APPLICATION AND VALIDATION OF GENETIC ALGORITHMS FOR PARAMETER ESTIMATION IN SELECTED STATISTICAL AND EPIDEMIOLOGIC EXAMPLES

Greiner M[1], Selhorst T[2]

[1]Institute for Parasitology and Tropical Veterinary Medicine, Department of Tropical Veterinary Medicine and Epidemiology, Freie Universität Berlin, Berlin, Germany, Koenigsweg 67, D-14163 Berlin
[2] Federal Research Centre for Virus Diseases of Animals, Institute for Epidemiology, Wusterhausen/Dosse, Germany, Seestr.55 D-16868 Wusterhausen/D.

Genetic algorithms (GAs) were introduced by Holland[1] in 1975 as a new class of parameter-search techniques that mimic evolutionary principles (e.g., heredity, mutation, cross-over, selection according to fitness scores). A principal advantage of GAs is their applicability for multi-dimensional parameter spaces and multimodal likelihood functions, in which classical algorithms often fail to converge to a global maximum. One objective of our study was to apply GAs to well-known estimation problems and thus to validate the results obtained by GAs against standard numeric or iterative solutions. The second objective was to apply GAs to situations in which no standard solutions exist.

Materials & Methods

The GAs were realized using Evolver™ (version 4.0, Palisade™) (examples 1–5) and a C++ program developed by Selhorst (1999 unpublished) (example 6).

(1) Linear regression
Least squares estimates for the parameters of the linear regression model $Y_i = \boldsymbol{b}_0 + \boldsymbol{b}_1 X_\mathrm{i} + e_\mathrm{i}$ are to be found using the paired observations $\{X_i, Y_i\}$, $i = 1, \ldots 11$. Standard approach (in matrix notation): $\underline{\boldsymbol{b}} = (\underline{X}'\underline{X})^{-1}\ \underline{X}'\underline{Y}$. GA approach: Find values for $\boldsymbol{b}_0$ and $\boldsymbol{b}_1$ that minimize the optimality criterion $\mathrm{S} = \Sigma_\mathrm{i}\ (Y_\mathrm{i} - \boldsymbol{b}_0 + \boldsymbol{b}_1 X_\mathrm{i})^2$.

(2) Exponential function
We were interested in finding zero of $y(x) = x^2/2 + \exp(-ax)$. Standard approach: For illustration, we skipped the numeric solution and used a Newton-Raphson and Random-walk algorithm to find x such that $y'(x) = 0$. GA approach: Find x such that $y(x)$ is minimized.

(3) Odds ratio estimation with misclassified data
Consider a 2x2 table for analysis of the association between risk factor (F) and disease (D) under non-differential misclassification with known sensitivity and specificity for diagnosis of F and D, given Se_F=0.9, Sp_F=1, Se_D=0.7 and Sp_D=0.85.

The odds ratio adjusted for misclassification (OR_{a}) is to be estimated using the observed table values $\underline{m}$' = [a, b, c, d] = [68, 49, 55, 59]. Standard approach: $OR_{\mathrm{a}} = (a_{\mathrm{a}}d_{\mathrm{a}})/(b_{\mathrm{a}}c_{\mathrm{a}}) = (96.45\times41.50)/(25.50\times 67.55)=2.32$, where $[a_{\mathrm{a}}, b_{\mathrm{a}}, c_{\mathrm{a}}, d_{\mathrm{a}}] = \underline{x}'$, $\underline{x} = \underline{A}^{-1}\underline{m}$ and

$$\underline{A} = \begin{bmatrix} \mathrm{Se}_{\mathrm{D}}Se_F & (1-Sp_D)Se_F & \mathrm{Se}_{\mathrm{D}}(1-Sp_F) & (1-Sp_D)(1-Sp_F) \\ (1-Se_D)Se_F & Sp_DSe_F & (1-Se_D)(1-Sp_F) & Sp_D(1-Sp_F) \\ \mathrm{Se}_{\mathrm{D}}(1-Se_F) & (1-Sp_D)(1-Se_F) & \mathrm{Se}_{\mathrm{D}}Sp_F & (1-Sp_D)Sp_F \\ (1-Se_D)(1-Se_F) & Sp_D(1-Se_F) & (1-Se_D)Sp_F & Sp_DSp_F \end{bmatrix}$$

denotes the correction matrix[3]. GA approach: Find values for $\underline{x}$ such that the sum of squares $S = (\underline{m}-\underline{A}\,\underline{x})'(\underline{m}-\underline{A}\,\underline{x})$ is minimized.

(5) Test validation without gold-standard (2 tests, 2 populations)
Hui and Walter[2] used cross-tabulated data $\underline{m}_1$' = [0, 0, 3, 129] and $\underline{m}_2$' = [3, 0, 24, 3], $\underline{m}_{\mathrm{i}}' = [a_{\mathrm{i}}, b_{\mathrm{i}}, c_{\mathrm{i}}, d_{\mathrm{i}}]$, where a_{i}, b_{i}, c_{i}, d_{i} i=1, 2 represent the observed frequency of (both tests +ve), (test 1 -ve, test 2 +ve), (test 1 +ve, test 2 -ve), (both tests –ve) outcomes, respectively, for the ith population for estimation of the (fixed) sensitivity and specificity of the tests and the prevalences in the two populations. Standard approach: EM algorithm for maximum likelihood (ML) estimation of $\underline{P} = [Se_1, Sp_1, Se_2, Sp_2, p_1, p_2]$ with respect to the likelihood function L. GA approach: Find $\hat{\underline{P}}$ such that L is maximized.

$$L = \sum_{i=1}^{2} \begin{pmatrix} a_i \log(p_iSe_1Se_2 + (1-p_i)(1-Sp_1)(1-Sp_2) + \\ b_i \log(p_i(1-Se_1)Se_2 + (1-p_i)Sp_1(1-Sp_2) + \\ c_i \log(p_iSe_1(1-Se_2) + (1-p_i)(1-Sp_1)Sp_2 + \\ d_i \log(p_i(1-Se_1)(1-Se_2) + (1-p_i)Sp_1Sp_2 \end{pmatrix}$$

(6) Extension of example (5):
Here we consider an extension of (5), where one constraint (Se_1 being equal in both populations) was taken away. We generated two 2x2 tables based on arbitrary values for $\underline{P} = [Se_{11}, Se_{12}, Sp_1, Se_2, Sp_2, p_1, p_2]$ (parameter values are given in Fig. 1) and used a GA to find $\hat{\underline{P}}$ such that L (updated to include the 7th parameter) was maximized. The last step was repeated 500 times and the distribution of parameter estimates was visualized using a kernel density estimate (KDE) (Fig. 1).

Results & discussion

In the first part of the study (examples 1–5) we found that the GAs converged to the correct solutions consistently (i.e. different starting values lead to reproducible results) but much slower than standard procedures. Applied to a test validation with excess number of free parameters (example 6) it was found that those parameters, which were estimated with extremely small variability, were virtually unbiased (Fig. 1). Using this information, the number of variable parameters can be reduced and standard procedures are applicable for estimation of the remaining parameters.

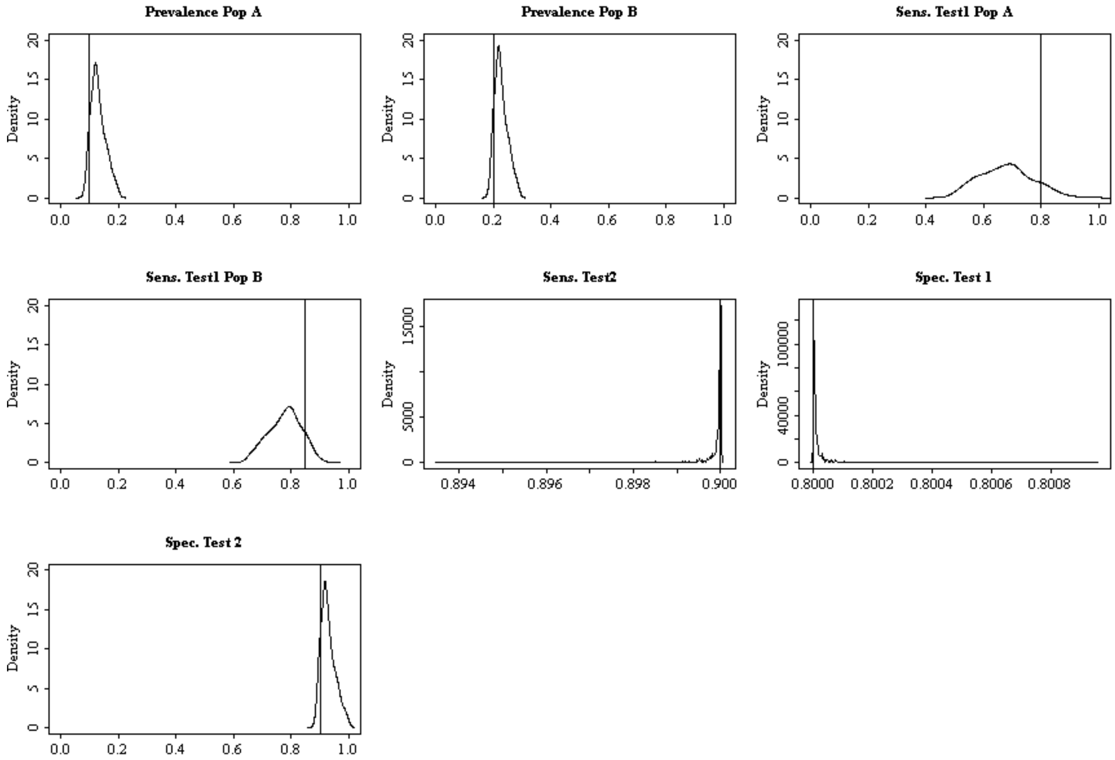

Figure 1: Example (6): Using a GA seven parameters (including different sensitivities in population A and B) were estimated. The results of 500 replications of the GA are displayed as density distribution (KDE). The true parameter values are indicated with a vertical line.

The conclusion of this preliminary study is that GAs potentially yield unbiased parameter estimates for standard problems and, more importantly, can be used in situations where linear parameter search strategies fail.

Reference

1 Holland JH. Adaptation in natural and artificial systems. Ann Harbor, Michigan: University of Michigan Press, 1975.

2 Hui SL, Walter SD. Estimation of error rates of diagnostic tests. Biometrics 1980; 36: 167-171.

3 Greiner M, Gardner I. Application of diagnostic tests in veterinary epidemiologic studies. Preventive Veterinary Medicine 2000; in press.