above) and also $\ell = 2, 3, 4$.

Figure 3 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.

4 GENO for decision processes

The initial impetus for this work comes from Erev, Roth, Slonim, and Barron (2007) from experimental game theory. Their observations, described in detail in Section 5, are not iid and involve decisions or strategies, and we adapt GENO to this situation.

4.1 The set-up

Let $Z_1, \ldots, Z_n$ be decisions 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 $t+1$ of the process, the information available to the decision maker (player) towards making the next decision is

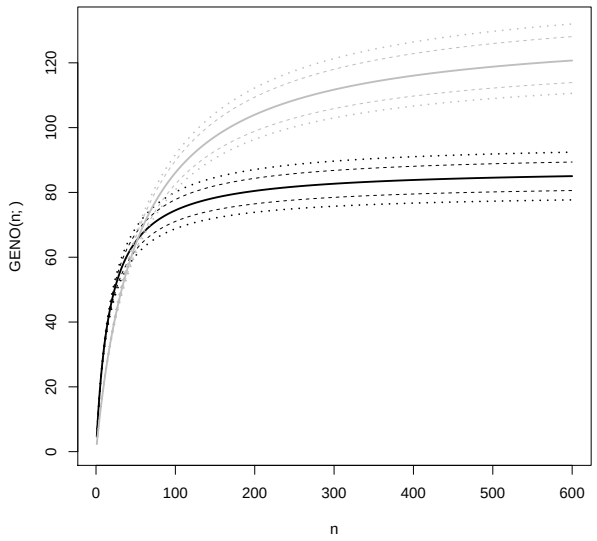
$$\mathcal{D}_t := (Z_1, \ldots, Z_t, V_1, \ldots, V_t). \tag{15}$$

The decision process is determined by a mixed strategy having probability $p_{\mathcal{D}_{t-1}}(z_t)$ of making the decision z_t at time t on the basis of $\mathcal{D}_{t-1}$ for $t = 1, 2, \ldots$. The likelihood of the player's actions is

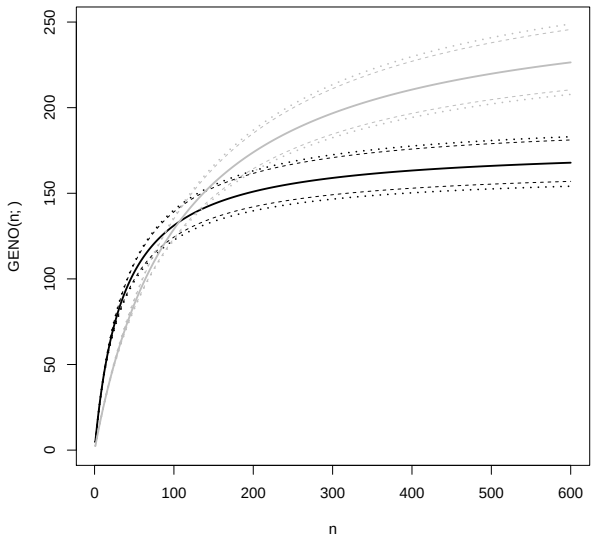
$$L(z_1, \ldots, z_n) = L(z_n|\mathcal{D}_{n-1})L(z_1, \ldots, z_{n-1}) = p_{\mathcal{D}_{n-1}}(z_n)L(z_1, \ldots, z_{n-1}) = \cdots = \prod_{t=1}^{n} p_{\mathcal{D}_{t-1}}(z_t). \tag{16}$$

A given player has a true strategy g, that is, $p_{\mathcal{D}_{t-1}}(z_t) = g(z_t; \mathcal{D}_{t-1})$.

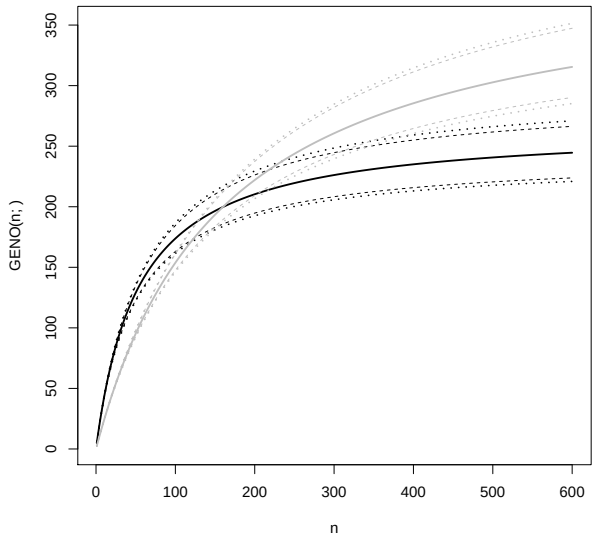
Consider K models for such a mixed strategy, where for $k = 1, \ldots, K$, $p_{\mathcal{D}_t}(z_t)$ is modeled by $f_k(z_t; s_t^{(k)}(\mathcal{D}_{t-1}), \theta^{(k)})$, where f_k and $s_t^{(k)}$ are known functions and the parameter $\theta^{(k)} \in \Theta^{(k)} \subseteq \mathbb{R}^{d_k}$. As usual, 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. Furthermore, we assume that under the true strategy played, the sequence $s_t^{(k)}(\mathcal{D}_{t-1})$ has a stationary distribution, whose density is denoted by



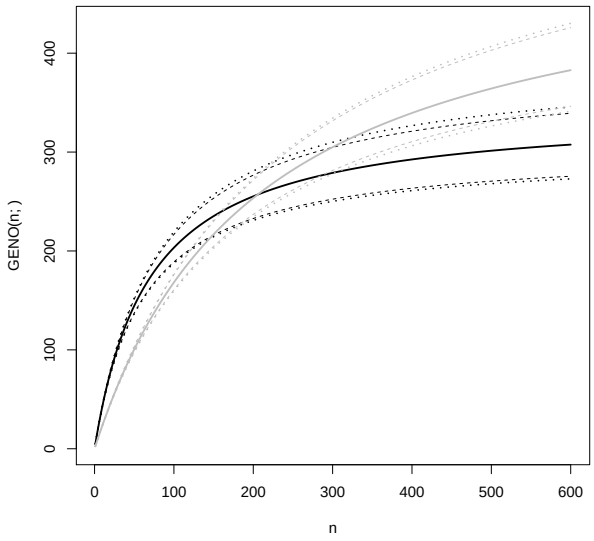
(a) $\ell = 1$



(b) $\ell = 2$



(c) $\ell = 3$



(d) $\ell = 4$

Figure 3: 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 (11).

q_k. This holds if, for example, the true strategy and the model are Markovian with possibly more than one step. It is well known that the Markov property can often be achieved by appropriately defining the state space. By (16), the log-likelihood of $z_1, \ldots, z_n$ under the k-th model for a given player becomes

$$\sum_{t=1}^{n} \log f_k(z_t; s_t^{(k)}(\mathcal{D}_{t-1}), \theta^{(k)}). \tag{17}$$

Let $\hat{\theta}_n^{(k)}$ be the MLE, i.e., the maximizer of (17), based on n observations. In the spirit of GENO above, imagine that a new player is playing the same game with the same strategies n^* times; let Z_t^* and $\mathcal{D}_t^*$ be the decision and the available data at time t. One wants to maximize the expected log-likelihood of the model based on the MLE, i.e., to maximize

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

where $\pi(z)$ is the stationary probability that $Z = z$, whose existence is assumed. It is known to exist if, for example, $s_t^{(k)}(\mathcal{D}_{t-1})$ is a function of the information from some time $t-1-M$ to t for certain M; that is, it is M-step Markovian, and so is the true strategy. We also consider a full model, which is high-dimensional. For example if our models are M-Markovian with some values of M, then the full model may be chosen as Markovian with the maximal M or larger. As before, we denote quantities of the full model in the same way as those for the the k-th model, but without the index.

4.2 Definition, approximation, and estimation of GENO

Similar to the iid case, GENO_1 of the k-th model is defined by

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

where f, $\hat{\theta}_m$, and q pertain to the full model. As in the iid case, we use the approximation

$$E \left[\sum_{z \in \mathcal{Z}} \pi(z) \int q_k(u) \log f_k \left(z; u, \hat{\theta}_n^{(k)}\right) du - \sum_{z \in \mathcal{Z}} \pi(z) \int q_k(u) \log f_k \left(z; u, \theta_0^{(k)}\right) du \right] \approx -\frac{d_k}{2n}, \tag{18}$$

where

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

As in (4), unless the model f_k is the true model, we get a trace rather than the dimension in (18); see Billingsley (1961). The approximation is justified when in addition to the existence of the stationary distribution q_k, we assume that the models considered are good enough so that the trace can be replaced by the dimension, as above.

Similarly, given data $\mathcal{D}_N$ on the process up to a certain time N, we assume that

$$E\Big[\sum_{t=1}^{N}\log f_k(Z_t;s_t^{(k)}(\mathcal{D}_{t-1}),\hat{\theta}_N^{(k)})-\sum_{z\in\mathcal{Z}}\pi(z)\int q_k(u)\log f_k\left(z;u,\theta_0^{(k)}\right)du\Big]\approx\frac{d_k}{2N}. \tag{19}$$

We assume further that (18) and (19) hold also for the full model f. Such approximations were made in Markov models by Tong (1975) in the context of AIC computations of M-step Markov chains. Under (18), GENO can be redefined by

$$\mathrm{GENO}(n;k):=\frac{d/2}{\sum_{z\in\mathcal{Z}}\pi(z)\left\{\int q(u)\log f(z;u,\theta_0)du-\int q_k(u)\log f_k(z;u,\theta_0^{(k)})du\right\}+\frac{d_k}{2n}}, \tag{20}$$

and (19) implies that the estimate of GENO is

$$\widehat{\mathrm{GENO}}(n;k):=\frac{d/2}{\frac{1}{N}\sum_{n=1}^{N}\log\left\{f(Z_t;s_t(\mathcal{D}_{t-1}),\hat{\theta}_N)\Big/f_k(Z_t;s_t^{(k)}(\mathcal{D}_{t-1}),\hat{\theta}_N^{(k)})\right\}-\frac{d-d_k}{2N}+\frac{d_k}{2n}}. \tag{21}$$

When the data comes from several players, this is generalized as in Section 2.5, equations (11) and (12), and 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 each game is played by 9 of these pairs, 500 times. In each two-player game, each player chooses an action; these choices determine the probabilities of winning a fixed amount or zero.

5.1 GENO for the motivating data

The choices of player 1 and player 2 at the t-th repetition of the game are X_t, Y_t, which are sequences of 0's and 1's since there are two possible actions in each game, denoted by 0 and 1. For $t = 1, 2, \ldots$ we set $Z_t = (X_t, Y_t)$. In the notation of Section 4 we view the pair of players together

as a single decision maker. Similarly, the two rewards to the players in each repetition are denoted by the vector V_t. Here $N = 500$. The likelihood for a given pair of the players' actions is

$$\prod_{t=1}^{N}(p^X_{\mathcal{D}_{t-1}})^{X_t}(1-p^X_{\mathcal{D}_{t-1}})^{1-X_t}(p^Y_{\mathcal{D}_{t-1}})^{Y_t}(1-p^Y_{\mathcal{D}_{t-1}})^{1-Y_t},$$

where $p^X_{\mathcal{D}_{t-1}}$ and $1-p^X_{\mathcal{D}_{t-1}}$ are the probabilities that the first player plays strategy 1 or 0, and likewise for $p^Y_{\mathcal{D}_{t-1}}$ for the second player, and $\mathcal{D}_t$ is defined as in (15).

Consider K models for strategies for the X-player (= player 1), where for the k-th model, $p^X_{\mathcal{D}_{t-1}} = h^X_k(s^{(k)}_t(\mathcal{D}_{t-1}), \alpha^{(k)})$, where h^X_k and $s^{(k)}_t$ are known functions and $\alpha^{(k)}$ is a parameter. With β replacing α, a similar notation is used for Y, and we set $\theta^{(k)} = (\alpha^{(k)}, \beta^{(k)}) \in \Theta^{(k)} \subseteq \mathbb{R}^{d_k}$.

For any u in the image of $s^{(k)}_t$ and $z = (x, y) \in \{0, 1\}^2$ define,

$$f_k(z; u, \theta^{(k)}) := \{h^X_k(u; \alpha^{(k)})\}^x \left\{1 - h^X_k(u; \alpha^{(k)})\right\}^{1-x} \{h^Y_k(u; \beta^{(k)})\}^y \left\{1 - h^Y_k(u; \beta^{(k)})\right\}^{1-y}.$$

Under this notation the log-likelihood under the k-th model coincides with (17).

For a single pair of players, playing N times, GENO(n, k) and its estimate are now defined by (20) and (21).

5.2 The models

Following ERSB we restrict attention to models that assume the same model of strategy for both players; however we allow different values of parameters for different players. Therefore, it is enough to describe the strategy of the X-player. The following models are considered:

1. Reinforcement learning (RL): the X-player chooses action 1 at round t with probability

$$p^X_{\mathcal{D}_{t-1}} = \frac{q^X_1(t)}{q^X_1(t) + q^X_0(t)},$$

where $q^X_i(t)$ is referred to as the propensity to select action i. The propensities are updated at each stage according to

$$q^X_i(t) = \begin{cases} (1-w)q^X_i(t-1) + w\nu_{t-1} & \text{if } X_{t-1} = i \\ q^X_i(t-1) & \text{if } X_{t-1} \neq i \end{cases}, \quad i = 0, 1, \tag{22}$$

where $0 < w < 1$ is a parameter of the model, and ν_{t-1} is player 1's previous reward. Initially $q^X_0(1) = q^X_1(1)$ are equal to the expected payoff when both players choose each strategy with equal probability. In the notation of the previous section, we have for this

model $s_t(\mathcal{D}_{t-1}) = (\mathcal{I}^X, \mathcal{I}^Y)$ where $\mathcal{I}^X := \left(q_1^X(t-1), q_0^X(t-1), \nu_{t-1}, X_{t-1}\right)$ and $\alpha^{(k)} = w$, and similarly for the Y-player.

2. Reinforcement learning lambda (RLL): at round t the X-player chooses action 1 with probability
$$p^X_{\mathcal{D}_{t-1}} = \frac{\lambda + q_1^X(t)}{2\lambda + q_1^X(t) + q_0^X(t)},$$
where $q_i^X(t)$ are updated as in (22) and $\lambda > 0$ is an unknown parameter. When λ is large, $p^X_{\mathcal{D}_{t-1}} \approx \frac{1}{2}$, and the propensities are down weighted.

3. Reinforcement learning stickiness (RLS): at round t the X-player chooses action 1 with probability
$$p^X_{\mathcal{D}_{t-1}} = (1-\xi)\frac{q_1^X(t)}{q_1^X(t) + q_0^X(t)} + \xi X_{n-1},$$
$0 < \xi < 1$ is a "stickiness" parameter; when ξ is close to 1 the player repeats his choice with high probability.

4. Toss: at each round, the X-player chooses action 1 with probability p independently of previous rounds.

5. Nash: at each round, the X-player chooses action 1 with the 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., $X_{t-M}, \ldots, X_{t-1}$ and the last reward ν_{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, which indicate the probability of choosing action 1 for each such sequence, and the dimension (i.e., d_k) is $2 \cdot 2^{M+1}$. We consider $M = 1, 2, 3$.

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 (10). The results are given in Table 3. We find that the M-step Markov models are preferred according to the AIC criterion. 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 full model.

Table 3: The mean (standard deviation) of the log-likelihood and the AIC number over the 90 pairs.

Model	d_k	log-likelihood	AIC
RL	2	-613.8 (67.9)	615.8 (67.9)
RLL	4	-588.4 (74.5)	592.4 (74.5)
RLS	4	-504.1 (133.1)	508.1 (133.1)
Toss	2	-567.0 (121.3)	569.0 (121.3)
Nash	0	-721.7 (184.4)	721.7 (184.4)
1-step Markov	8	-459.7 (126.6)	467.7 (126.6)
2-step Markov	16	-443.4 (126.3)	459.4 (126.3)
3-step Markov	32	-427.5 (124.7)	459.5 (124.7)

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 much smaller AIC numbers and therefore are preferred.

5.3 GENO: results

Table 4 shows $\widehat{\text{GENO}}(n,k)$ for different n's and for the models mentioned in the previous section. Confidence intervals at the 95% level based on (13) are also provided.

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 and smaller than approximately 510, 2-step Markov is preferred. For larger n, M-step Markov with $M = 3$, which is defined as the full model, is the best.

Table 4: The 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 (40.6,52.2)
RLL	47.9 (42.3,55)	51 (44.6,59)	52.1 (45.5,60.5)
RLS	96.9 (88,107.4)	110.2 (98.9,124)	115.5 (103.2,130.7)
Toss	59.5 (51.5,69.4)	61.8 (53.2,72.5)	62.6 (53.8,73.6)
Nash	28.8 (25.1,33.2)	28.8 (25.1,33.2)	28.8 (25.1,33.2)
1-step Markov	**132.8 (123.9,142.2)**	**198.9 (179.4,220.7)**	**238.4 (210.9,270.4)**
2-step Markov	91.1 (88.6,93.2)	167.2 (159.1,174.7)	231.8 (216.4,246.4)
Model k	GENO$(200, k)$	GENO$(400, k)$	GENO$(600, k)$
RL	46 (40.8,52.5)	46.4 (41,52.9)	46.5 (41.1,53)
RLL	52.7 (45.9,61.3)	53.6 (46.6,62.5)	53.9 (46.8,62.9)
RLS	118.4 (105.4,134.4)	122.9 (109 , 140.3)	124.5 (110.3,142.4)
Toss	63 (54.1,74.2)	63.6 (54.6,75.1)	63.8 (54.7,75.4)
Nash	28.8 (25.1,33.2)	28.8 (25.1,33.2)	28.8 (25.1,33.2)
1-step Markov	264.7 (231.3,304.8)	317.2 (270.3,376.5)	339.6 (286.5,408.6)
2-step Markov	**287.4 (264.1,310)**	**448.4 (394.2,506.2)**	**551.5 (471.7,641.5)**

Appendix: The approximation $\mathrm{GENO}_1(n;k) \approx \mathrm{GENO}_2(n;k)$

$\mathrm{GENO}_1(n;k)$ is the maximal solution in m of

$$\frac{Tr}{2m} + \frac{C_m}{m^{3/2}} = \int_{-\infty}^{\infty} g(y) \log\left[f(y,\theta_0)\Big/ f_k(y,\theta_0^{(k)})\right] dy + \frac{Tr_k}{2n} + \frac{D_n}{n^{3/2}} := A,$$

where $|C_m| \le C,\ |D_n| \le D$ for some constants C, D, and

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

Define $\tilde{G} := \frac{Tr/2}{A}$. We show below that $\mathrm{GENO}_1(n;k) \approx \tilde{G}$ and $\tilde{G} \approx \mathrm{GENO}_2(n;k)$ for large GENO's.

The approximation $\mathrm{GENO}_1(n;k) \approx \tilde{G}$.

Set $G_1 := \mathrm{GENO}_1(n;k)$. Consider the linear function $f(m) := Am - Tr/2$. We have that $f(\tilde{G}) = 0$ and $f(G_1) = \frac{C_m}{G_1^{1/2}}$. For any m_1, m_2 we have that $\frac{f(m_2)-f(m_1)}{m_2-m_1} = A$. Therefore,

$$\frac{f(G_1) - f(\tilde{G})}{G_1 - \tilde{G}} = \frac{-C_m/G_1^{1/2}}{G_1 - \tilde{G}} = A.$$

Since $\tilde{G} = \frac{Tr/2}{A}$, we obtain $G_1 - \tilde{G} = \frac{-C_m}{Tr/2}\frac{\tilde{G}}{G_1^{1/2}}$. Hence,

$$\frac{G_1}{\tilde{G}} = \frac{G_1 - \tilde{G} + \tilde{G}}{\tilde{G}} = \frac{-C_m/(Tr/2)\cdot \tilde{G}/G_1^{1/2} + \tilde{G}}{\tilde{G}} = 1 - \frac{-C_m/(Tr/2)}{G_1^{1/2}}.$$

We have that C_m is bounded and $Tr/2 \approx d/2$ is bounded from below; therefore, $\frac{G_1}{\tilde{G}} \approx 1$ for large G_1.

The approximation $\mathrm{GENO}_2(n;k) \approx \tilde{G}$.

Set $G_2 = \mathrm{GENO}_2(n;k)$. By definition,

$$\frac{G_2}{\tilde{G}} = \frac{\int_{-\infty}^{\infty} g(y) \log\left[f(y,\theta_0)\Big/ f_k(y,\theta_0^{(k)})\right] dy + \frac{Tr_k}{2n} + \frac{D_n}{n^{3/2}}}{\int_{-\infty}^{\infty} g(y) \log\left[f(y,\theta_0)\Big/ f_k(y,\theta_0^{(k)})\right] dy + \frac{Tr_k}{2n}} = 1 + \frac{D_n/(Tr/2)}{n^{3/2}} G_2;$$

further,

$$\frac{G_2}{n} = \frac{Tr/2}{n\int_{-\infty}^{\infty} g(y) \log\left[f(y,\theta_0)\Big/ f_k(y,\theta_0^{(k)})\right] dy + Tr_k/2} \le \frac{Tr}{Tr_k}.$$

Summing up,

$$\left|\frac{G_2}{\tilde{G}} - 1\right| \le \frac{D/(2Tr_k)}{n^{1/2}},$$

and when $Tr_k \approx d_k$ is bounded from below, $\frac{\tilde{G}}{G_2}$ converges to 1 as n goes to infinity.

Acknowledgement: This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 1474/10). 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. (1961). 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.

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.

Schwarz, Gideon E. (1978) Estimating the dimension of a model. *Annals of Statistics* **6** 461–464.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B* **39** 44–47.

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