

CRC FOCUS



SWARM INTELLIGENCE METHODS FOR STATISTICAL REGRESSION

Soumya D. Mohanty



CRC Press
Taylor & Francis Group

Swarm Intelligence Methods for Statistical Regression



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Swarm Intelligence Methods for Statistical Regression

Soumya D. Mohanty



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2019 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper
Version Date: 20181128

International Standard Book Number-13: 978-1-138-55818 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged, please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Dedication

To my parents, Rama Ranjan and Krishna



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

Preface	xv
---------	----

Conventions and Notation	xv
--------------------------	----

CHAPTER 1 ■ Introduction	1
--------------------------	---

1.1	OPTIMIZATION IN STATISTICAL ANALYSIS	1
1.2	STATISTICAL ANALYSIS: BRIEF OVERVIEW	3
1.3	STATISTICAL REGRESSION	6
1.3.1	Parametric regression	6
1.3.2	Non-parametric regression	8
1.4	HYPOTHESES TESTING	11
1.5	NOTES	15
1.5.1	Noise in the independent variable	15
1.5.2	Statistical analysis and machine learning	16

CHAPTER 2 ■ Stochastic Optimization Theory	19
--	----

2.1	TERMINOLOGY	20
2.2	CONVEX AND NON-CONVEX OPTIMIZATION PROBLEMS	21

2.3	STOCHASTIC OPTIMIZATION	24
2.4	EXPLORATION AND EXPLOITATION	28
2.5	BENCHMARKING	30
2.6	TUNING	32
2.7	BMR STRATEGY	34
2.8	PSEUDO-RANDOM NUMBERS AND STOCHASTIC OPTIMIZATION	35
2.9	NOTES	36
CHAPTER	3 ■ Evolutionary Computation and Swarm Intelligence	37
3.1	OVERVIEW	37
3.2	EVOLUTIONARY COMPUTATION	39
3.3	SWARM INTELLIGENCE	41
3.4	NOTES	42
CHAPTER	4 ■ Particle Swarm Optimization	45
4.1	KINEMATICS: GLOBAL-BEST PSO	46
4.2	DYNAMICS: GLOBAL-BEST PSO	48
4.2.1	Initialization and termination	49
4.2.2	Interpreting the velocity update rule	49
4.2.3	Importance of limiting particle velocity	51
4.2.4	Importance of proper randomization	53
4.2.5	Role of inertia	54
4.2.6	Boundary condition	56
4.3	KINEMATICS: LOCAL-BEST PSO	56
4.4	DYNAMICS: LOCAL-BEST PSO	58
4.5	STANDARDIZED COORDINATES	59

4.6	RECOMMENDED SETTINGS FOR REGRESSION PROBLEMS	60
4.7	NOTES	60
4.7.1	Additional PSO variants	61
4.7.2	Performance example	63
CHAPTER	5 ■ PSO Applications	65
5.1	GENERAL REMARKS	66
5.1.1	Fitness function	66
5.1.2	Data simulation	67
5.1.3	Parametric degeneracy and noise	68
5.1.4	PSO variant and parameter settings	70
5.2	PARAMETRIC REGRESSION	70
5.2.1	Tuning	70
5.2.2	Results	74
5.3	NON-PARAMETRIC REGRESSION	77
5.3.1	Reparametrization in regression spline	78
5.3.2	Results: Fixed number of breakpoints	81
5.3.3	Results: Variable number of breakpoints	82
5.4	NOTES AND SUMMARY	84
5.4.1	Summary	87
APPENDIX	A ■ Probability Theory	89
A.1	RANDOM VARIABLE	89
A.2	PROBABILITY MEASURE	90
A.3	JOINT PROBABILITY	92
A.4	CONTINUOUS RANDOM VARIABLES	94
A.5	EXPECTATION	97

A.6	COMMON PROBABILITY DENSITY FUNCTIONS	98
APPENDIX	B ■ Splines	101
B.1	DEFINITION	101
B.2	B-SPLINE BASIS	103
APPENDIX	C ■ Analytical Minimization	107
C.1	QUADRATIC CHIRP	108
C.2	SPLINE-BASED SMOOTHING	109
Bibliography		111
Index		117

Preface

This book is based on a set of lectures on big data analysis delivered at the BigDat International Winter School held at Bari, Italy, in 2017. The lectures focused on a very practical issue encountered in the statistical regression of non-linear models, namely, the numerical optimization of the fitting function. The optimization problem in statistical analysis, especially in big data applications, is often a bottleneck that forces either the adoption of simpler models or a shift to linear models even where non-linearity is known to be a better option. The goal of the lectures was to introduce the audience to a set of relatively recent biology inspired stochastic optimization methods, collectively called swarm intelligence (SI) methods, that are proving quite effective in tackling the optimization challenge in statistical analysis.

It was clear from the audience response at these lectures that, despite their collective background in very diverse areas of data analysis ranging from the natural sciences to the social media industry, many had not heard of and none had seriously explored SI methods. The root causes behind this lacuna seem to be (a) a lack of familiarity, within the data analysis community, of the latest literature in stochastic optimization, and (b) lack of experience and guidance in tuning these methods to work well in real-world problems. Matters are not helped by the fact that there are not a whole lot of papers in the optimization community examining the role of SI methods in statistical analysis, most of the focus in that field being on optimization problems in engineering.

I hope this small book helps in bridging the current divide between the two communities. As I have seen within my own

research, the statistical data analyst will find that success in solving the optimization challenge spurs the contemplation of better, more sophisticated models for data. Students and researchers in the SI community reading this book will find that statistical data analysis offers a rich source of challenging test beds for their methods.

The aim of the book is to arm the reader with practical tips and rules of thumb that are observed to work well, not to provide a deeper than necessary theoretical background. In particular, the book does not delve deep into the huge range of SI methods out there, concentrating rather on one particular method, namely particle swarm optimization (PSO). My experience in teaching SI methods to students has shown that it is best to learn these methods by starting with one and understanding it well. For this purpose, PSO provides the simplest entry point. Similarly, this book does not provide a more than superficial background in optimization theory, choosing to only highlight some its important results.

It is assumed that the reader of this book has a basic background in probability theory and function approximation at the level of undergraduate or graduate courses. Nonetheless, appendices are provided that cover the required material succinctly. Instead of a problem set, two realistic statistical regression problems covering both parametric and non-parametric approaches form the workhorse exercises in this book. The reader is highly encouraged to independently implement these examples and reproduce the associated results provided in the book.

References are primarily provided in the “Notes” section at the end of each chapter. While it was tempting to include in the book a more complete review of the technical literature than what is currently provided, I decided to point the reader to mostly textbooks or review articles. This was done keeping in mind the expected readership of this book, which I assume would be similar to the student-heavy makeup of the BigDat participants. This inevitably means that many key references have been left out but I hope that the ones included will give

readers a good start in seeking out the technical material that is appropriate for their application areas.

Acknowledgements: It is a pleasure to acknowledge colleagues who supported the inception and development of this book. I thank Carlos Martin-Vide, Donato Malerba, and other organizers of the BigDat schools for inviting me as a lecturer and for their hospitality. This annual school provides a wonderful and refreshing opportunity to interact with data scientists and students from a broad spectrum of fields, and I hope that it continues to have several future editions.

Before embarking on the writing of this book, I had the pleasure of discussing the core content of the lectures with Innocenzo Pinto and Luigi Troiano at the University of Sannio at Benevento, Italy. I had similar opportunities, thanks to Runqiu Liu and Zong-kuan Guo at the Chinese Academy of Sciences, Beijing, to present the material to undergraduate and graduate students during my lectures there on gravitational wave data analysis. I am greatly indebted to Yan Wang at Huazhong University of Science and Technology, Wuhan, for a thorough reading of the draft and many invaluable comments and suggestions. Several useful comments from Ram Valluri at the University of Western Ontario are appreciated as well. Finally, I thank Randi Cohen at CRC press for contacting me and initiating this project.