Table of Contents

Preface vii

Chapter 1: Introduction 1

What is simulation and where is it applied? 3
Why use simulation? 6
Simulation and big data 7
Choosing the right simulation technique 8
Summary 11
References 11

Chapter 2: R and High-Performance Computing 13

The R statistical environment 14
Basics in R 15
Some very basic stuff about R 15
Installation and updates 16
Help 17
The R workspace and the working directory 18
Data types 18
Vectors in R 19
Factors in R 21
list 21
data.frame 22
array 24
Missing values 25
Generic functions, methods, and classes 26
Data manipulation in R 28
Apply and friends with basic R 28
Basic data manipulation with the dplyr package 31
dplyr – creating a local data frame 32
dplyr – selecting lines 33
dplyr – order 34

dplyr – selecting columns 35
dplyr – uniqueness 37
dplyr – creating variables 37
dplyr – grouping and aggregates 38
dplyr – window functions 41
Data manipulation with the data.table package 42
data.table – variable construction 42
data.table – indexing or subsetting 43
data.table – keys 44
data.table – fast subsetting 44
data.table – calculations in groups 46
High performance computing 47
Profiling to detect computationally slow functions in code 47
Further benchmarking 49
Parallel computing 56
Interfaces to C++ 58
Visualizing information 60
The graphics system in R 61
The graphics package 62
Warm-up example – a high-level plot 62
Control of graphics parameters 64
The ggplot2 package 66
References 71
Chapter 3: The Discrepancy between Pencil-Driven Theory and Data-Driven Computational Solutions 73
Machine numbers and rounding problems 74
Example – the 64-bit representation of numbers 77
Convergence in the deterministic case 77
Example – convergence 78
Condition of problems 86
Summary 87
References 87
Chapter 4: Simulation of Random Numbers 89
Real random numbers 90
Simulating pseudo random numbers 92
Congruential generators 93
Linear and multiplicative congruential generators 94
Lagged Fibonacci generators 98
More generators 98
Simulation of non-uniform distributed random variables 101
The inversion method 101

The alias method 105
Estimation of counts in tables with log-linear models 106
Rejection sampling 108
Truncated distributions 116
Metropolis - Hastings algorithm 117
A few words on Markov chains 118
The Metropolis sampler 126
The Gibbs sampler 129
The two-phase Gibbs sampler 129
The multiphase Gibbs sampler 131
Application in linear regression 132
The diagnosis of MCMC samples 134
Tests for random numbers 141
The evaluation of random numbers – an example of a test 142
Summary 146
References 146
Chapter 5: Monte Carlo Methods for Optimization Problems 149
Numerical optimization 153
Gradient ascent/descent 154
Newton-Raphson methods 154
Further general-purpose optimization methods 157
Dealing with stochastic optimization 159
Simplified procedures (Star Trek, Spaceballs, and Spaceballs princess) 159
Metropolis-Hastings revisited 163
Gradient-based stochastic optimization 165
Summary 170
References 171
Chapter 6: Probability Theory Shown by Simulation 173
Some basics on probability theory 173
Probability distributions 174
Discrete probability distributions 174
Continuous probability distributions 175
Winning the lottery 176
The weak law on large numbers 178
Emperor penguins and your boss 178
Limits and convergence of random variables 180
Convergence of the sample mean – weak law of large numbers 181
Showing the weak law of large numbers by simulation 182
The central limit theorem 190
Properties of estimators 195
Properties of estimators 196

Confidence intervals 197
A note on robust estimators 200
Summary 201
References 201
Chapter 7: Resampling Methods 203
The bootstrap 204
A motivating example with odds ratios 205
Why the bootstrap works 208
A closer look at the bootstrap 211
The plug-in principle 212
Estimation of standard errors with bootstrapping 213
An example of a complex estimation using the bootstrap 216
The parametric bootstrap 218
Estimating bias with bootstrap 221
Confidence intervals by bootstrap 222
The jackknife 226
Disadvantages of the jackknife 229
The delete-d jackknife 230
Jackknife after bootstrap 232
Cross-validation 235
The classical linear regression model 235
The basic concept of cross validation 236
Classical cross validation – 70/30 method 238
Leave-one-out cross validation 240
k-fold cross validation 242
Summary 244
References 245
Chapter 8: Applications of Resampling Methods and Monte Carlo Tests 247
The bootstrap in regression analysis 247
Motivation to use the bootstrap 248
The most popular but often worst method 253
Bootstrapping by draws from residuals 258
Proper variance estimation with missing values 263
Bootstrapping in time series 269
Bootstrapping in the case of complex sampling designs 273
Monte Carlo tests 278
A motivating example 278
The permutation test as a special kind of MC test 287

A Monte Carlo test for multiple groups 290
Hypothesis testing using a bootstrap 294
A test for multivariate normality 295
Size of the test 297
Power comparisons 298
Summary 298
References 299
Chapter 9: The EM Algorithm 301
The basic EM algorithm 301
Some prerequisites 302
Formal definition of the EM algorithm 303
Introductory example for the EM algorithm 304
The EM algorithm by example of k-means clustering 305
The EM algorithm for the imputation of missing values 312
Summary 318
References 318
Chapter 10: Simulation with Complex Data 321
Different kinds of simulation and software 322
Simulating data using complex models 324
A model-based simple example 324
A model-based example with mixtures 327
Model-based approach to simulate data 328
An example of simulating high-dimensional data 329
Simulating finite populations with cluster or hierarchical structures 330
Model-based simulation studies 333
Latent model example continued 334
A simple example of model-based simulation 336
A model-based simulation study 341
Design-based simulation 347
An example with complex survey data 348
Simulation of the synthetic population 349
Estimators of interest 350
Defining the sampling design 351
Using stratified sampling 353
Adding contamination 354
Performing simulations separately on different domains 356
Inserting missing values 357
Summary 359
References 359

Chapter 11: System Dynamics and Agent-Based Models 363
Agent-based models 364
Dynamics in love and hate 368
Dynamic systems in ecological modeling 371
Summary 374
References 374
Index 375