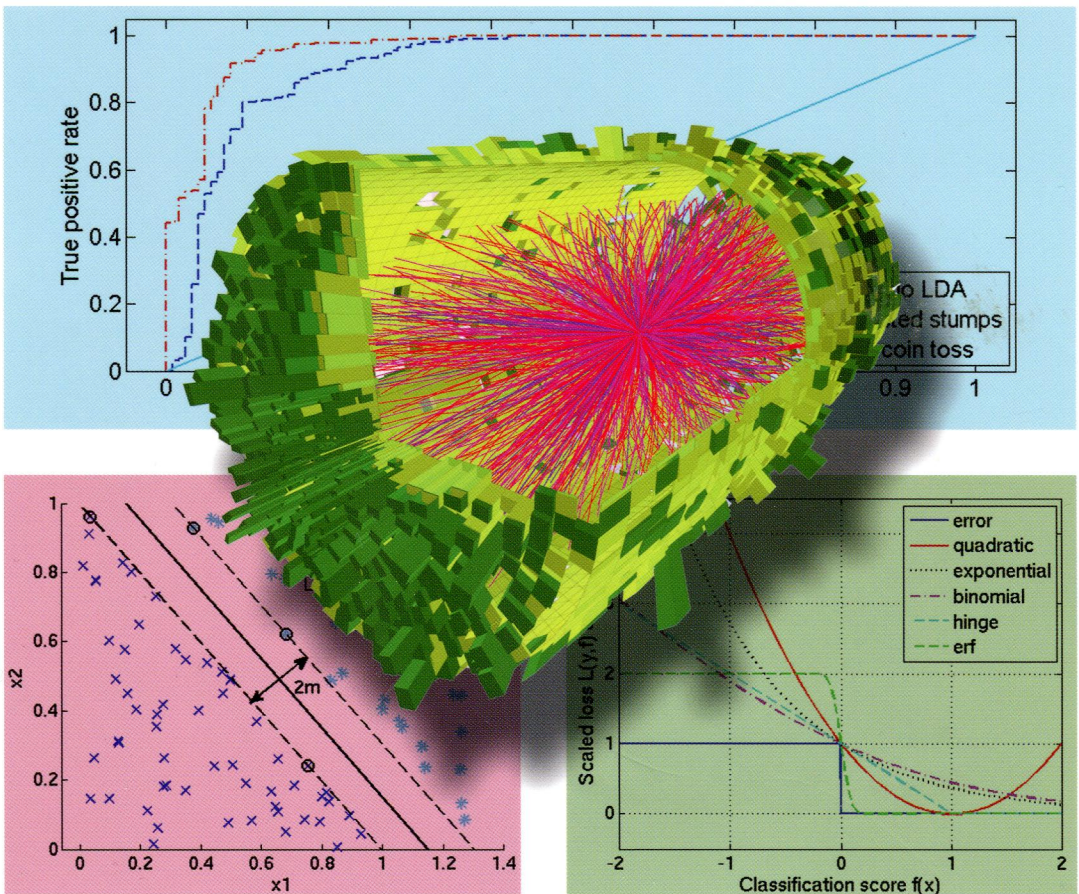


# Statistical Analysis Techniques in Particle Physics

Fits, Density Estimation and Supervised Learning



*Ilya Narsky and Frank C. Porter*

# **Statistical Analysis Techniques in Particle Physics**

Fits, Density Estimation and Supervised Learning



## Authors

### **Dr. Ilya Narsky**

MathWorks  
Natick  
United States of America  
inarsky@yahoo.com

### **Prof. Frank C. Porter**

California Inst. of Technology  
Physics Department  
Pasadena  
United States of America  
fcp@hep.caltech.edu

## Cover

CMS Experiment at LHC, CERN  
Data recorded Mon Nov 8, 11:30:43 2010 CEST  
Run/Event 150431/630470  
Lumi section 173

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

**Library of Congress Card No.:**  
applied for

**British Library Cataloguing-in-Publication Data:**  
A catalogue record for this book is available from the British Library.

## **Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

© 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Boschstr. 12, 69469 Weinheim, Germany

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

**Print ISBN** 978-3-527-41086-6

**ePDF ISBN** 978-3-527-67731-3

**ePub ISBN** 978-3-527-67729-0

**Mobi ISBN** 978-3-527-67730-6

**oBook ISBN** 978-3-527-67732-0

**Cover Design** Grafik-Design Schulz

**Typesetting** le-tex publishing services GmbH, Leipzig

**Printing and Binding** Markono Print Media Pte Ltd, Singapore

Printed on acid-free paper

## Contents

	<b>Acknowledgements</b>	<i>XIII</i>
	<b>Notation and Vocabulary</b>	<i>XV</i>
<b>1</b>	<b>Why We Wrote This Book and How You Should Read It</b>	<i>1</i>
<b>2</b>	<b>Parametric Likelihood Fits</b>	<i>5</i>
2.1	Preliminaries	<i>5</i>
2.1.1	Example: CP Violation via Mixing	<i>7</i>
2.1.2	The Exponential Family	<i>9</i>
2.1.3	Confidence Intervals	<i>10</i>
2.1.4	Hypothesis Tests	<i>11</i>
2.2	Parametric Likelihood Fits	<i>12</i>
2.2.1	Nuisance Parameters	<i>16</i>
2.2.2	Confidence Intervals from Pivotal Quantities	<i>17</i>
2.2.3	Asymptotic Inference	<i>19</i>
2.2.4	Profile Likelihood	<i>20</i>
2.2.5	Conditional Likelihood	<i>20</i>
2.3	Fits for Small Statistics	<i>21</i>
2.3.1	Sample Study of Coverage at Small Statistics	<i>22</i>
2.3.2	When the pdf Goes Negative	<i>25</i>
2.4	Results Near the Boundary of a Physical Region	<i>26</i>
2.5	Likelihood Ratio Test for Presence of Signal	<i>28</i>
2.6	Ⓢ sPlots	<i>31</i>
2.7	Exercises	<i>35</i>
	References	<i>37</i>
<b>3</b>	<b>Goodness of Fit</b>	<i>39</i>
3.1	Binned Goodness of Fit Tests	<i>41</i>
3.2	Statistics Converging to Chi-Square	<i>46</i>
3.3	Univariate Unbinned Goodness of Fit Tests	<i>49</i>
3.3.1	Kolmogorov–Smirnov	<i>49</i>
3.3.2	Anderson–Darling	<i>50</i>
3.3.3	Watson	<i>51</i>
3.3.4	Neyman Smooth	<i>51</i>

3.4	Multivariate Tests	52
3.4.1	Energy Tests	53
3.4.2	Transformations to a Uniform Distribution	54
3.4.3	Local Density Tests	55
3.4.4	Kernel-based Tests	56
3.4.5	Mixed Sample Tests	57
3.4.6	Using a Classifier	58
3.5	Exercises	59
	References	61
<b>4</b>	<b>Resampling Techniques</b>	<b>63</b>
4.1	Permutation Sampling	63
4.2	Bootstrap	65
4.2.1	Bootstrap Confidence Intervals	68
4.2.2	Smoothed Bootstrap	70
4.2.3	Parametric Bootstrap	70
4.3	Jackknife	70
4.4	BC <sub>a</sub> Confidence Intervals	76
4.5	Cross-Validation	78
4.6	⊗ Resampling Weighted Observations	82
4.7	Exercises	86
	References	86
<b>5</b>	<b>Density Estimation</b>	<b>89</b>
5.1	Empirical Density Estimate	90
5.2	Histograms	90
5.3	Kernel Estimation	92
5.3.1	Multivariate Kernel Estimation	92
5.4	Ideogram	93
5.5	Parametric vs. Nonparametric Density Estimation	93
5.6	Optimization	94
5.6.1	Choosing Histogram Binning	97
5.7	Estimating Errors	100
5.8	The Curse of Dimensionality	102
5.9	Adaptive Kernel Estimation	103
5.10	Naive Bayes Classification	105
5.11	Multivariate Kernel Estimation	106
5.12	Estimation Using Orthogonal Series	108
5.13	Using Monte Carlo Models	111
5.14	Unfolding	112
5.14.1	Unfolding: Regularization	116
5.15	Exercises	120
	References	120
<b>6</b>	<b>Basic Concepts and Definitions of Machine Learning</b>	<b>121</b>
6.1	Supervised, Unsupervised, and Semi-Supervised	121

6.2	Tall and Wide Data	123
6.3	Batch and Online Learning	124
6.4	Parallel Learning	125
6.5	Classification and Regression	127
	References	128
<b>7</b>	<b>Data Preprocessing</b>	<b>129</b>
7.1	Categorical Variables	129
7.2	Missing Values	132
7.2.1	Likelihood Optimization	134
7.2.2	Deletion	135
7.2.3	Augmentation	137
7.2.4	Imputation	137
7.2.5	Other Methods	139
7.3	Outliers	139
7.4	Exercises	141
	References	142
<b>8</b>	<b>Linear Transformations and Dimensionality Reduction</b>	<b>145</b>
8.1	Centering, Scaling, Reflection and Rotation	145
8.2	Rotation and Dimensionality Reduction	146
8.3	Principal Component Analysis (PCA)	147
8.3.1	Theory	148
8.3.2	Numerical Implementation	149
8.3.3	Weighted Data	150
8.3.4	How Many Principal Components Are Enough?	151
8.3.5	Example: Apply PCA and Choose the Optimal Number of Components	154
8.4	Independent Component Analysis (ICA)	158
8.4.1	Theory	158
8.4.2	Numerical implementation	161
8.4.3	Properties	162
8.5	Exercises	163
	References	163
<b>9</b>	<b>Introduction to Classification</b>	<b>165</b>
9.1	Loss Functions: Hard Labels and Soft Scores	165
9.2	Bias, Variance, and Noise	168
9.3	Training, Validating and Testing: The Optimal Splitting Rule	173
9.4	Resampling Techniques: Cross-Validation and Bootstrap	177
9.4.1	Cross-Validation	177
9.4.2	Bootstrap	179
9.4.3	Sampling with Stratification	181
9.5	Data with Unbalanced Classes	182
9.5.1	Adjusting Prior Probabilities	183
9.5.2	⊗ Undersampling the Majority Class	184

9.5.3	☞ Oversampling the Minority Class	185
9.5.4	Example: Classification of Forest Cover Type Data	186
9.6	Learning with Cost	190
9.7	Exercises	191
	References	192
<b>10</b>	<b>Assessing Classifier Performance</b>	<b>195</b>
10.1	Classification Error and Other Measures of Predictive Power	195
10.2	Receiver Operating Characteristic (ROC) and Other Curves	196
10.2.1	Empirical ROC curve	196
10.2.2	Other Performance Measures	198
10.2.3	Optimal Operating Point	198
10.2.4	Area Under Curve	200
10.2.5	Smooth ROC Curves	200
10.2.6	Confidence Bounds for ROC Curves	205
10.3	Testing Equivalence of Two Classification Models	210
10.4	Comparing Several Classifiers	215
10.5	Exercises	217
	References	218
<b>11</b>	<b>Linear and Quadratic Discriminant Analysis, Logistic Regression, and Partial Least Squares Regression</b>	<b>221</b>
11.1	Discriminant Analysis	221
11.1.1	Estimating the Covariance Matrix	223
11.1.2	Verifying Discriminant Analysis Assumptions	225
11.1.3	Applying LDA When LDA Assumptions Are Invalid	226
11.1.4	Numerical Implementation	228
11.1.5	Regularized Discriminant Analysis	228
11.1.6	LDA for Variable Transformation	229
11.2	Logistic Regression	231
11.2.1	Binomial Logistic Regression: Theory and Numerical Implementation	231
11.2.2	Properties of the Binomial Model	233
11.2.3	Verifying Model Assumptions	233
11.2.4	Logistic Regression with Multiple Classes	234
11.3	Classification by Linear Regression	235
11.4	☞ Partial Least Squares Regression	236
11.5	Example: Linear Models for MAGIC Telescope Data	239
11.6	Choosing a Linear Classifier for Your Analysis	247
11.7	Exercises	247
	References	248
<b>12</b>	<b>Neural Networks</b>	<b>251</b>
12.1	Perceptrons	251
12.2	The Feed-Forward Neural Network	254
12.3	Backpropagation	256

- 12.4 Bayes Neural Networks 260
- 12.5 Genetic Algorithms 262
- 12.6 Exercises 263
- References 263
- 13 Local Learning and Kernel Expansion 265**
  - 13.1 From Input Variables to the Feature Space 266
    - 13.1.1 Kernel Regression 269
  - 13.2 Regularization 270
    - 13.2.1 Kernel Ridge Regression 274
  - 13.3 Making and Choosing Kernels 278
  - 13.4 Radial Basis Functions 279
    - 13.4.1 Example: RBF Classification for the MAGIC Telescope Data 280
  - 13.5 Support Vector Machines (SVM) 283
    - 13.5.1 SVM with Weighted Data 286
    - 13.5.2 SVM with Probabilistic Outputs 288
    - 13.5.3 ☞ Numerical Implementation 288
    - 13.5.4 ☞ Multiclass Extensions 293
  - 13.6 Empirical Local Methods 293
    - 13.6.1 Classification by Probability Density Estimation 294
    - 13.6.2 Locally Weighted Regression 295
    - 13.6.3 Nearest Neighbors and Fuzzy Rules 298
  - 13.7 Kernel Methods: The Good, the Bad and the Curse of Dimensionality 302
  - 13.8 Exercises 303
  - References 304
- 14 Decision Trees 307**
  - 14.1 Growing Trees 308
  - 14.2 Predicting by Decision Trees 312
  - 14.3 Stopping Rules 312
  - 14.4 Pruning Trees 313
    - 14.4.1 Example: Pruning a Classification Tree 317
  - 14.5 Trees for Multiple Classes 319
  - 14.6 ☞ Splits on Categorical Variables 320
  - 14.7 Surrogate Splits 321
  - 14.8 ☞ Missing Values 323
  - 14.9 Variable importance 324
  - 14.10 Why Are Decision Trees Good (or Bad)? 327
  - 14.11 Exercises 328
  - References 329
- 15 Ensemble Learning 331**
  - 15.1 Boosting 332
    - 15.1.1 Early Boosting 332
    - 15.1.2 AdaBoost for Two Classes 333



- 15.1.3 Minimizing Convex Loss by Stagewise Additive Modeling 336
- 15.1.4 Maximizing the Minimal Margin 343
- 15.1.5 Nonconvex Loss and Robust Boosting 351
- 15.1.6 Boosting for Multiple Classes 357
- 15.2 Diversifying the Weak Learner: Bagging, Random Subspace and Random Forest 358
  - 15.2.1 Measures of Diversity 359
  - 15.2.2 Bagging and Random Forest 361
  - 15.2.3 Random Subspace 363
  - 15.2.4 Example:  $K/\pi$  Separation for BaBar PID 364
- 15.3 Choosing an Ensemble for Your Analysis 365
- 15.4 Exercises 367
  - References 367
- 16 Reducing Multiclass to Binary 371**
  - 16.1 Encoding 372
  - 16.2 Decoding 375
  - 16.3 Summary: Choosing the Right Design 378
    - References 379
- 17 How to Choose the Right Classifier for Your Analysis and Apply It Correctly 381**
  - 17.1 Predictive Performance and Interpretability 381
  - 17.2 Matching Classifiers and Variables 382
  - 17.3 Using Classifier Predictions 382
  - 17.4 Optimizing Accuracy 383
  - 17.5 CPU and Memory Requirements 383
- 18 Methods for Variable Ranking and Selection 385**
  - 18.1 Definitions 386
    - 18.1.1 Variable Ranking and Selection 386
    - 18.1.2 Strong and Weak Relevance 386
  - 18.2 Variable Ranking 389
    - 18.2.1 Filters: Correlation and Mutual Information 390
    - 18.2.2 Wrappers: Sequential Forward Selection (SFS), Sequential Backward Elimination (SBE), and Feature-based Sensitivity of Posterior Probabilities (FSPP) 394
    - 18.2.3 Embedded Methods: Estimation of Variable Importance by Decision Trees, Neural Networks, Nearest Neighbors, and Linear Models 400
  - 18.3 Variable Selection 401
    - 18.3.1 Optimal-Set Search Strategies 401
    - 18.3.2 Multiple Testing: Backward Elimination by Change in Margin (BECM) 403
    - 18.3.3 Estimation of the Reference Distribution by Permutations: Artificial Contrasts with Ensembles (ACE) Algorithm 410
  - 18.4 Exercises 413

	References	414
<b>19</b>	<b>Bump Hunting in Multivariate Data</b>	417
19.1	Voronoi Tessellation and SLEUTH Algorithm	418
19.2	Identifying Box Regions by PRIM and Other Algorithms	420
19.3	Bump Hunting Through Supervised Learning	422
	References	423
<b>20</b>	<b>Software Packages for Machine Learning</b>	425
20.1	Tools Developed in HEP	425
20.2	R	426
20.3	MATLAB	427
20.4	Tools for Java and Python	428
20.5	What Software Tool Is Right for You?	429
	References	430
	<b>Appendix A: Optimization Algorithms</b>	431
A.1	Line Search	431
A.2	Linear Programming (LP)	432
	<b>Index</b>	435