**Statistics and Computing**

*Series Editors:*
J. Chambers
D. Hand
W. Härdle

For other titles published in this series, go to
http://www.springer.com/series/3022

Roberto Baragona · Francesco Battaglia · Irene Poli

# Evolutionary Statistical Procedures

## An Evolutionary Computation Approach to Statistical Procedures Designs and Applications

Springer

Prof. Roberto Baragona
Sapienza University of Rome
Department of Communication
and Social Research
Via Salaria 113
00198 Rome
Italy
roberto.baragona@uniroma1.it

Prof. Francesco Battaglia
Sapienza University of Rome
Department of Statistical Sciences
Piazzale Aldo Moro 5
00100 Roma
Italy
francesco.battaglia@uniroma1.it

Prof. Irene Poli
Ca' Foscari University of Venice
Department of Statistics
Cannaregio 873
30121 Venice
Italy
irenpoli@unive.it



ISSN 1431-8784
ISBN 978-3-642-16217-6 e-ISBN 978-3-642-16218-3
DOI 10.1007/978-3-642-16218-3
Springer Heidelberg Dordrecht London New York



*Cover design*: SPI Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

In many application fields, artificial intelligence, data mining, pattern recognition operations research, to name but a few, often problems arise that may be reduced at their very essence to optimization problems.

Unfortunately, neither the objective function nor the solution search space display that nice properties that may be conveniently exploited by widespread familiar numerical analysis tools. Though these latter offer powerful devices to cope with a great deal of both theoretical and practical problems in so many disciplines, the hypotheses on which they rely are far from being fulfilled within the frameworks that so often constitute the background of the application fields we mentioned so far. Here well behaved analytic functions and compact domains are not commonplace and only raw hazardous simplifying assumptions may constrain so many classes of problems in such a way they may be treated by means of comfortable numerical procedures.

It may happen that too much simplification does not allow the actual problem to be solved satisfactorily, that is we obtain useless though well grounded solutions. Heuristic methods have been developed to yield reliable solutions to many particular problems, but only the development of general heuristics offered a theoretical framework for dealing with a large class of problems.

One outstanding approach in this kind of methods proved to be evolutionary computing. This rapidly growing research field came nowadays to a well established discipline on its own enjoying solid theoretical foundations and large evidence of effectiveness as far as complex non conventional optimization problems are concerned. History dates back to the fifties and since then an enormous body of theory has been developed which makes evolutionary computing a suitable framework for building applied methodology while it is an active and thriving research field. Its influence spread out through so many disciplines, from biology to informatics and engineering and to economics and social sciences.

We shall try to make an account of the influence of evolutionary computing in Statistics and close related fields. Evolutionary computation is particularly useful in Statistics, in all cases when the statistician has to select, inside a very large discrete set just one element, be it a method, a model, a parameter value, or such.

Therefore a common application of evolutionary computation is to the selection of variables, both in regression problems and in time series linear models. In time series analysis it has been proposed also for building non linear models. For the same reason, evolutionary computation may be employed in the problem of outlier detection, and several papers were published both for the independent observations case and for time series.

A recent, very promising application is in the design of experiments, where the optimal choice of a combination of factors, and their levels, is needed, and cost constraints are very strong.

Finally, a typical field of application of evolutionary computation to Statistics is cluster analysis. Here, the use of an evolutionary algorithm provides a valid alternative when the number of units and variables is large, and the standard cluster techniques allow only approximate solutions.

This book brings together most literature on the use of evolutionary computation methods in statistical analysis, and contains also some unpublished material. It is based on the over 15 years experience and research work of the authors in this field.

This book requires a basic knowledge of mathematics and statistics, and may be useful to research students and professionals to appreciate the suitability for solving complex statistical problems of evolutionary computation. We believe that these algorithms will become a common standard in Statistics in a few years.

Much of the research work included in this book has been possible due to generous funding from Institutions that we are happy to acknowledge with thanks. F. Battaglia and R. Baragona gratefully acknowledge funding from MIUR, Italy (PRIN 2007) and European Commission through Marie Curie Research and Training Network COMISEF – Computational Methods in Statistics, Econometrics and Finance (contract MRTN-CT-2006-034270). I. Poli also would like to acknowledge the MIUR for the PRIN 2007 project, the European Commission for the PACE integrated project (IST-FET, www.istpace.org), and the Fondazione di Venezia for the DICE project. She also would like to thank the researchers at the European Centre for Living Technology for the very fruitful collaboration at the development of the evolutionary perspective in Statistics, in particular Davide de March, Debora Slanzi, Laura Villanova, Matteo Borrotti, Michele Forlin.

We are willing to express our gratitude to many colleagues who commented on large parts of the manuscript letting some obscurities and subtleties be disclosed. In particular we need to mention Antonietta di Salvatore and Domenico Cucina who patiently and carefully read all of the versions of the manuscript, and Debora Slanzi who provided us with lot of computer programs and graphical displays, specially as regards design of experiments. We are indebted for useful and animated discussions to Matt Protopapas, specially concerned with issues related to non linear time series, and to Sanghamitra Bandyopadhyay and Ujjwal Maulik who pointed out in particular some important aspects related to multiobjective optimization and cluster analysis. We have to thank the Springer Editors who successively took care of the manuscript for their assiduous assistance and encouragement, Lilith Braun and Peter

Niels Thomas. Also our thanks are deservedly due to the Desk Editor Alice Blanck for her accurate preparation of the final version of the manuscript, and to Samuel Roobesh, Project Manager at Integra Software Services Pvt. Ltd., for his attentive care in handling the production of this book.

Rome, Italy — Roberto Baragona
Rome, Italy — Francesco Battaglia
Venice, Italy — Irene Poli
September 2010

# Contents