

*Edited by*

*Olaf Behnke, Kevin Kröninger,  
Grégory Schott, and  
Thomas Schörner-Sadenius*

**Data Analysis  
in High Energy Physics**

## **Related Titles**

Brock, I., Schörner-Sadenius, T. (eds.)

### **Physics at the Terascale**

2011

ISBN: 978-3-527-41001-9

Griffiths, D.

### **Introduction to Elementary Particles**

2008

ISBN: 978-3-527-40601-2

Russenschuck, S.

### **Field Computation for Accelerator Magnets**

Analytical and Numerical Methods for Electromagnetic Design and Optimization

2010

ISBN: 978-3-527-40769-9

Reiser, M.

### **Theory and Design of Charged Particle Beams**

2008

ISBN: 978-3-527-40741-5

Halpern, P.

### **Collider**

The Search for the World's Smallest Particles

2009

ISBN: 978-0-470-28620-3

Wangler, T.P.

### **RF Linear Accelerators**

2008

ISBN: 978-3-527-40680-7

Martin, B., Shaw, G.

### **Particle Physics**

2008

ISBN: 978-0-470-03294-7

Padamsee, H., Knobloch, J., Hays, T.

### **RF Superconductivity for Accelerators**

2008

ISBN: 978-3-527-40842-9

Talman, R.

### **Accelerator X-Ray Sources**

2006

ISBN: 978-3-527-40590-9

# **Data Analysis in High Energy Physics**

A Practical Guide to Statistical Methods

*Edited by*

*Olaf Behnke, Kevin Kröninger, Grégory Schott, and  
Thomas Schörner-Sadenius*



WILEY-VCH Verlag GmbH & Co. KGaA

**The Editors****Dr. Olaf Behnke**

DESY  
Hamburg  
Germany  
olaf.behnke@desy.de

**Dr. Kevin Kröninger**

Universität Göttingen  
II. Physikalisches Institut  
Göttingen, Germany  
kevin.kroeninger@phys.uni-goettingen.de

**Dr. Grégory Schott**

Karlsruher Institut für Technologie  
Institut für Experimentelle Kernphysik  
Karlsruhe, Germany  
gregory.schott@cern.ch

**Dr. Thomas Schörner-Sadenius**

DESY  
Hamburg, Germany  
thomas.schoerner@desy.de

**The Cover Picture**

represents a hypothetical invariant-mass distribution. The markers with error bars represent the experimental data, the blue area the estimated background and the green regions possible signals for  $M = 200$ ,  $M = 300$  and  $M = 400$  (in arbitrary units). The inset shows the negative logarithm of the likelihood function used to identify a resonance in the mass spectrum.

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>.

© 2013 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 photostriking, 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** 978-3-527-41058-3

**ePDF ISBN** 978-3-527-65344-7

**ePub ISBN** 978-3-527-65343-0

**mobi ISBN** 978-3-527-65342-3

**oBook ISBN** 978-3-527-65341-6

**Cover Design** Grafik-Design Schulz,  
Fußgönheim

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

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

Printed in Singapore

Printed on acid-free paper

## Contents

**Preface** XV

**List of Contributors** XIX

<b>1</b>	<b>Fundamental Concepts</b>	<b>1</b>
	<i>Roger Barlow</i>	
1.1	Introduction	1
1.2	Probability Density Functions	2
1.2.1	Expectation Values	2
1.2.2	Moments	3
1.2.2.1	Variance	3
1.2.2.2	Skew and Kurtosis	3
1.2.2.3	Covariance and Correlation	4
1.2.2.4	Marginalisation and Projection	4
1.2.2.5	Other Properties	5
1.2.3	Associated Functions	5
1.3	Theoretical Distributions	5
1.3.1	The Gaussian Distribution	6
1.3.2	The Poisson Distribution	8
1.3.3	The Binomial Distribution	9
1.3.4	Other Distributions	10
1.3.4.1	The Uniform Distribution	10
1.3.4.2	The Cauchy, or Breit–Wigner, or Lorentzian Distribution	10
1.3.4.3	The Landau Distribution	11
1.3.4.4	The Negative Binomial Distribution	12
1.3.4.5	Student’s <i>t</i> Distribution	12
1.3.4.6	The $\chi^2$ Distribution	14
1.3.4.7	The Log-Normal Distribution	15
1.3.4.8	The Weibull Distribution	15
1.4	Probability	16
1.4.1	Mathematical Definition of Probability	17
1.4.2	Classical Definition of Probability	17
1.4.3	Frequentist Definition of Probability	17
1.4.4	Bayesian Definition of Probability	18

1.4.4.1	Bayes' Theorem	19
1.5	Inference and Measurement	20
1.5.1	Likelihood	20
1.5.2	Frequentist Inference	20
1.5.3	Bayesian Inference	21
1.5.3.1	Use of Different Priors	22
1.5.3.2	Jeffreys Priors	23
1.5.3.3	The Correct Prior?	24
1.6	Exercises	24
	References	25
<b>2</b>	<b>Parameter Estimation</b>	27
	<i>Olaf Behnke and Lorenzo Moneta</i>	
2.1	Parameter Estimation in High Energy Physics: Introductory Words	27
2.2	Parameter Estimation: Definition and Properties	27
2.3	The Method of Maximum Likelihood	29
2.3.1	Maximum-Likelihood Solution	30
2.3.2	Properties of the Maximum-Likelihood Estimator	31
2.3.3	Maximum Likelihood and Bayesian Statistics	31
2.3.3.1	Averaging of Measurements with Gaussian Errors	32
2.3.4	Variance of the Maximum-Likelihood Estimator	33
2.3.4.1	Confidence Region Evaluation with $\chi^2$ Function Quantiles	37
2.3.4.2	Profile Likelihood	37
2.3.5	Minimum-Variance Bound and Experiment Design	38
2.4	The Method of Least Squares	40
2.4.1	Linear Least-Squares Method	42
2.4.1.1	Averaging of Measurements with Gaussian Errors	43
2.4.1.2	Averaging Correlated Measurements	44
2.4.1.3	Straight-Line Fit	46
2.4.2	Non-linear Least-Squares Fits	48
2.4.2.1	Non-linear Least Squares: Mass-Peak Fit (Signal Position)	50
2.5	Maximum-Likelihood Fits:	
	Unbinned, Binned, Standard and Extended Likelihood	52
2.5.1	Unbinned Maximum-Likelihood Fits	52
2.5.1.1	Unbinned MLE: Fitting Fractions of Processes	54
2.5.2	Extended Maximum Likelihood	55
2.5.2.1	Unbinned Extended MLE: Fitting Rates of Processes	56
2.5.2.2	Unbinned Extended MLE: Fitting a Signal with Position-Dependent Normalisation	57
2.5.3	Binned Maximum-Likelihood Fits	59
2.5.4	Least-Squares Fit to a Histogram	60
2.5.4.1	Binned Mass-Peak Fit: Practical Considerations	61
2.5.5	Special Topic: Averaging Data with Inconsistencies	63
2.6	Bayesian Parameter Estimation	67
2.7	Exercises	69
	References	72

<b>3</b>	<b>Hypothesis Testing</b>	75
	<i>Gr��gory Schott</i>	
3.1	Basic Concepts	75
3.1.1	Statistical Hypotheses	75
3.1.2	Test Statistic	76
3.1.3	Critical Region	77
3.1.4	Type I and Type II Errors	79
3.1.5	Summary: the Testing Process	80
3.2	Choosing the Test Statistic	80
3.3	Choice of the Critical Region	82
3.4	Determining Test Statistic Distributions	82
3.5	<i>p</i> -Values	83
3.5.1	Significance Levels	84
3.5.2	Inclusion of Systematic Uncertainties	86
3.5.3	Combining Tests	87
3.5.4	Look-Elsewhere Effect	88
3.6	Inversion of Hypothesis Tests	89
3.7	Bayesian Approach to Hypothesis Testing	92
3.8	Goodness-of-Fit Tests	92
3.8.1	Pearson's $\chi^2$ Test	93
3.8.2	Run Test	96
3.8.3	$\chi^2$ Test with Unbinned Measurements	98
3.8.4	Test Using the Maximum-Likelihood Estimate	98
3.8.5	Kolmogorov-Smirnov Test	99
3.8.6	Smirnov-Cram��r-von Mises Test	101
3.8.7	Two-Sample Tests	101
3.9	Conclusion	102
3.10	Exercises	102
	References	104
<b>4</b>	<b>Interval Estimation</b>	107
	<i>Luc Demortier</i>	
4.1	Introduction	107
4.2	Characterisation of Interval Constructions	108
4.3	Frequentist Methods	110
4.3.1	Neyman's Construction	110
4.3.1.1	Ingredient 1: the Estimator	111
4.3.1.2	Ingredient 2: the Reference Ensemble	112
4.3.1.3	Ingredient 3: the Ordering Rule	113
4.3.1.4	Ingredient 4: the Confidence Level	115
4.3.2	Test Inversion	116
4.3.3	Pivoting	118
4.3.3.1	Gaussian Means and Standard Deviations	118
4.3.3.2	Exponential Lifetimes	120
4.3.3.3	Binomial Efficiencies	120

4.3.3.4	Poisson Means	122
4.3.4	Asymptotic Approximations	123
4.3.5	Bootstrapping	124
4.3.5.1	The Bootstrap- <i>t</i> Interval	125
4.3.5.2	Percentile Intervals	127
4.3.5.3	Bootstrap Calibration	128
4.3.6	Nuisance Parameters	128
4.4	Bayesian Methods	133
4.4.1	Binomial Efficiencies	137
4.4.2	Poisson Means	138
4.5	Graphical Comparison of Interval Constructions	140
4.6	The Role of Intervals in Search Procedures	142
4.6.1	Coverage	143
4.6.2	Sensitivity	144
4.7	Final Remarks and Recommendations	146
4.8	Exercises	146
	References	150
<b>5</b>	<b>Classification</b>	153
	<i>Helge Voss</i>	
5.1	Introduction to Multivariate Classification	153
5.2	Classification from a Statistical Perspective	155
5.2.1	Receiver-Operating-Characteristic Curve and the Neyman–Pearson Lemma	157
5.2.1.1	Neyman–Pearson Lemma	158
5.2.2	Supervised Machine Learning	159
5.2.3	Bias–Variance Trade-Off	159
5.2.4	Cross-Validation	161
5.3	Multivariate Classification Techniques	162
5.3.1	Likelihood (Naive Bayes Classifier)	162
5.3.2	<i>k</i> -Nearest Neighbour and Multi-dimensional Likelihood	163
5.3.3	Fisher Linear Discriminant	165
5.3.4	Artificial Neural Networks – Feed-Forward Multi-layer Perceptrons	168
5.3.5	Support Vector Machines	172
5.3.6	(Boosted) Decision Trees	178
5.3.7	Boosting and Bagging	179
5.3.7.1	Boosting	179
5.3.7.2	Adaptive Boost (ADABoost)	180
5.3.7.3	Bagging	181
5.4	General Remarks	182
5.4.1	Pre-processing	182
5.5	Dealing with Systematic Uncertainties	183
5.6	Exercises	184
	References	186

<b>6</b>	<b>Unfolding</b>	187
	<i>Volker Blobel</i>	
6.1	Inverse Problems	187
6.1.1	Direct and Inverse Processes	187
6.1.2	Discretisation and Linear Solution	189
6.1.3	Unfolding Poisson-Distributed Data	192
6.1.4	Convolution and Deconvolution	193
6.1.5	Parametrised Unfolding	195
6.2	Solution with Orthogonalisation	196
6.2.1	Singular Value and Eigenvalue Decomposition	196
6.2.2	Unfolding Using the Least Squares Method	199
6.2.2.1	Least-Squares Method Using the SVD	199
6.2.2.2	Null Space and Truncated SVD	199
6.2.2.3	Truncation and Positive Correlations	202
6.2.3	Folding Versus Unfolding	202
6.3	Regularisation Methods	203
6.3.1	Norm and Derivative Regularisation	203
6.3.1.1	Regularisation	203
6.3.1.2	Norm Regularisation	204
6.3.1.3	Regularisation Based on Derivatives	206
6.3.1.4	Determination/Selection of the Regularisation Parameter	207
6.4	The Discrete Cosine Transformation and Projection Methods	209
6.4.1	Discrete Cosine Transformation	210
6.4.2	Projection Methods	211
6.4.3	Low-Pass Regularisation	212
6.5	Iterative Unfolding	213
6.6	Unfolding Problems in Particle Physics	215
6.6.1	Particle Physics Experiments	215
6.6.2	Unfolding Smooth Distributions	218
6.6.3	Unfolding Non-smooth Distributions	219
6.6.4	Presentation of Regularisation Results	220
6.7	Programs Used for Unfolding in High Energy Physics	221
6.8	Exercise	223
	References	223
<b>7</b>	<b>Constrained Fits</b>	227
	<i>Benno List</i>	
7.1	Introduction	227
7.2	Solution by Elimination	230
7.2.1	Statistical Interpretation	231
7.3	The Method of Lagrange Multipliers	232
7.3.1	Lagrange Multipliers	233
7.3.2	Unmeasured Parameters	236
7.4	The Lagrange Multiplier Problem with Linear Constraints and Quadratic Objective Function	237

7.4.1	Error Propagation	239
7.4.2	Error Propagation in the Presence of Unmeasured Quantities	241
7.5	Iterative Solution of the Lagrange Multiplier Problem	244
7.5.1	Choosing a Direction	245
7.5.1.1	Coping with a Rank-Deficient Matrix $\mathbf{M}$	248
7.5.1.2	Normalisation of Parameters and Constraint Functions	249
7.5.2	Controlling the Step Length	250
7.5.2.1	Merit Function	251
7.5.2.2	Line Searches	252
7.5.2.3	Stopping Conditions	253
7.5.2.4	Practical Choice of the Step Length	254
7.5.2.5	The Maratos Effect	255
7.5.3	Detecting Convergence	256
7.5.3.1	Stopping the Iterations	257
7.5.3.2	Verifying Convergence	258
7.5.4	Finding Initial Values	258
7.5.5	Error Calculation	259
7.6	Further Reading and Web Resources	259
7.7	Exercises	260
	References	261
<b>8</b>	<b>How to Deal with Systematic Uncertainties</b>	263
	<i>Rainer Wanke</i>	
8.1	Introduction	263
8.2	What Are Systematic Uncertainties?	264
8.3	Detection of Possible Systematic Uncertainties	265
8.3.1	Top-Down Approach	265
8.3.2	Bottom-Up Approach	266
8.3.3	Examples for Detecting Systematics	267
8.3.3.1	Background Systematics	267
8.3.3.2	Detector Acceptances	268
8.3.3.3	Splitting Data into Independent Subsets	269
8.3.3.4	Evaluating the Result in Intervals of an Analysis Parameter	270
8.3.3.5	Analysis Software and Fit Routines	271
8.4	Estimation of Systematic Uncertainties	272
8.4.1	Some Simple Cases	273
8.4.1.1	External Input Parameters	273
8.4.1.2	Tolerances	273
8.4.1.3	Small Systematics	273
8.4.2	Educated Guesses	273
8.4.2.1	Background Estimation	274
8.4.2.2	Detector Resolutions	275
8.4.2.3	Theory Uncertainties	276
8.4.2.4	Discrepancies between Data and Simulation	277
8.4.2.5	Analyses with Small Statistics	278

8.4.3	Cut Variations	279
8.4.3.1	Cut Variations in Multi-Dimensional Analyses	283
8.4.4	Combination of Systematic Uncertainties	285
8.4.4.1	Combination of Covariance Matrices	287
8.5	How to Avoid Systematic Uncertainties	288
8.5.1	Choice of Selection Criteria	288
8.5.2	Avoiding Biases	289
8.5.2.1	Do not Expect a Certain Result	289
8.5.2.2	Do not Look at Signal Events	291
8.5.3	Blind Analyses	292
8.6	Conclusion	293
8.7	Exercise	295
	References	296
<b>9</b>	<b>Theory Uncertainties</b>	297
	<i>Markus Diehl</i>	
9.1	Overview	297
9.2	Factorisation: A Cornerstone of Calculations in QCD	298
9.2.1	The Perturbative Expansion and Uncertainties from Higher Orders	301
9.2.1.1	The Renormalisation Scale	301
9.2.1.2	The Factorisation Scale	305
9.2.1.3	Combining Different Orders	307
9.2.1.4	Multi-Scale Problems and Resummation Methods	307
9.3	Power Corrections	308
9.3.1	Operator Product Expansion	308
9.3.2	Power Corrections in Cross Sections	309
9.4	The Final State	310
9.4.1	Underlying Event and Multi-Parton Interactions	310
9.4.2	From Partons to Hadrons	311
9.4.3	Monte Carlo Event Generators	313
9.5	From Hadrons to Partons	314
9.5.1	Parametric PDF Uncertainties	318
9.5.1.1	The Hessian Matrix	319
9.5.1.2	Lagrange Multipliers	321
9.5.1.3	The NNPDF Approach	321
9.5.2	A Comparison of Recent PDF Sets	322
9.6	Exercises	324
	References	326
<b>10</b>	<b>Statistical Methods Commonly Used in High Energy Physics</b>	329
	<i>Carsten Hensel and Kevin Kröninger</i>	
10.1	Introduction	329
10.2	Estimating Efficiencies	329
10.2.1	Motivation	330
10.2.2	Trigger Efficiencies and Their Estimates	330
10.2.3	The Counting Method	331

10.2.4	The Tag-and-Probe Method	331
10.2.5	The Bootstrap Method	332
10.2.6	Calculating Uncertainties on Trigger Efficiencies	333
10.3	Estimating the Contributions of Processes to a Dataset: The Matrix Method	334
10.3.1	Estimating the Background Contributions to a Data Sample	334
10.3.2	Extension to Distributions	336
10.3.3	Limitations of the Matrix Method	337
10.4	Estimating Parameters by Comparing Shapes of Distributions: The Template Method	337
10.4.1	Template Shapes	340
10.4.2	Including Prior Knowledge	340
10.4.3	Including Efficiencies	341
10.4.4	Including Systematic Uncertainties	341
10.4.5	Systematic Uncertainties Due to the Fitting Procedure	343
10.4.6	Alternative Fitting Methods and Choice of Parameters	344
10.4.7	Extension to Multiple Channels and Multi-Dimensional Templates	344
10.5	Ensemble Tests	345
10.5.1	Generation of Ensembles	346
10.5.2	Results of Ensemble Tests	347
10.6	The Experimenter's Role and Data Blinding	351
10.6.1	The Experimenter's Preconception	352
10.6.2	Variants of Blind Analyses	352
10.7	Exercises	354
	References	356
<b>11</b>	<b>Analysis Walk-Throughs</b>	<b>357</b>
	<i>Aart Heijboer and Ivo van Vulpen</i>	
11.1	Introduction	357
11.2	Search for a $Z'$ Boson Decaying into Muons	357
11.2.1	Counting Experiment	358
11.2.1.1	Quantifying the Sensitivity: $p$ -Values and Significance	358
11.2.1.2	Optimising the Mass Window	360
11.2.1.3	Estimating the Background from Data Using Sidebands	360
11.2.1.4	Scanning over the Full Dimuon Mass Range: The ‘Look-Elsewhere Effect’	361
11.2.2	Profile Likelihood Ratio Analysis	362
11.2.2.1	Profile Likelihood Test Statistic	362
11.2.2.2	Properties of the Test Statistic Distributions for the $b$ -only and $s + b$ Hypotheses	363
11.2.2.3	Rules for Discovery and Exclusion	364
11.2.2.4	Results from Data	366
11.2.2.5	Probing the Sensitivity Limits: Enhanced Luminosity and Signal Cross Sections	366
11.2.2.6	Scanning the Full Mass Region	367

11.3	Measurement	369
11.3.1	Introduction	369
11.3.2	Unbinned Likelihood	370
11.3.2.1	Likelihood Ingredients	371
11.3.3	Extracting a Measurement in the Presence of Nuisance Parameters	372
11.3.4	Mass Measurement	373
11.3.5	Testing for Bias and Coverage	373
11.3.6	Systematic Uncertainties	374
11.3.7	Constraints and Combining Measurements	375
11.4	Exercises	377
	References	379
<b>12</b>	<b>Applications in Astronomy</b>	<b>381</b>
	<i>Harrison B. Prosper</i>	
12.1	Introduction	381
12.2	A Survey of Applications	382
12.2.1	The On/Off Problem	383
12.2.1.1	A Bayesian Approach	385
12.2.1.2	Conclusion	390
12.2.2	Image Reconstruction	390
12.2.2.1	Of Monkeys and Kangaroos	391
12.2.2.2	Maximum Entropy Method in Practice	393
12.2.3	Fitting Cosmological Parameters	394
12.2.3.1	Cosmology in a Nutshell	394
12.2.3.2	Statistical Analysis of Supernovae Data	397
12.2.3.3	Model Complexity	399
12.3	Nested Sampling	401
12.4	Outlook and Conclusions	404
12.5	Exercises	405
	References	405
	<b>The Authors</b>	<b>409</b>
	<b>Index</b>	<b>413</b>

## Preface

Statistical inference plays a crucial role in the exact sciences. In fact, many results can only be obtained with the help of sophisticated statistical methods. In our field of experimental particle physics, statistical reasoning enters into basically every step of our data analysis work.

Recent years have seen the development of many new statistical techniques and of complex software packages implementing these. Consequently, the requirements on the statistics knowledge for scientists in high energy physics have increased dramatically, as have the needs for education and documentation in this field. This book aims at contributing to this purpose. It targets a broad readership at all career levels, from students to senior researchers, and is intended to provide comprehensive and practical advice for the various statistical analysis tasks typically encountered in high energy physics. To achieve this, the book is split into 12 chapters, all written by a different expert author or team of two authors and focusing on a well-defined topic:

- *Fundamental Concepts* introduces the basics of statistical data analyses, such as probability density functions and their properties, theoretical distributions (Gaussian, Poisson and many others) and concepts of probability (frequentist and Bayesian reasoning).

The next chapters elucidate the basic tools used to infer results from data:

- *Parameter Estimation* illustrates how to determine the best parameter values of a model from fitting data, for example how to estimate the strength of a signal.
- *Hypothesis Testing* lays out the framework that can be used to decide on hypotheses such as ‘the data can be explained by fluctuations of the known background sources alone’ or ‘the model describes the data reasonably well’.
- *Interval Estimation* discusses how to determine confidence or credibility intervals for parameter values, for example upper limits on the strength of a signal.

The following chapters deal with more advanced tasks encountered frequently:

- *Classification* presents various methods to optimally discriminate different event classes, for example signal from background, using multivariate data input. These methods can be very useful to enhance the sensitivity of a measurement, for example to find and measure a signal in the data that is otherwise drowned in background.
- *Unfolding* describes strategies and methods for correcting data for the usually inevitable effects of detector bias, acceptance, and resolution, which in particular can be applied in measurements of differential distributions.
- *Constrained Fits* discusses how to exploit physical constraints, such as energy–momentum conservation, to improve measurements or to determine unknown parameters.

The determination of systematic uncertainties is a key task for any measurement that is often performed as the very last step of a data analysis. We feel that it is worthwhile to discuss this – often neglected – topic in two chapters:

- *How to Deal with Systematic Uncertainties* elucidates how to detect and avoid sources of systematic uncertainties and how to estimate their impact.
- *Theory Uncertainties* illuminates various aspects of theoretical uncertainties, in particular for the strong interaction.

The following three chapters complete the book:

- *Statistical Methods Commonly Used in High Energy Physics* introduces various practical analysis tools and methods, such as the template and matrix methods for the estimation of sample compositions, or the determination of biases of analysis procedures by means of ensemble tests.
- *Analysis Walk-Throughs* provides a synopsis of the book by going through two complete analysis examples – a search for a new particle and a measurement of the properties of this hypothetical new particle.
- *Applications in Astronomy* takes us on a journey to the field of astronomy and illustrates, with several examples, the sophisticated data analysis techniques used in this research area.

In all chapters, care has been taken to be as practical and concrete as the material allows – for this purpose many specifically designed examples have been inserted into the text body of the chapters. A further deepening of the understanding of the book material can be achieved with the dedicated exercises at the end of all chapters. Hints and solutions to the exercises, together with some necessary software, are available from a webpage provided by the publisher. Here, we will also collect feedback, corrections and other information related to this volume; please check [www.wiley.com](http://www.wiley.com) for the details.

Many people have contributed to this book, and we would like to thank all of them. First of all, we thank the authors of the individual chapters for the high-quality material they provided.

Besides the authors, a number of people are needed to successfully conclude a book project like this one: numerous colleagues contributed by means of discussion, by providing expert advice and answers to our questions. We cannot name them all.

Katarina Brock spent many hours editing and polishing all the figures and providing a unified layout for them. Konrad Kieling from Wiley provided valuable support in typesetting the book. Vera Palmer and Ulrike Werner from Wiley provided constant support in all questions related to this book. We thank Tatsuya Nakada for his permission to use his exercise material.

Our last and very heartfelt thanks goes to our friends, partners and families who endured, over a considerable period, the very time- and also nerve-consuming genesis of this book. Without their support and tolerance this book would not exist today.

All comments, criticisms and questions you might have on the book are welcome – please send them to the authors via email:

*olaf.behnke@desy.de,  
kevin.kroeninger@phys.uni-goettingen.de,  
thomas.schoerner@desy.de,  
gregory.schott@cern.ch.*

Hamburg  
Göttingen  
Karlsruhe  
November 2012

*Olaf Behnke  
Kevin Kröninger  
Thomas Schörner-Sadenius and  
Grégory Schott*

## List of Contributors

*Roger Barlow*  
 University of Huddersfield  
 Huddersfield  
 United Kingdom

*Olaf Behnke*  
 DESY  
 Hamburg  
 Germany

*Volker Blobel*  
 Universität Hamburg  
 Hamburg  
 Germany

*Luc Demortier*  
 The Rockefeller University  
 New York, New York  
 United States of America

*Markus Diehl*  
 DESY  
 Hamburg  
 Germany

*Aart Heijboer*  
 Nikhef  
 Amsterdam  
 Netherlands

*Carsten Hensel*  
 Universität Göttingen  
 II. Physikalisches Institut  
 Göttingen  
 Germany

*Kevin Kröninger*  
 Universität Göttingen  
 II. Physikalisches Institut  
 Göttingen  
 Germany

*Benno List*  
 DESY  
 Hamburg  
 Germany

*Lorenzo Moneta*  
 CERN  
 Geneva  
 Switzerland

*Harrison B. Prosper*  
 Florida State University  
 Tallahassee, Florida  
 United States of America

**Grégory Schott**

Karlsruher Institut für Technologie  
Institut für Experimentelle Kernphysik  
Karlsruhe  
Germany

**Ivo van Vulpen**

Nikhef  
Amsterdam  
Netherlands

**Helge Voss**

Max-Planck-Institut für Kernphysik  
Heidelberg  
Germany

**Rainer Wanke**

Institut für Physik  
Universität Mainz  
Mainz  
Germany

# 1

## Fundamental Concepts

*Roger Barlow*

### 1.1 Introduction

Particle physics is all about random behaviour. When two particles collide, or even when a single particle decays, we can't predict with certainty what will happen, we can only give probabilities of the various different outcomes. Although we measure the lifetimes of unstable particles and quote them to high precision – for the  $\tau$  lepton, for example, it is  $0.290 \pm 0.001$  ps – we cannot say exactly when a particular  $\tau$  will decay: it may well be shorter or longer. Although we know the probabilities (called, in this context, branching ratios) for the different decay channels, we can't predict how any particular  $\tau$  will decay – to an electron, or a muon, or various hadrons.

Then, when particles travel through a detector system they excite electrons in random ways, in the gas molecules of a drift chamber or the valence band of semi-conducting silicon, and these electrons will be collected and amplified in further random processes. Photons and phototubes are random at the most basic quantum level. The experiments with which we study the properties of the basic particles are random through and through, and a thorough knowledge of that fundamental randomness is essential for machine builders, for analysts, and for the understanding of the results they give.

It was not always like this. Classical physics was deterministic and predictable. Laplace could suggest a hypothetical demon who, aware of all the coordinates and velocities of all the particles in the Universe, could then predict all future events. But in today's physics the demon is handicapped not only by the uncertainties of quantum mechanics – the impossibility of knowing both coordinates and velocities – but also by the greater understanding we now have of chaotic systems. For predicting the flight of cannonballs or the trajectories of comets it was assumed, as a matter of common sense, that although our imperfect information about the initial conditions gave rise to increasing inaccuracy in the predicted motion, better information would give rise to more accurate predictions, and that this process could continue without limit, getting as close as one needed (and could afford) to

perfect prediction. We now know that this is not true even for some quite simple systems, such as the compound pendulum.

That is only one of the two ways that probability comes into our experiments. When a muon passes through a detector it may, with some probability, produce a signal in a drift chamber: the corresponding calculation is a *prediction*. Conversely a drift chamber signal may, with some probability, have been produced by a muon, or by some other particle, or just by random noise. To interpret such a signal is a process called *inference*. Prediction works forwards in time and inference works backwards. We use the same mathematical tool – *probability* – to cover both processes, and this causes occasional confusion. But the statistical processes of inference are, though less visibly dramatic, of vital concern for the analysis of experiments. Which is what this book is about.

## 1.2

### Probability Density Functions

The outcomes of random processes may be described by a variable (or variables) which can be *discrete* or *continuous*, and a discrete variable can be *quantitative* or *qualitative*. For example, when a  $\tau$  lepton decays it can produce a muon, an electron, or hadrons: that's a qualitative difference. It may produce one, three or five charged particles: that's quantitative and discrete. The visible energy (i.e. not counting neutrinos) may be between 0 and 1777 MeV: that's quantitative and continuous.

The probability prediction for a variable  $x$  is given by a function: we can call it  $f(x)$ . If  $x$  is discrete then  $f(x)$  is itself a probability. If  $x$  is continuous then  $f(x)$  has the dimensions of the inverse of  $x$ : it is  $\int f(x)dx$  that is the dimensionless probability, and  $f(x)$  is called a *probability density function* or *pdf*.<sup>1)</sup> There are clearly an infinite number of different pdfs and it is often convenient to summarise the properties of a particular pdf in a few numbers.

#### 1.2.1

##### Expectation Values

If the variable  $x$  is quantitative then for any function  $g(x)$  one can form the average

$$E[g] = \int g(x) f(x)dx \quad \text{or, as appropriate,} \quad E[g] = \sum g(x) f(x), \quad (1.1)$$

where the integral (for continuous  $x$ ) or the sum (for discrete  $x$ ) covers the whole range of possible values. This is called the *expectation value*. It is also sometimes written  $\langle g \rangle$ , as in quantum mechanics. It gives the mean, or average, value of  $g$ , which is not necessarily the most likely one – particularly if  $x$  is discrete.

1) The parton density functions of QCD, PDFs, share the abbreviation and are indeed pdfs in both senses.

### 1.2.2

#### Moments

For any pdf  $f(x)$ , the integer powers of  $x$  have expectation values. These are called the (algebraic) *moments* and are defined as

$$\alpha_n = E[x^n]. \quad (1.2)$$

The first moment,  $\alpha_1$ , is called the *mean* or, more properly, *arithmetic mean* of the distribution; it is usually called  $\mu$  and often written  $\bar{x}$ . It acts as a key measure of *location*, in cases where the variable  $x$  is distributed with some known shape about a particular point.

Conversely there are cases where the shape is what matters, and the absolute location of the distribution is of little interest. For these it is useful to use the *central moments*

$$m_n = E[(x - \mu)^n]. \quad (1.3)$$

##### 1.2.2.1 Variance

The second central moment is also known as the *variance*, and its square root as the *standard deviation*:

$$V[x] = \sigma^2 = m_2 = E[(x - \mu)^2]. \quad (1.4)$$

The variance is a measure of the width of a distribution. It is often easier to deal with algebraically whereas the standard deviation  $\sigma$  has the same dimensions as the variable  $x$ ; which to use is a matter of personal choice. Broadly speaking, statisticians tend to use the variance whereas physicists tend to use the standard deviation.

##### 1.2.2.2 Skew and Kurtosis

The third and fourth central moments are used to build shape-describing quantities known as *skew* and *kurtosis* (or *curtosis*):

$$\gamma_1 = \frac{m_3}{\sigma^3}, \quad (1.5)$$

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3. \quad (1.6)$$

Division by the appropriate power of  $\sigma$  makes these quantities dimensionless and thus independent of the scale of the distribution, as well as of its location. Any symmetric distribution has zero skew: distributions with positive skew have a tail towards higher values, and conversely negative skew distributions have a tail towards lower values. The Poisson distribution has a positive skew, the energy recorded by a calorimeter has a negative skew. A Gaussian has a kurtosis of zero – by definition, that's why there is a '3' in the formula. Distributions with positive

kurtosis (which are called *leptokurtic*) have a wider tail than the equivalent Gaussian, more centralised or *platykurtic* distributions have negative kurtosis. The Breit–Wigner distribution is leptokurtic, as is Student's *t*. The uniform distribution is platykurtic.

### 1.2.2.3 Covariance and Correlation

Suppose you have a pdf  $f(x, y)$  which is a function of two random variables,  $x$  and  $y$ . You can not only form moments for both  $x$  and  $y$ , but also for combinations, particularly the *covariance*

$$\text{cov}[x, y] = E[x y] - E[x]E[y]. \quad (1.7)$$

If the joint pdf is factorisable:  $f(x, y) = f_x(x) \cdot f_y(y)$ , then  $x$  and  $y$  are independent, and the covariance is zero (although the converse is not necessarily true: a zero covariance is a necessary but not a sufficient condition for two variables to be independent).

A dimensionless version of the covariance is the *correlation*  $\rho$ :

$$\rho = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}. \quad (1.8)$$

The magnitude of the correlation lies between 0 (uncorrelated) and 1 (completely correlated). The sign can be positive or negative: amongst a sample of students there will probably be a positive correlation between height and weight, and a negative correlation between academic performance and alcohol consumption.

If there are several (i.e. more than two) variables,  $x_1, x_2, \dots, x_N$ , one can form the *covariance* and *correlation matrices*:

$$V_{ij} = \text{cov}[x_i, x_j] = E[x_i x_j] - E[x_i]E[x_j], \quad (1.9)$$

$$\rho_{ij} = \frac{V_{ij}}{\sigma_i \sigma_j}, \quad (1.10)$$

and  $V_{ii}$  is just  $\sigma_i^2$ .

### 1.2.2.4 Marginalisation and Projection

Mathematically, any pdf  $f(x, y)$  is a function of two variables  $x$  and  $y$ . They can be similar in nature, for example the energies of the two electrons produced by a converting high energy photon, or they can be different, for example the position and direction of particles undergoing scattering in material.

Often we are really interested in one parameter (say  $x$ ) while the other (say  $y$ ) is just a *nuisance parameter*. We want to reject the extra information shown in the two-dimensional function (or scatter plot). This can be done in two ways: the *projection* of  $x$ ,  $f(x)|_y$  is obtained by choosing a particular value of  $y$ , the *marginal distribution*  $f(x) = \int f(x, y)dy$  is found by integrating over  $y$ .

Projections can be useful for illustration, otherwise to be meaningful you have to have a good reason for choosing that specific value of  $y$ . Marginalisation requires that the distribution in  $y$ , like that of  $x$ , is properly normalised.

### 1.2.2.5 Other Properties

There are many other properties that can be quoted, depending on the point we want to bring out, and on the established usage of the field.

The mean is not always the most helpful measure of location. The *mode* is the value of  $x$  at which the pdf  $f(x)$  is maximum, and if you want a typical value to quote it serves well. The *median* is the midway point, in the sense that half the data lie above and half below. It is useful in describing very skewed distributions (particularly financial income) in which fluctuations in a small tail would give a big change in the mean.

We can also specify dispersion in ways that are particularly useful for non-Gaussian distributions by using *quantiles*: the upper and lower *quartiles* give the values above which, and below which, 25% of the data lie. *Deciles* and *percentiles* are also used.

### 1.2.3

#### Associated Functions

The *cumulative distribution function*

$$F(a) = \int_{-\infty}^a f(x)dx \quad \text{or, as appropriate,} \quad F(a) = \sum f(x_i) \Theta(a - x_i), \quad (1.11)$$

where  $\Theta$  is the Heaviside or step function ( $\Theta(x) = 1$  for  $x \geq 0$  and 0 otherwise), giving the probability that a variable will take a value up to  $a$ , is occasionally useful.

The *characteristic function*

$$\phi(u) = E[e^{iux}] = \int e^{iux} f(x)dx, \quad (1.12)$$

which is just (up to factors of  $2\pi$ ) the Fourier transform of the pdf, is also met with sometimes as it has useful properties.

### 1.3

#### Theoretical Distributions

A pdf is a mathematical function. It involves a variable (or variables) describing the random quantity concerned. This may be a discrete integer or a continuous real number. It also involves one or more parameters. In what follows we will denote a random variable by  $x$  for a real number and  $r$  for an integer. Parameters generally have their traditional symbols for particular pdfs: where we refer to a generic parameter we will call it  $\theta$ . It is often helpful to write a function as  $f(x; \theta)$  or  $f(x|\theta)$ , separating this way more clearly the random variable(s) from the adjustable parameter(s). The semicolon is preferred by some, the line has the advan-

tage that it matches the notation used for conditional probabilities, described in Section 1.4.4.1.

There are many pdfs in use to model the results of random processes. Some are based on physical motivations, some on mathematics, and some are just empirical forms that happen to work well in particular cases.

The overwhelmingly most useful form is the *Gaussian* or *normal* distribution. The *Poisson* distribution is also encountered very often, and the *binomial* distribution is not uncommon. So we describe these in some detail, and then some other distributions rather more briefly.

### 1.3.1

#### The Gaussian Distribution

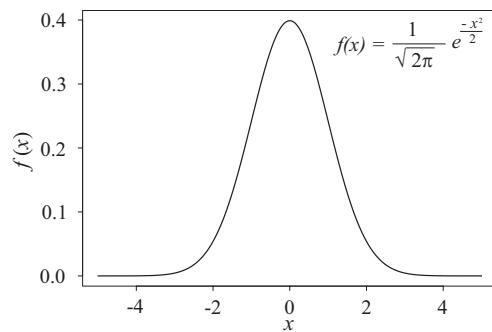
The Gaussian, or normal, distribution for a continuous random variable  $x$  is given by

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (1.13)$$

It has two parameters; the function is manifestly symmetrical about the location parameter  $\mu$ , which is the mean (and mode, and median) of the distribution. The scale parameter  $\sigma$  is also the standard deviation of the distribution. So there is, in a sense, only one Gaussian, the *unit Gaussian* or *standard normal distribution*  $f(x; 0, 1)$  shown in Figure 1.1. Any other Gaussian can be obtained from this by scaling by a factor  $\sigma$  and translating by an amount  $\mu$ . The Gaussian distribution is sometimes denoted  $\mathcal{N}(x; \mu, \sigma)$ .

The Gaussian is ubiquitous (hence the name ‘normal’) because of the *central limit theorem*, which states that if any distribution is convoluted with itself a large number of times, the resulting distribution tends to a Gaussian form. For a proof, see for example Appendix 2 in [1].

Gaussian random numbers are much used in simulation, and a suitable random number generator is available on most systems. If it is not, then you can generate a unit Gaussian by taking two uniformly generated random numbers  $u_1, u_2$ , set



**Figure 1.1** The unit Gaussian or standard normal distribution.

$\theta = 2\pi u_1$ ,  $r = \sqrt{-2 \ln u_2}$ , and then  $r \cos \theta$  and  $r \sin \theta$  are independent samples from a unit Gaussian.

The product of two independent Gaussians gives a two-dimensional function

$$f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}, \quad (1.14)$$

but the most general quadratic form in the exponent must include the cross term and can be written as

$$f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}, \quad (1.15)$$

where the parameter  $\rho$  is the correlation between  $x$  and  $y$ . For  $N$  variables, for which we will use the vector  $\mathbf{x}$ , the full form of the multivariate Gaussian can be compactly written using matrix notation:

$$f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V}) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}|} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x}-\boldsymbol{\mu})}. \quad (1.16)$$

Here,  $\mathbf{V}$  is the covariance matrix described in Section 1.2.2.3.

The *error function* and the *complementary error function* are basically closely related to the cumulative Gaussian

$$\text{erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-x^2} dx, \quad (1.17)$$

$$\text{erfc}(y) = \frac{2}{\sqrt{\pi}} \int_y^\infty e^{-x^2} dx. \quad (1.18)$$

Their main use is in calculating Gaussian *p-values* (see Section 1.3.4.6). The probability that a Gaussian random variable will lie within one standard deviation, or ‘1  $\sigma$ ’, of the mean is 68% obtained by calculating  $\text{erf}(y = 1)$ . Conversely, the chance that a variable drawn from a Gaussian random process will lie outside 1  $\sigma$  is 32%. Given such a process – say a mean of 10.2 and a standard deviation of 3.1 – then if you confront a particular measurement – say 13.3 – it is quite plausible that it was produced by the process. One says that its *p-value*, the probability that the process would produce a measurement this far, or further, from the ideal mean, is 32%. Conversely, if the number were 25.7 rather than 13.3, that would be 5  $\sigma$  rather than 1  $\sigma$ , for which the *p-value* is only  $5.7 \cdot 10^{-7}$ . In discussion of discoveries (or otherwise) of new particles and new effects this language is turned round, and a discovery with a *p-value* of  $5.7 \cdot 10^{-7}$  is referred to as a ‘5  $\sigma$  result’<sup>2)</sup>. A translation is given in Table 1.1 – although for practical purposes it is easier to use functions such as `pnorm` and `qnorm` in the programming language `R` [2], or `TMath::Prob` in `ROOT` [3].

2) Note that there is a subtle difference between a one-sided and two-sided *p-value*. Details will be discussed in Chapter 3

**Table 1.1** Two-sided Gaussian  $p$ -values for  $1\sigma$  to  $5\sigma$  deviations.

Deviation	$p$ -value (%)
$1\sigma$	31.7
$2\sigma$	4.56
$3\sigma$	0.270
$4\sigma$	0.00633
$5\sigma$	0.0000573

## 1.3.2

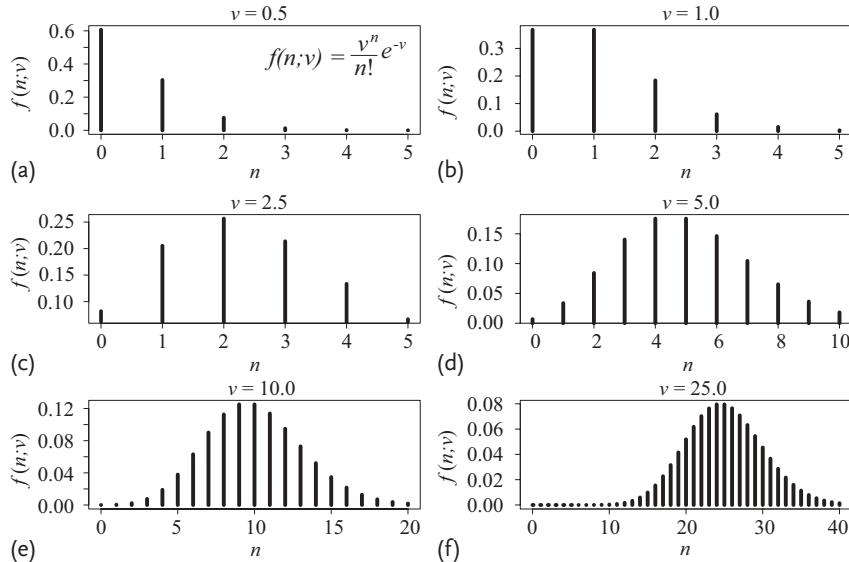
**The Poisson Distribution**

The Poisson distribution

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu} \quad (1.19)$$

describes the probability of  $n$  events occurring when the mean expected number is  $\nu$ ;  $n$  is discrete and  $\nu$  is continuous. Typical examples are the number of clicks produced by a Geiger counter in an interval of time, or, famously, the number of Prussian cavalrymen killed by horse-kicks [4]. Some examples are shown in Figure 1.2.

The Poisson distribution has a mean of  $\nu$  and a standard deviation  $\sigma = \sqrt{\nu}$ . This property – that the standard deviation is the square root of the mean – is a key



**Figure 1.2** Poisson distributions for (a)  $\nu = 0.5$ , (b)  $\nu = 1.0$ , (c)  $\nu = 2.5$ , (d)  $\nu = 5.0$ , (e)  $\nu = 10.0$ , (f)  $\nu = 25.0$ .

fact about distributions generated by a Poisson process, which is important as this includes most cases where a number of samples is taken, including the contents of the bin of a histogram.

### Example 1.1 Counting cosmic muons

In an experiment built to measure cosmic muons, the expected rate of muons in one run of the experiment is 0.45 events. This means that you have a 64% probability of observing no decays, a 29% probability of a single decay, 6% chance of seeing two and less than 1% of seeing three.

#### 1.3.3

#### The Binomial Distribution

The binomial distribution describes a generalisation of the simple problem of the numbers of heads and tails that can arise from spinning a coin several times. The probability for getting  $r$  ‘successes’ from  $N$  ‘trials’ given an intrinsic probability of success  $p$  is

$$f(r; N, p) = \frac{N!}{r!(N-r)!} p^r (1-p)^{n-r}. \quad (1.20)$$

Sometimes one writes  $q$  instead of  $1 - p$ , which makes the algebra prettier. The distribution has a mean of  $Np$  and a standard deviation  $\sigma = \sqrt{Np(1-p)} = \sqrt{Npq}$ . The factor  $N!/[r!(N-r)!]$  is the number of ways that  $r$  objects may be chosen from  $N$ , and is often written  $\binom{N}{r}$ .

### Example 1.2 Tracking chambers

A charged particle in an experiment goes through a set of six tracking chambers, which measure its position. Each of them is 95% efficient. If you require all six chambers to register a hit in order to define a reconstructed track, the efficiency of the system will clearly be  $0.95^6 = 73.5\%$ . If you are satisfied with five or more hits the efficiency is 96.7%. If at least four hits are enough, the track efficiency is 99.8%.

If  $p$  is small then the distribution can be approximated by a Poisson distribution<sup>3)</sup> of mean  $Np$ . This is often used implicitly when analysing Monte Carlo samples: if you generate 1 000 000 Monte Carlo events, of which 100 end up in some particular histogram bin, then strictly speaking this is described by a binomial process rather than a Poisson. In practice you can take the error as the Poisson  $\sqrt{100}$  rather than a binomial  $\sqrt{1 000 000 \cdot 0.0001 \cdot 0.9999}$ . This doesn’t work if  $p$  is large. If 9 out of 10 events are accepted by the trigger, the error on the trigger efficiency of 90% is not

3) Indeed the Poisson can be derived as the limit of the binomial as  $N \rightarrow \infty$ ,  $p \rightarrow 0$  with  $Np$  constant.

$\sqrt{9}/10 = 30\%$  but  $\sqrt{0.9 \cdot 0.1/10} = 9.5\%$  (in such a case the shortcut is to take the one lost event as approximately Poisson, giving the error as 10%, which is close).

If  $N$  is large and  $p$  is not small then the distribution is approximately a Gaussian.

If there are not just two possible outcomes but  $n$ , with probabilities  $\{p_1, p_2, \dots, p_n\}$ , then the total probability of getting  $r_1$  of the first outcome,  $r_2$  of the second, and so on, is

$$f(r_1, r_2, \dots, r_n; N, p_1, p_2, \dots, p_n) = \frac{N!}{\prod r_i!} \prod p_i^{r_i}. \quad (1.21)$$

This is the *multinomial distribution*.

### 1.3.4

#### Other Distributions

There are many, many other possible distribution functions, and it is worth listing some of those more often met with.

##### 1.3.4.1 The Uniform Distribution

The *uniform distribution*, also known as the *rectangular* or *top-hat* distribution, is constant inside some range – call this range  $-a/2$  to  $+a/2$ , so the width is  $a$ ; if the range is not central about zero but about some other value this is easily done by a translation. The mean, clearly, is zero, and the standard deviation is  $a/\sqrt{12}$ . This can be used in position measurements by a hodoscope: if a rectangular slab of scintillator gives a signal, you know that a track went through it but you do not know where. It is reasonable to assume a uniform distribution for the pdf of the hit position.

This can be relevant in considering some systematic uncertainties on the total result, as is also discussed in Section 8.4.1.2. For example, if you set up an experiment to run overnight, counting events with some efficiency  $E_1$ , and when you arrive in the morning you find a component has tripped so the efficiency is  $E_2$ , with no information about when this happened, your efficiency has to be quoted as  $(E_1 + E_2)/2 \pm (|E_1 - E_2|)/\sqrt{12}$ . It can also be applied to theoretical models: when two models give different predictions you are justified in using their mean as your prediction, with a (systematic) error which is the difference divided by  $\sqrt{12}$ , if (and only if) these two models represent absolute extremes and you really have no feeling as to where between the two extremes the truth may lie.

##### 1.3.4.2 The Cauchy, or Breit–Wigner, or Lorentzian Distribution

In nuclear and particle physics the function

$$f(E; M, \Gamma) = \frac{1}{2\pi} \frac{\Gamma}{(E - M)^2 + (\Gamma/2)^2} \quad (1.22)$$

gives the variation with the energy  $E$  of a cross section produced by the formation of a state with mass  $M$  and width  $\Gamma$ . It can be written more neatly in dimensionless

form as

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad (1.23)$$

where  $x = (E - M)/(\Gamma/2)$ . The mean is clearly  $M$ . It does not have a variance: the integral  $\int x^2 f(x)dx$  is divergent. If you have to compare this curve and with that of a Gaussian, the *full width at half maximum* (FWHM) is clearly  $\Gamma$  for this curve and for a Gaussian it is  $2\sqrt{2 \ln 2}\sigma = 2.35\sigma$ .

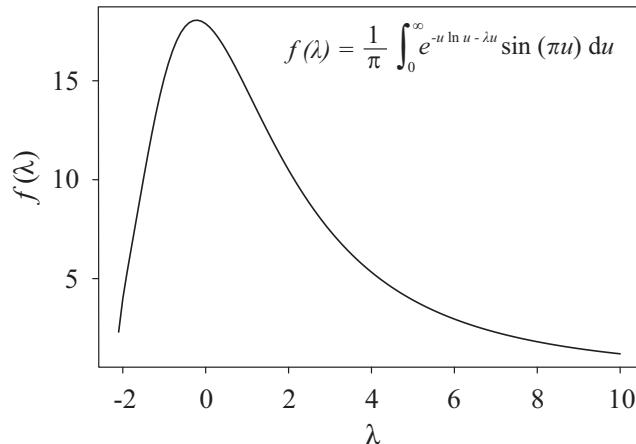
This distribution is used in fitting resonance peaks (provided the width is much larger than the measurement error on  $E$ ). It also has an empirical use in fitting a set of data which is almost Gaussian but has wider tails. This often arises in cases where a fraction of the data is not so well measured as the rest. A double Gaussian may give a good fit, but it often turns out that this form does an adequate job without the need to invoke extra parameters.

#### 1.3.4.3 The Landau Distribution

When a charged particle passes an atom, its electrons experience a changing electromagnetic field and acquire energy. The amount of energy may be large; on rare occasions it will be large enough to create a delta ray. The probability distribution for the energy loss was computed by Landau [5] and is given by

$$f(\lambda) = \frac{1}{\pi} \int_0^\infty e^{-u \ln u - \lambda u} \sin(\pi u) du, \quad (1.24)$$

where  $\lambda = (\Delta - \Delta_0)/\xi$ . Here,  $\Delta$  is the actual energy loss,  $\Delta_0$  is a location parameter, and  $\xi$  is a scale, exact values for which depend on the material. This distribution has a peak at  $\Delta_0$ , cuts off quickly below that, and has a very large long positive tail. The function is shown in Figure 1.3.



**Figure 1.3** The Landau distribution.

The Landau distribution has very unpleasant mathematical properties. Some of its integrals diverge, for example it has no variance (like the Cauchy distribution), and, worse than that, it does not even have a mean. The ensuing complications can be avoided on a case-by-case basis by imposing an upper limit on the energy loss, as a particle cannot lose more than 100% of its energy.

There is a function  $1/(\sqrt{2\pi})e^{-1/2\lambda + e^{-\lambda}}$  which is described in some places as ‘the Landau distribution’. It is not. It is an approximation to the Landau distribution [6], and not a very good one at that.

#### 1.3.4.4 The Negative Binomial Distribution

This considers the familiar binomial, but with a twist. As before, some process has a random probability  $p$  of success and  $q \equiv 1 - p$  of failure, and is repeated for many trials. But now instead of asking the probability of  $r$  successes from a fixed number of trials  $n$ , we ask for the probability of  $r$  successes before encountering a fixed number  $k$  of failures. This is given by

$$f(r; k, p) = \frac{(k+r-1)!}{r!(k-1)!} q^k p^r . \quad (1.25)$$

It is the probability for  $r$  successes and  $k-1$  failures in any permutation, followed by a final  $k$ th failure. The combinatorial factor can also be written  $(-1)^r \binom{-k}{r}$ , hence the name ‘negative binomial’. This can readily be extended to non-integer values by writing it as

$$f(r; k, p) = \frac{\Gamma(k+r)}{\Gamma(k)r!} q^k p^r , \quad (1.26)$$

although it is not clear what physical meaning this may have.  $\Gamma$  is the Gamma function, defined as

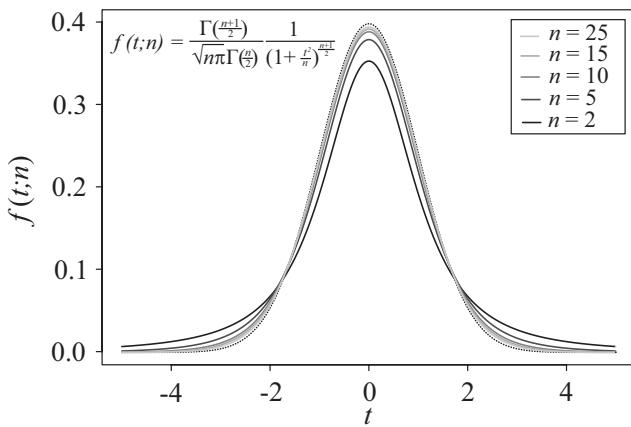
$$\Gamma(k) = \int_0^{+\infty} e^{-t} t^{k-1} dt . \quad (1.27)$$

The negative binomial distribution has a mean  $\mu = (p/q)k$  and a variance  $V = (p/q^2)k$ . The negative binomial approaches the Poisson as  $k$  becomes large and  $p$  small with constant  $p k \equiv \mu$ .

#### 1.3.4.5 Student's $t$ Distribution

If you take a sample of  $n$  values,  $\{x_1, \dots, x_n\}$ , from a Gaussian and histogram their differences from the true mean, divided by the standard deviation (a quantity often called the *pull distribution*), then this gives a unit Gaussian, that is a Gaussian with  $\mu = 0$ ,  $\sigma = 1$ , which can be a useful check that you have your errors right. If, as often happens, the true mean is unknown, then the spread about the measured mean is slightly smaller than 1, by a factor  $\sqrt{(n-1)/n}$ .

If the standard deviation  $\sigma$  is also unknown, then you can use instead the estimated  $\hat{\sigma} = \sqrt{(x-\mu)^2}$  if  $\mu$  is known or  $\hat{\sigma} = \sqrt{n/(n-1)(x-\bar{x})^2}$  if it is not. Now,



**Figure 1.4** Student's  $t$  distribution for  $n = 2, 5, 10, 15, 25$  with the Gaussian (dotted) for comparison.

for small  $n$  especially, this is not a very good estimator, and because you are dividing the differences from the mean by this bad estimate, the distribution for

$$t = \frac{x - \mu}{\hat{\sigma}} \quad (1.28)$$

is not given by a Gaussian, but by *Student's t distribution* for  $n - 1$  degrees of freedom, where Student's  $t$  distribution is given by

$$f(t; n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} . \quad (1.29)$$

This tends to a unit Gaussian as  $n$  becomes large, but for small  $n$  it has tails which are significantly wider (see Figure 1.4): large  $t$  values can result if  $\hat{\sigma}$  is an underestimate of the true value. The mean is clearly zero; the variance is not one, as it would be for a unit Gaussian, but  $n/(n - 2)$ .

### Example 1.3 Light yields in scintillators

You have five samples of scintillator from a manufacturer with light yields measured (in some units) as 1.23, 1.42, 1.35, 1.29 and 1.40. A second, cheaper, manufacturer provides a sample whose yield is 1.19. Does this give reason to believe that the cheaper sample has an inferior light yield?

The sample mean is 1.338 and the estimated standard deviation is 0.079, so the cheaper sample is 1.90 standard deviations below the mean. If this were a Gaussian distribution then the probability of a value lying this far below the mean is only 2.9% – so you would take this as strong evidence that the cheaper process was not so good. But for Student's  $t$  with four degrees of freedom the probability is (consulting the tables or evaluating a function) only 6.5%, so your evidence would be weaker (the calculations were done using the R function `pt(x,ndf)`).

### 1.3.4.6 The $\chi^2$ Distribution

In describing the agreement between a predictive function  $g(x)$  and a set of  $n$  measurements  $\{(x_i, \gamma_i)\}$ , it is useful to form the total squared deviation

$$\chi^2 = \sum_{i=1}^n \left[ \frac{\gamma_i - g(x_i)}{\sigma_i} \right]^2, \quad (1.30)$$

where  $\sigma_i$  is the Gaussian error on measurement  $i$ : if these errors are the same for all measurements then the factor can, of course, be taken outside the summation.

Each term will clearly contribute an amount of order one to the sum, and it is no surprise that  $E[f(\chi^2; n)] = n$ . The distribution is given by

$$f(\chi^2; n) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} \chi^{n-2} e^{-\chi^2/2}. \quad (1.31)$$

Some examples for different  $n$  are shown in Figure 1.5.

The  $\chi^2$  distribution is used a great deal in considering the question of whether a particular set of measurements (with their errors) and a particular model are compatible. This is addressed through the *cumulative  $\chi^2$  distribution*. For a given value of  $\chi^2$ , the complement of the cumulative distribution gives the *p-value*, the probability that, given that the model is indeed correct, a measurement would give a result with a  $\chi^2$  this large, or larger. If the value of  $\chi^2$  obtained is large compared to  $n$  then the *p-value* is small, that is the probability that a set of measurements truly described by this model would give such a large disagreement is small, and doubt is cast on the model, or the data (or both). The mean of  $f(\chi^2, n)$  is just  $n$ , and the standard deviation is  $\sqrt{2n}$ . For large  $n$  the distribution converges to the Gaussian, as it must by the central limit theorem. However, the convergence is actually rather slow, and this approximation is not often used. Instead the *p-value*

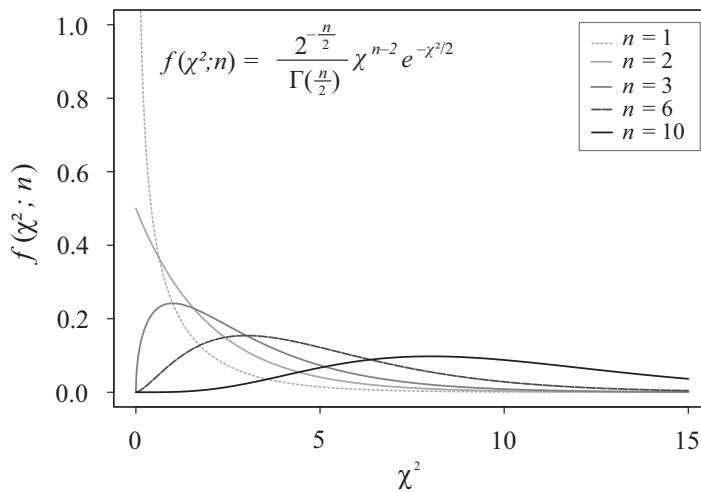


Figure 1.5  $\chi^2$  distributions for  $n = 1, 2, 3, 6$  and  $10$ .

should be obtained accurately from functions such as `TMath::Prob` in `root` or `pchisq` in `R`.

If the model has free parameters  $\theta$  which are not given, but were found by fitting the data, then the same  $\chi^2$  test can be used, but for  $n$  one takes the number of data points minus the number of fitted parameters. This is called the number of degrees of freedom. Strictly speaking this is only true if the model is a linear one (i.e. linear in the parameters). This is often the case, either exactly or to a good approximation, but there are some instances where this condition does not hold, leading to the computation of deceptively small and inaccurate  $p$ -values.

#### Example 1.4 Resistance measurements

A series of ten measurements are made of resistance  $R$  as a function of temperature  $T$ . The temperature is controlled very accurately, but the resistance is only measured with an accuracy of  $2 \Omega$ . A theoretical model predicts a value for  $R$  of  $(10.3 + 0.047 \cdot T) \Omega$ . The evaluation of  $\chi^2$  gives a value of 25.1. What can you say?

In this case one would use  $n = 10$ . Evaluating (using the `R` function `pchisq`) the probability of getting a value as large as 25.1 from  $n = 10$  gives 0.5%. It seems very implausible that the model really describes this data. (This does not necessarily mean the model is wrong. It could be that the data are badly measured. Or that the measurement accuracy has been estimated too optimistically.)

You will occasionally obtain  $\chi^2$  values that seem very small:  $\chi^2 \ll n$ . There is no standard procedure for rejecting these, but you should treat them with some suspicion and consider whether the model may have been formulated after the data had been measured ('retrospective prediction'), or whether perhaps the errors have been over-generously estimated.

##### 1.3.4.7 The Log-Normal Distribution

If the logarithm of the variable is given by a Gaussian distribution  $f(\ln x; \mu, \sigma)$  then the distribution for  $x$  itself is the *log-normal distribution*

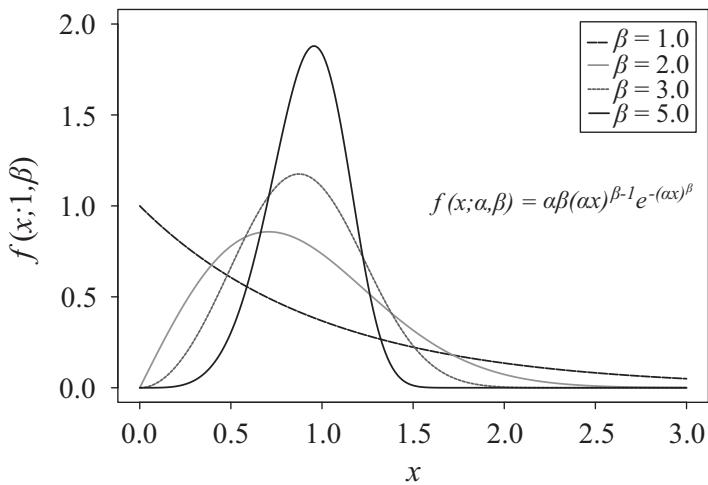
$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\left[\frac{(\ln x - \mu)^2}{2\sigma^2}\right]}. \quad (1.32)$$

Just as the central limit theorem dictates that any variable which is the sum of a large number of random components is described by a Gaussian distribution, any variable which is the product of a large number of random factors, none of which dominates the behaviour, is described by the log-normal. For instance, the signal registered by an electron in a calorimeter may be described by a log-normal distribution, as a certain fraction of the energy may be lost to dead material, a fraction to lost photons, a fraction to neutron production, and so on. The mean is given by  $e^{\mu + \sigma^2/2}$ , and the standard deviation is  $\sqrt{e^{\sigma^2} - 1}$ .

##### 1.3.4.8 The Weibull Distribution

The *Weibull distribution* is:

$$f(x; \alpha, \beta) = \alpha\beta(\alpha x)^{\beta-1} e^{-(\alpha x)^\beta}. \quad (1.33)$$



**Figure 1.6** Weibull distributions for  $\alpha = 1.0$ , progressively more peaked for  $\beta = 1.0, 2.0, 3.0$  and  $5.0$ .

This gives a shape which rises from zero to a peak and then falls back to zero again. It was originally invented to describe the failure rates in aging light bulbs. There are no failures at small times (because they are new and fresh) or at long times (because they have all failed). It is a rather more realistic modelling of real-life ‘lifetime’ than the simple exponential decay law for which the failure probability is constant.

The parameter  $\alpha$  is just a scale factor and  $\beta$  controls the shape. The case  $\beta = 1$  corresponds to the simple exponential decay law, whereas  $\beta > 1$  describes the behaviour when the failure probability increases with age, and gives successively sharper peaks. A case where the failure probability falls with time (perhaps because of initial burn-in) is described by  $\beta < 1$ . Examples are shown in Figure 1.6. The mean is  $\frac{1}{\alpha} \Gamma[1 + (1/\beta)]$  and the variance is  $1/\alpha^2 \left\{ \Gamma[1 + (2/\beta)] - \Gamma[1 + (1/\beta)]^2 \right\}$ . A location parameter  $x_0$  may also be needed in some problems, replacing  $x$  by  $x - x_0$ .

#### 1.4 Probability

We use probability every day, in both our work as physicists and our everyday lives. Sometimes this is a matter of precise calculation, when we buy an insurance policy or decide whether to publish a result, sometimes it is more intuitive, as when we decide to take an umbrella to work in the morning.

But although we are familiar with the concept of probability, on closer inspection it turns out that there are subtleties. When we get into technicalities there turn out to be different definitions of the concept which are not always compatible.

### 1.4.1

#### **Mathematical Definition of Probability**

Let  $A$  be an event. Then the probability  $P(A)$  is a number obeying three conditions, the *Kolmogorov axioms* [7]:

1.  $P(A) \geq 0$ ;
2.  $P(U) = 1$ , where  $U$  is the set of all  $A$ , the sample space;
3.  $P(A \cup B) = P(A) + P(B)$  for any  $A, B$  which are exclusive, that is  $A \cap B = \emptyset$ .

From these axioms a whole system of theorems and properties can be derived. However, the theory contains no statement as to what the numbers actually mean. For mathematicians this is, of course, not a problem, but it does not help us to apply the results.

### 1.4.2

#### **Classical Definition of Probability**

The probability of a coin landing heads or tails is clearly  $1/2$ . Symmetry dictates that it cannot be anything else. Likewise the chance of drawing a particular card from a pack has to be  $1/52$ . The original development of probability by Laplace, Pascal and their contemporaries, to aid the gambling fraternity, was founded on this equally likely construction. ‘Probability’ could be defined by taking fundamental symmetry where all cases were equally likely (say, the six sides on a dice), and extended to more complex cases (say, rolling two dice) by counting combinations.

Unfortunately this definition does not generalise to cases of continuous variables, where there is no fundamental symmetry: if you ‘draw a line at random’ from a given point, this could be done by taking coordinates of the endpoint from a uniform distribution, or by drawing an angle uniformly taken between  $0$  and  $360^\circ$ , the results are incomparably different. This approach thus leads to a dead end.

### 1.4.3

#### **Frequentist Definition of Probability**

Problems with the classical definition led to the alternative definition of probability as the limit of frequency by Venn, von Mises [8] and others. If a selection is made  $N$  times under identical circumstances, then the fraction of cases resulting in a particular outcome  $A$  tends to a limit, and this limit is what is meant by the probability:

$$P(A) = \lim_{N \rightarrow \infty} \frac{N(A)}{N}. \quad (1.34)$$

This is the generally adopted definition, taught in most elementary courses and textbooks. It satisfies, of course, the Kolmogorov axioms.

Where the classical definition is valid it leads to the same results. But there is an important philosophical difference. The probability  $P(A)$  is not some intrinsic property of  $A$ , it also depends on the way the sampling is done: on how the collective or ensemble of total possible outcomes has been constructed.

Thus, to use von Mises' example: the life insurance companies determine that the probability of one of their (male) clients dying between the ages of 40 and 41 is 1.1%. This is a hard and verifiable number, essential for the correct adjustment of the premium paid. However, it is not an intrinsic probability of the person concerned: you cannot say that a particular client has this number attached to them as a property in the same way that their height and weight are. The client belongs not just to this ensemble (insured 40-yr-old males) but to many others: 40-yr-old males, non-smoking 40-yr-old males, non-smoking professional lion tamers – and for each of these ensembles there will be a different number.

So there are cases with several possible ensembles, and the value of  $P(A)$  is ambiguous until the ensemble is specified. There are also cases where there is no ensemble, as the event is unique. The Big Bang is an obvious example, but others can be found much nearer home. For example, what is the probability  $P(\text{rain})$  that it will rain tomorrow? Now, there is only one tomorrow, and it will either rain or it will not, so  $P(\text{rain})$  is either 0 or 1. Von Mises condemns any further discussion as 'unscientific' use of language. This is further discussed (and resolved) in Section 1.5.2.

#### 1.4.4

##### **Bayesian Definition of Probability**

Another way of extending the unsatisfactory classical definition of probability was made by de Finetti [9] and others. De Finetti's starting point is the provocative 'Probability does not exist.' It has no objective status: it is something the human mind has constructed.

He shows that one can consistently define a personal probability (or *degree-of-belief*)  $P(A)$  in  $A$  by establishing the odds of a bet whereby you lose €1 if  $A$  subsequently turns out to be false, and you receive € $G$  if it turns out to be true. If  $P(A) > 1/(1 + G)$  you will accept the bet; if  $P(A) < 1/(1 + G)$  you will decline it.

Such personal probability is indeed something we use every day: when you decide whether or not to take an umbrella to work in the morning your decision is based on your personal probability of there being rain (and also the 'costs' involved in (a) getting wet and (b) having something extra to carry). However, there is no need for my personal probability to be the same as yours, or anyone else's. It is thus often referred to as a *subjective* probability. Subjective probability is also generally known as *Bayesian probability*, because of the great use it makes of Bayes' theorem [10]. This is a simple and fundamental result which is actually valid for any of the probability definitions being used.

#### 1.4.4.1 Bayes' Theorem

Suppose  $A$  and  $B$  are two events, and introduce the conditional probability  $P(A|B)$ , the probability of event  $A$  given that  $B$  is true (for instance: the probability that a card is the six of spades, given that it is black,  $P(\text{six of spades}|\text{black})$  is  $1/26$ ).

The probability of both  $A$  and  $B$  occurring,  $P(A \cap B)$  is clearly  $P(A|B)P(B)$ . But it is also  $P(B|A)P(A)$ . Equating these two quantities gives

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A). \quad (1.35)$$

This is used in problems like the famous ‘taxi colour’ example.

#### Example 1.5 Taxi colour

In some city, 15% of taxi cabs are yellow, and 85% are green. A taxi is involved in a hit-and-run accident, and an eyewitness says it was a yellow cab. The police have established that such eyewitness statements get the colour correct in 80% of cases and wrong in 20%. What is the probability that the cab was yellow?

The arithmetic is simple: just plug the numbers into Bayes’ theorem. Note that the  $P(B)$  term in the denominator can be helpfully written as  $P(B|A)P(A) + P(B|\bar{A})[1 - P(A)]$ , where  $\bar{A}$  denotes ‘not  $A$ '. If the cab’s true colour is denoted by  $Y$  or  $G$ , and the colour the witness says they saw by  $y$  or  $g$ , then

$$\begin{aligned} P(Y|y) &= \frac{P(y|Y)P(Y)}{P(y|Y)P(Y) + P(y|G)P(G)} P(Y) \\ &= \frac{0.8}{0.8 \cdot 0.15 + 0.2 \cdot 0.8} \cdot 0.15 = 0.4. \end{aligned}$$

So the cab was more likely (60% probability) to have been green – despite the witness saying exactly the opposite.

The ‘Bayesian’ use of Bayes’ theorem uses the same algebra but applies it to cases where  $B$  represents some experimental result and  $A$  some theory.  $P(B|A)$  is the probability of the result occurring if the theory is true, and  $P(A)$  is the personal probability you ascribe to the theory being true before the experiment is done – the *prior probability*.  $P(A|B)$  is the probability you ascribe to the theory in the light of the experiment – the *posterior probability* (the prior  $P(A)$  and the posterior  $P(A|B)$  are meaningless in the frequentist definition).

This all works neatly. If a result is forbidden in some theory,  $P(B|A) = 0$ , then its observation must lead to the theory being discarded. If a result is favoured in some theory, then observation of that result increases our degree-of-belief in that theory, although this increase is tempered by the probability of observing the result in any case.

## 1.5

### Inference and Measurement

Standard probability calculations are all about getting from the theory to the data. They address questions like: under such-and-such conditions, what is the probability that a specified random event will happen?

*Inference* is the reverse process: getting from the data to the theory. There is a theoretical model, containing some parameter (or parameters)  $\theta$ , which predicts the probability of getting a certain result (or set of results)  $x$ . What does the observation of a particular value of  $x$  tell you about  $\theta$ ? A simple example would be a particle of true energy  $E_{\text{true}} \equiv \theta$  giving a measured energy of  $E_{\text{meas}} \equiv x$  in the calorimeter. A less simple example would be the existence of a Higgs particle with mass  $m_H$  giving a set of events with particular characteristics in different channels.

#### 1.5.1

##### Likelihood

When you make a measurement, then  $f(x, \theta)$ , the probability of obtaining a result  $x$  given the value of a model parameter  $\theta$ , can also be written as the *likelihood*  $L(x; \theta)$ . This change is purely cosmetic: the actual algebra of the function is the same. Taking the Poisson as an example, and contemplating the observation of five events and a prediction of 3.4, one can write  $f(5; 3.4) \equiv L(5; 3.4) \equiv (3.4^5 / 5!) e^{-3.4}$ .

Given a result  $x$ , the value of  $L(x; \theta)$  tells you the probability that  $\theta$  would lead to  $x$ , which in turn tells you something about the plausibility that a particular value of  $\theta$  is the true one. The latter statement is purposefully made vague: it will be considered in proper detail later.

For practical purposes one often uses the logarithm of the likelihood, as, if you had a set of independent results  $x = \{x_1, \dots, x_n\}$ , then  $\ln L(x; \theta) = \sum_i \ln f(x_i; \theta)$ , and sums are easier to handle than products.

The *likelihood principle* states that if you have a result  $x$  then the likelihood function  $L(x; \theta)$  contains all the information relevant to your measurement of  $\theta$ . This principle is regarded by some as an irrefutable axiom, and by others as an irrelevance. Bayesian inference generally satisfies this, whereas frequentist inference generally violates it as the frequentist also has to consider the ensemble of experimental results that might have been obtained.

#### 1.5.2

##### Frequentist Inference

As von Mises points out, the probability of rain tomorrow is either 0 or 1, and no more can be said. However, you can construct an ensemble for something that looks very similar. Suppose that the pressure is falling and the clouds are gathering. A local weather forecast (perhaps made by a professional meteorologist, perhaps by the ache in your grandmother's left elbow) predicts rain. If you consider the track record of this particular prediction and count the number of times it has proved

correct, that gives a probability which is valid in the frequentist sense. So although you cannot say ‘It will probably rain tomorrow’, you can say ‘The statement “It will rain tomorrow.” is probably true.’

Indeed, if your weather prophet has been correct nine times out of ten, you can say ‘The statement “It will rain tomorrow.” has a 90% probability of being true.’ Again notice that the number is a property not just of the event (rain) but of the ensemble, in the form of the weather forecaster.

Now apply this approach to the interpretation of a measurement. Suppose your measurement process is known to give a result  $x$  which differs from the true value  $\mu$  with a probability distribution which is Gaussian with some known  $\sigma$ . You quote the result, whether it is the mass of the top quark determined from years of collider data, or a measurement of a resistance on a lab bench, as

$$x \pm \sigma. \quad (1.36)$$

This seems to say that  $\mu$  lies in the range  $[x - \sigma, x + \sigma]$  with 68% probability. But it can’t. The top mass,  $m_t$ , for which we currently quote  $173.2 \pm 0.9$  GeV, either lies in the range [172.3, 174.1] GeV or it does not. It is our measurement which is random, not the true value. So, as a frequentist, you make a statement about statements. ‘The statement “ $172.3 < m_t$  [GeV]  $< 174.1$ ” has a 68% chance of being true.’ Or, to put it another way, you make the statement ‘ $172.3 < m_t$  [GeV]  $< 174.1$ ’ with 68% confidence. There is a trade-off between the accuracy of the statement and the confidence you have in it. You could have played safer, and said with 95% confidence ‘ $171.4 < m_t$  [GeV]  $< 175.0$ ’. In other cases one-sided (upper or lower) limits may be appropriate.

### 1.5.3 Bayesian Inference

The Bayesian has no need of such mental gymnastics:

- $\pi(\theta)$  is the pdf describing my prior belief in the value of  $\theta$ .
- After a result  $x$ , Bayes’ theorem then gives my posterior belief  $f(\theta|x)$  as  $f(x|\theta)/f(x)\pi(\theta)$  where  $f(x) = \int f(x|\theta)\pi(\theta)d\theta$ .
- The denominator does not contain  $\theta$ , so we can write  $f(\theta|x) \propto f(x|\theta)\pi(\theta)$ .
- If you also decide that  $\pi(\theta)$  is uniform, just a constant, then  $f(\theta|x) \propto f(x|\theta)$ .
- The proportionality constant can be fixed by the normalisation.

In particular, the Bayesian interpretation of a Gaussian measurement, assuming a flat prior, equates the likelihood with the posterior probability:  $f(\mu|x) = 1/(\sigma\sqrt{2\pi})e^{-(x-\mu)^2/2\sigma^2}$ . This interpretation of the likelihood  $L(x;\mu)$  as a pdf in the parameter  $\mu$  looks especially plausible in the case of a Gaussian measurement: one has to remember that it is only valid for Bayesians and not for frequentists. Actually, the depiction of Bayesians and frequentists as different and rival schools of thought is not really correct. Yes, some statisticians can be fairly described as one or the other, but most of us adopt the approach most appropriate for a particular

problem. But care must be taken not to use concepts that are inapplicable in the framework chosen.

The uniform prior  $\pi(\theta) = \text{const}$  has the problem that, if the range of  $\theta$  is infinite, then to preserve  $\int \pi(\theta)d\theta = 1$ , the constant must vanish. However, one normally just works with priors which do not integrate to unity – so-called *improper priors*<sup>4)</sup> – relying on the final normalisation of the posterior.

Information from further measurements can be neatly incorporated into this framework. The posterior distribution from the first experiment is taken as the prior for the second, and its posterior forms the prior for the third (the order in which the combination is done is irrelevant).

#### 1.5.3.1 Use of Different Priors

The simplicity of a uniform prior is misleading. In the first place, it probably does not represent your true personal belief. Consider some hypothetical  $X$  particle predicted by some far-beyond-the-Standard-Model theory, and suppose you are convinced that it does exist. If you say  $\pi(m_X) = \text{const}$ , then that implies that your prior expectation that  $m_X$  lies between 1 and 2 GeV is the same as your prior expectation that it lies between 100 001 and 100 002 GeV, which frankly is not credible (given the choice, would you rather work on an experiment that could detect an  $X$  between 1 and 2 GeV, or between 100 001 and 100 002 GeV?).

Secondly, uniformity does not survive reparameterisation. If an angle has a uniform pdf in  $\theta$ , then the distribution in  $\cos \theta$  is very non-uniform, and in  $\sin \theta$  and  $\tan \theta$  it is different again. You cannot claim to have an objective analysis through having a uniform prior, as the choice of which variable to make uniform will affect the result.

Once the data have had a chance to constrain the results, then the effects of the choice of prior are reduced. Suppose we change from a prior uniform in  $m_X$  to one uniform in  $\ln m_X$  (this corresponds to a prior for  $m_X$  proportional to  $1/m_X$ ). If your measurements cover a range of  $m_X$  from 100 to 200 GeV, then this is a big change, but if it is only from 171 to 173 GeV then the difference is small – and we gloss over such differences in everyday statistics when we write expressions like  $\sigma_{\ln x} = \sigma_x/x$ .

So a sound measurement does not depend (much) on the choice of prior. This is called a *robust measurement*. In presenting a Bayesian result you may justify it by any of the following:

- showing that the result is robust: that the arbitrary choice of prior makes no great difference;
- justifying the prior in some way as being correct – or, perhaps, showing that the uniform prior is uniform in the correct variable;
- saying that you have chosen this prior and it represents your personal belief.

But just saying ‘We took a uniform prior.’ is not doing a proper job.

4) All possible jokes have already been made. Don’t go there, OK?

### 1.5.3.2 Jeffreys Priors

One attempt to systematise the choice of priors was made by Jeffreys [11]. His argument is based on the idea that an impartial prior should be ‘uninformative’ – it should not prefer any particular value or values.

Still speaking loosely: if the log-likelihood  $\ln L(x; \theta)$  has a nice sharp peak, then the data are telling you something about  $\theta$ , and if it is just a broad spread then it’s not being much help. The ‘peakiness’ of a distribution can be expressed using the second differential (with a helpful minus sign). On, or near, a sharp peak,  $-(\partial^2 \ln L) / (\partial \theta^2)$  will be large and positive.

Now we take a step back and forget any measurements made, and ask: given some value of  $\theta$ , what would we expect, on average, from a measurement? This quantity is called the *Fisher information*:

$$I(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right), \quad (1.37)$$

where the expectation value is evaluated by multiplying by  $f(x; \theta)$  and integrating over all possible results  $x$  for this particular value of  $\theta$ . A large value of  $I(\theta)$  means that if you make a measurement it will (probably) provide useful knowledge about the true value of  $\theta$ , and a small value of  $I(\theta)$  tells you that the measurement will not tell you much and is hardly worth doing. It can easily be shown (see e.g. Eq. (5.8) in [1]) that

$$I(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]. \quad (1.38)$$

Jeffreys answers the question ‘Should we use  $\theta$  or  $\ln \theta$  or  $\sqrt{\theta}$  as our fundamental variable?’ by saying that we should choose a form such that no particular value will yield more informative results than any other. He prescribes a parameterisation  $\theta'(\theta)$  for which  $-(\partial^2 \ln L) / (\partial \theta'^2)$  is constant. This is a variable in which all values are (from the Fisher information viewpoint) equal, and if we make the prior in this variable flat we are clearly being fair and even-handed.

In practice one does not have to find  $\theta'$  explicitly. If  $\pi(\theta')$  is the prior for  $\theta'$  and is constant, and as  $I(\theta')$  is constant by construction, then

$$\begin{aligned} \pi(\theta) &= \pi(\theta') \left| \frac{\partial \theta'}{\partial \theta} \right| \propto \sqrt{E\left[\left(\frac{\partial \ln L}{\partial \theta'}\right)^2\right]} \left| \frac{\partial \theta'}{\partial \theta} \right| \\ &= \sqrt{E\left[\left(\frac{\partial \ln L}{\partial \theta'} \frac{\partial \theta'}{\partial \theta}\right)^2\right]} \\ &= \sqrt{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]} \\ &= \sqrt{I(\theta)}. \end{aligned} \quad (1.39)$$

That means that for any  $\theta$ , one should take the prior as  $\pi(\theta) = \sqrt{I(\theta)}$ . For a location parameter, the *Jeffreys prior* is indeed just uniform in  $[-\infty, \infty]$  but for a scale parameter the prior is proportional to  $1/\theta$  – equivalently, the prior is uniform if you use  $\ln \theta$  as the fundamental form. For a Poisson mean it is  $1/\sqrt{\theta}$  where  $\theta = \nu$ .

Jeffreys' method rests on the idea that the prior should not prejudge the result: that it should be as 'uninformative' as possible. But it also provides a structure for giving a unique answer. Whether you choose the fundamental parameter to be  $\nu$  or  $\sqrt{\nu}$  does not change your final quoted result, thanks to the different priors you would have to use. This is why such priors are often termed 'objective', in that the dependence on your personal choice is removed.

Extension to more than one parameter is difficult but not impossible through a technique called *reference priors* [12].

Although the Jeffreys prior offers a way to getting unambiguous results, it has not been universally taken up. Partly because some are too lazy to consider anything other than a uniform prior in their favourite variable. Partly because of the difficulty of applying it to more than one parameter. Partly because it violates the likelihood principle. Partly because the prior does depend on the likelihood function and thus on the experimental technique, so you would invoke a different prior for (say) the Higgs mass, as determined with ATLAS through  $H \rightarrow \gamma\gamma$  than you would for the Higgs mass, as determined by CMS through  $H \rightarrow W^+W^-$ .

#### 1.5.3.3 The Correct Prior?

So what prior, or what collection of priors, should you use in a Bayesian analysis, if you are forced to do so? The answer, clearly, is: whatever happens to be your personal belief. But although a prior is subjective, it should not be arbitrary. Other data, measurements of this quantity or similar ones, can be used for guidance. Asking (theorist) colleagues can be useful, but if you do that be sure to ask for a wide range.

The 'quest for the correct prior' is an issue for physicists, who are conditioned to expect problems for which there is a unique correct answer, rather than statisticians, who know better. There is no unique correct prior for a problem. There is a range of sensible priors, and you should use these to check the robustness of your result. If it is stable, then the choice does not matter. If it is unstable, then the measurements cannot tell you anything honestly useful.

## 1.6

### Exercises

#### Exercise 1.1 Uniform distributions

Show, by integration, that the standard deviation of a uniform distribution of width  $w$  is  $w/\sqrt{12}$ .

**Exercise 1.2 Poisson distributions 1**

Show that the characteristic function of the Poisson distribution is  $\phi(u) = e^{\lambda(e^{iu}-1)}$ . Using the fact that the characteristic function of a convolution is the product of the individual characteristic functions, show that the convolution of two Poisson distributions of means  $\lambda_1$  and  $\lambda_2$  is also a Poisson, of mean  $\lambda_1 + \lambda_2$ . Now prove the result without using characteristic functions.

**Exercise 1.3 Poisson distributions 2**

A Poisson distribution has a mean of 3.7. Calculate ‘by hand’ the probability that it will give two events or less. Then calculate the same result using `ppois` in R or `poisson_cdf` in ROOT, or the equivalent in your favourite math program.

**Exercise 1.4 Bayes’ theorem**

If some object X exists, it may be found with equal probability in one of  $N$  locations. Show, using Bayes’ theorem, that if  $P$  is your prior belief in the existence of X, then after a location is unsuccessfully investigated your belief changes to  $P' = (N-1)/(N-P)P$ . If  $N = 10$  and there are nine unsuccessful searches, calculate and contrast the final posteriors for a prior of 0.9 and of 0.99.

**Exercise 1.5 *p*-values**

Find the number of standard deviations corresponding to *p*-values of 10%, 5% and 1% for a Gaussian distribution. Consider both one-sided and two-sided *p*-values.

**Exercise 1.6 Jeffreys prior**

Prove the result given for the Jeffreys prior of a Poisson distribution of mean  $\nu$ . Do this by writing down the log-likelihood, differentiating twice and negating, taking the expectation value and then taking the square root.

**References**

- 1 Barlow, R.J. (1989) *Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons.
- 2 R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.
- 3 Antcheva, I. et al. (2009) ROOT – A C++ framework for petabyte data stor-
- age, statistical analysis and visualization, *Comput. Phys. Commun.*, **180**, 2499.
- 4 von Bortkewitsch, L. (1898) Das Gesetz der kleinen Zahlen. *Monatsh. Math.*, **9**, 39.
- 5 Landau, L.D. (1944) On the energy loss of fast particles by ionization. *J. Phys. (USSR)*, **8**, 201.

- 6 Kolbig, K.S. and Schorr, B. (1984) A program package for the Landau distribution. *Comp. Phys. Commun.*, **31**, 97. Erratum: (2008) *Comp. Phys. Commun.*, **178**, 972.
- 7 Kolmogorov, A.N. (1950) *Foundations of the Theory of Probability*, Chelsea Publishing Company.
- 8 von Mises, R. (1957) *Probability, Statistics and Truth*, Dover Publications.
- 9 de Finetti, B. (1974) *Theory of Probability*, John Wiley & Sons.
- 10 Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.*, **53**, 370.
- 11 Jeffreys, H. (1966) *Theory of Probability*, Oxford University Press.
- 12 Berger, J.O., Bernardo, J.M., and Sun, D. (2009) The formal definition of reference priors. *Ann. Stat.*, **37**, 905.

## 2

### Parameter Estimation

*Olaf Behnke and Lorenzo Moneta*

#### 2.1

##### Parameter Estimation in High Energy Physics: Introductory Words

The estimation of parameters from observed distributions, a process also called *fitting*, is one of the fundamental data analysis tasks in experimental high energy physics; it is applied in almost every step of a measurement. In collider experiments, for instance, a typical goal is to identify new particles which are produced in hard collisions of two incident beam particles. The first task is to reconstruct each collision from the raw detector information. The trajectories and momenta of produced particles are determined by track fits to the detector hits associated with a particle. For obtaining good fit results the detectors have to be well calibrated, for example the subdetector positions and energy responses have to be known accurately. The calibration is often done by fitting the relevant detector parameters using large datasets of particles recorded during physics data-taking or in test-beam experiments. Finally, the spectra obtained from a large number of events are analysed. For instance, the mass, width and signal strength of a new particle can be inferred by fitting the invariant-mass distribution of the detected daughter particles. Typically, the spectra consist of both signal and background events, and both contributions are fitted simultaneously. The obtained results can be combined with those from other experiments. Technically this can also be achieved in a fitting procedure.

#### 2.2

##### Parameter Estimation: Definition and Properties

Parameter estimation consists of two basic ingredients: the estimation of the best approximation of the true parameter values ('best guess') – the so-called *point estimation* – and the estimation of the *uncertainties* of the estimated parameters, which are usually expressed as *confidence intervals*. The two most commonly used frequentist approaches to parameter estimation, namely the methods of *maximum*

*likelihood* and *least squares*, are discussed in detail in Sections 2.3 to 2.4. Bayesian methods are described in Section 2.6. The discussion of confidence intervals in this chapter is restricted to the case of fitting functions to data, and in particular estimating the standard deviations of the fitted parameters, while the general case is discussed in depth in Chapter 4.

Let us assume we have a dataset of  $N$  observations  $\mathbf{x} = (x_1, x_2, \dots, x_N)$ , where the measurements  $x_i$  are statistically independent and each follow a potentially unknown probability density  $f(x)$ . One can try to estimate the features of the function  $f(x)$ , such as its mean value or spread, or one may have a specific hypothesis for a functional form  $f(x; \boldsymbol{\theta})$  with unknown values of the parameter vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$  of dimension  $m$ . For example, the function  $f$  could be a straight line, in which case the parameters to be determined are its offset and slope.

An *estimator* is a function of the observed data  $\mathbf{x}$  which provides numerical values, the estimate  $\hat{\boldsymbol{\theta}}$ , for the parameter vector  $\boldsymbol{\theta}$ .

There are different ways to construct an estimator, and one usually tries to choose one that has the best properties. The most important *estimator properties* are

- *Consistency:* An estimator  $\hat{\boldsymbol{\theta}}$  is consistent (or asymptotically consistent) if it converges to the true value  $\boldsymbol{\theta}$  as the number of measurements  $N$  increases:

$$\lim_{N \rightarrow \infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}. \quad (2.1)$$

- *Bias:* The bias  $\mathbf{b}$  is defined as the difference between the expectation value of the estimator and the true parameter value  $\boldsymbol{\theta}$ :

$$\mathbf{b} = E[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}. \quad (2.2)$$

The expectation value is taken over a hypothetical infinite set of similar experiments. Unbiased estimators fulfil  $\mathbf{b} = 0$ . If the estimator  $\hat{\boldsymbol{\theta}}$  has a known bias  $\mathbf{b}$ , one can construct a new unbiased estimator  $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}} - \mathbf{b}$ .

- *Efficiency:* An estimator is efficient if its variance  $V[\hat{\boldsymbol{\theta}}]$  is small. For a parameter  $\boldsymbol{\theta}$  the minimum possible variance of an unbiased estimator is given, under rather general conditions, by the Rao–Cramér–Frechet *minimum-variance bound* (MVB):

$$V[\hat{\boldsymbol{\theta}}] \geq I(\boldsymbol{\theta})^{-1} \quad \text{with} \quad I_{jk}(\boldsymbol{\theta}) = -E \left[ \sum_{i=1}^N \frac{\partial^2 \ln f(x_i; \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right]. \quad (2.3)$$

The sum is over all data events, which are assumed to be independent and distributed according to the probability density function (pdf)  $f(x; \boldsymbol{\theta})$ . The allowed range of  $x$  must not depend on the parameters  $\boldsymbol{\theta}$ . The quantity  $I_{jk}(\boldsymbol{\theta})$  is called the *information matrix* and can also be expressed as

$$\begin{aligned} I_{jk}(\boldsymbol{\theta}) &= -N \int \frac{\partial^2 \ln f}{\partial \theta_j \partial \theta_k} f dx = N \int \frac{\partial \ln f}{\partial \theta_j} \frac{\partial \ln f}{\partial \theta_k} f dx \\ &= N \int \frac{1}{f} \frac{\partial f}{\partial \theta_j} \frac{\partial f}{\partial \theta_k} dx. \end{aligned} \quad (2.4)$$

Thus, any of these three integral representations can be used to determine the information matrix. For a one-dimensional parameter,  $I(\theta)$  is called the *information* (see (1.37) in Section 1.5.3.2). A proof of the MVB is given for example in [1] or [2].

### Example 2.1 Properties of estimator

A simple illustrative example of the properties for a one-parameter fit is the estimation of the mean lifetime  $\tau$  of a particle from  $N$  recorded decays at times  $t_1, t_2, \dots, t_N$ . Some properties of three simple estimators for this example are given in Table 2.1.

**Table 2.1** Three estimators of the mean lifetime of a particle and some of their properties.

Estimator	Consistent?	Unbiased?	Efficient?
$\hat{\tau} = \frac{t_1 + t_2 + \dots + t_N}{N}$	yes	yes	yes
$\hat{\tau} = \frac{t_1 + t_2 + \dots + t_N}{N - 1}$	yes	no	no
$\hat{\tau} = t_1$	no	yes	no

In general there is no such thing as an ‘ideal’ or ‘best’ estimator. In particular, there are many cases where the most efficient estimator is slightly biased. For certain estimators, the above properties are exactly known. However, in many cases the properties can only be evaluated by means of so-called *ensemble tests*, as discussed in Section 4.3.5 and in a broader scope in Section 10.5. These tests are often based on Monte Carlo simulations.

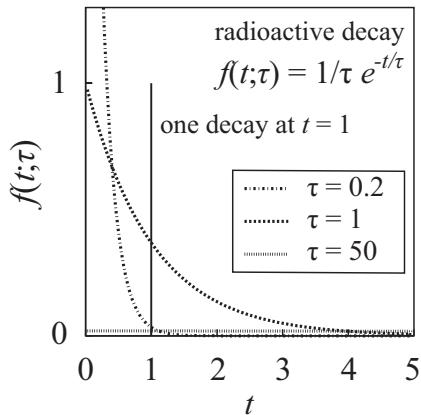
## 2.3

### The Method of Maximum Likelihood

Let us suppose we have a dataset of  $N$  measured quantities  $\mathbf{x} = (x_1, x_2, \dots, x_N)$ , where the measurements  $x_i$  are statistically independent and each follow the probability density  $f(x; \boldsymbol{\theta})$ . Here  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$  is a set of  $m$  parameters with unknown values to be estimated. The joint probability density function for the observed values  $\mathbf{x}$  is given by the likelihood function

$$L(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^N f(x_i; \boldsymbol{\theta}). \quad (2.5)$$

The maximum-likelihood estimate (MLE) of the parameters  $\boldsymbol{\theta}$  are the values  $\hat{\boldsymbol{\theta}}$  for which the likelihood function  $L(\mathbf{x}; \boldsymbol{\theta})$  has its *global maximum*. An intuitive reasoning for this approach is: if the assumed probability density function and its parameters are correct, we expect higher values of the likelihood function than for



**Figure 2.1** Likelihood curves  $L(t; \tau) = f(t; \tau)$  for three different values of the lifetime parameter  $\tau$ , and the observed decay at  $t = 1$ , indicated by a vertical line.

wrong parameter values. This is illustrated in Figure 2.1 for the case of radioactive decays following the exponential decay law  $f(t; \tau) = e^{-t/\tau}/\tau$ . One decay is observed at the time  $t = 1$ . The likelihood function  $L(t; \tau) = f(t; \tau)$  is shown for three different hypothetical lifetime parameter values  $\tau$  as a function of the observed time  $t$ . At the observed time  $t = 1$  the curve with  $\tau = 1$  has a much larger value than those with  $\tau = 50$  (the almost flat curve) and  $\tau = 0.2$  (the most steeply falling curve), and indeed  $\hat{\tau} = 1$  is the maximum-likelihood solution.

The maximum-likelihood method goes back to R.A. Fisher, see [3, 4].

### 2.3.1

#### Maximum-Likelihood Solution

The estimated values  $\hat{\boldsymbol{\theta}}$  of the parameters are obtained by finding the global maximum of the likelihood function. In practice, it is often more convenient to work with the logarithm of the likelihood function, called the *log-likelihood* and to search for the minimum of the negative log-likelihood function:<sup>1)</sup>

$$-\ln L(\mathbf{x}; \boldsymbol{\theta}) = -\sum_{i=1}^N \ln f(x_i; \boldsymbol{\theta}). \quad (2.6)$$

Unless the minimum occurs at the boundary of the allowed range of values for  $\boldsymbol{\theta}$ , a *necessary condition* for the minimum is that the negative log-likelihood satisfies

1) It is more convenient to work with the sum of events than with the product. Another reason is that one often deals with functions  $f(\mathbf{x}; \boldsymbol{\theta})$  that vary rapidly over many orders of magnitude with the fit parameter values, for example Gaussians, in which case the likelihood function may be (numerically) non-zero only in a small parameter region close to the estimator value, while the log-likelihood function can still be accurately calculated.

the following  $m$  equations:

$$-\frac{\partial \ln L(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} = 0 \quad \text{for } j = 1, \dots, m. \quad (2.7)$$

The likelihood function must be constructed using normalised probability density functions  $f(\mathbf{x}; \boldsymbol{\theta})$ :

$$\int f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = 1, \quad \text{so that} \quad \int L(\mathbf{x}; \boldsymbol{\theta}) dx_1 dx_2 \dots dx_m = 1. \quad (2.8)$$

In other words it is essential that the integral of the likelihood function does not depend on the parameters  $\boldsymbol{\theta}$ .

Apart from exceptional cases (see for example Sections 2.3.3.1, 2.4.1 and Example 2.2 below) the minimum of the negative log-likelihood function of  $m$  parameters cannot be found analytically as a function of the data  $\mathbf{x}$  and one has to use a numerical procedure. One example of such an algorithm is given in Section 2.4.2. In practice numerical minimisation is quite involved, and the reader is advised to use standard tools. In high energy physics it is very popular to use the MINUIT program [5] for parameter estimation, which comes with several built-in minimisation algorithms. A general discussion of numerical minimisation is beyond the scope of this book and the reader is referred to the expert literature such as [6].

### 2.3.2

#### Properties of the Maximum-Likelihood Estimator

In the asymptotic limit, that is when the number of measurements  $N$  goes to infinity, the maximum-likelihood estimator is consistent: as stated in (2.1), for each parameter  $\theta$  the estimate  $\hat{\theta}$  converges to its true value  $\theta$ . In this asymptotic limit, the MLE is unbiased and reaches the minimum-variance bound (2.3), meaning that no other estimator can be more efficient. For a finite number of events  $N$ , however, the MLE is in general a biased estimator, with a bias proportional to  $1/N$ .

Another important property of the MLE is the *invariance under parameter transformations*. If we apply a transformation,  $\psi = g(\theta)$ , the maximum-likelihood estimate  $\hat{\psi}$  is given as

$$\hat{\psi} = g(\hat{\theta}). \quad (2.9)$$

### 2.3.3

#### Maximum Likelihood and Bayesian Statistics

It should be stressed that the likelihood function is not a probability density function for the parameters  $\boldsymbol{\theta}$ . A ‘probability density function’ is defined only within Bayesian statistics (see Section 1.4.4) and corresponds to the posterior distribution  $p(\boldsymbol{\theta}; \mathbf{x})$  which involves the product of the likelihood function and the prior probability  $\pi(\boldsymbol{\theta})$  of the parameters (see Bayes’ theorem (1.35) in Section 1.4.4.1):

$$p(\boldsymbol{\theta}; \mathbf{x}) = \frac{L(\mathbf{x}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\int L(\mathbf{x}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \quad (2.10)$$

In Bayesian statistics, the MLE of the parameters coincides with the maximum of the posterior distribution, when the prior function in the parameter is uniform. The Bayesian parameter estimation is further elaborated in Section 2.6.

### 2.3.3.1 Averaging of Measurements with Gaussian Errors

A classical and illustrative application of the MLE is the averaging of  $N$  measurements of the quantity  $\theta$ . Let us assume that the individual measurements  $x_i$  are spread around the unknown true value  $\theta$  according to a Gaussian distribution with known width  $\sigma_i$ . Then, the probability density function is given by

$$f(x_i; \theta, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}. \quad (2.11)$$

This leads to the likelihood function

$$L(\theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}. \quad (2.12)$$

The log-likelihood function then becomes

$$\ln L(\theta) = -\frac{1}{2} \sum_{i=1}^N \frac{(x_i - \theta)^2}{\sigma_i^2} + \text{constant}. \quad (2.13)$$

This is the equation of a parabola. Since we add only negative terms, this parabola will have a maximum at some (yet to be found) value  $\hat{\theta}$ . Furthermore, it will have a constant second derivative  $-h = \partial^2 \ln L / \partial \theta^2$ . Thus, the log-likelihood function can be rewritten as

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{h}{2}(\theta - \hat{\theta})^2, \quad (2.14)$$

which leads to

$$L(\theta) \propto e^{-\frac{h}{2}(\theta - \hat{\theta})^2}. \quad (2.15)$$

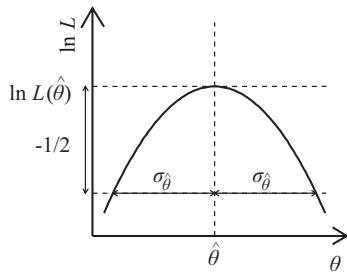
This means that for repeated experiments<sup>2)</sup> the estimated average  $\hat{\theta}$  will be spread around the true value  $\theta$  according to a Gaussian distribution with variance  $h^{-1}$ . The width of the Gaussian defines the standard deviation (uncertainty)  $\sigma_{\hat{\theta}}$ , which can be retrieved either from

$$\sigma_{\hat{\theta}} = (h)^{-1/2} = \left( -\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1/2} \quad (2.16)$$

or from the point where  $\ln L$  drops by 1/2 from its maximum value,

$$\ln L(\hat{\theta} \pm \sigma_{\hat{\theta}}) - \ln L(\hat{\theta}) = -1/2, \quad (2.17)$$

2) *Repeated experiment* means a new (independent) set of measurements, obtained under the same conditions.



**Figure 2.2** Exemplary  $\ln L$  curve for the averaging of several measurements with Gaussian uncertainties. The (exact) uncertainty  $\sigma_{\hat{\theta}}$  can be read off from the two points where  $\ln L$  drops by  $1/2$  from its maximum value at  $\hat{\theta}$ .

as illustrated in Figure 2.2. The solutions for  $\hat{\theta}$ ,  $h$  and  $\sigma_{\hat{\theta}}$  for the averaging problem can be calculated analytically by

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta} = 0 &= \sum_{i=1}^N \frac{(x_i - \theta)}{\sigma_i^2} = \sum_{i=1}^N \frac{x_i}{\sigma_i^2} - \theta \sum_{i=1}^N \frac{1}{\sigma_i^2} \\ \Rightarrow \quad \hat{\theta} &= \sum_{i=1}^N \left( \frac{x_i}{\sigma_i^2} \right) \Big/ \sum_{i=1}^N \left( \frac{1}{\sigma_i^2} \right) \end{aligned} \quad (2.18)$$

and by

$$h = -\frac{\partial^2 \ln L}{\partial \theta^2} = \sum_{i=1}^N \frac{1}{\sigma_i^2} \quad \Rightarrow \quad \sigma_{\hat{\theta}} = h^{-1/2} = \left( \sum_{i=1}^N \frac{1}{\sigma_i^2} \right)^{-1/2}. \quad (2.19)$$

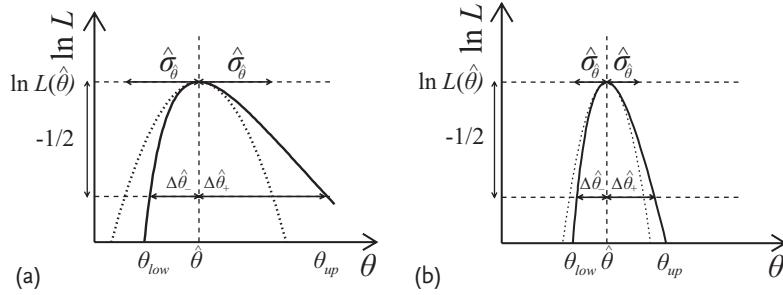
The results in (2.18) and (2.19) are the well-known formulæ for the weighted average.

Note that  $\sigma_{\hat{\theta}}$  obtained from (2.19) is an exact uncertainty, while for the most general MLE case one can only estimate uncertainties, as discussed in the next section (where correspondingly  $\hat{\sigma}_{\hat{\theta}}$  or other symbols are used to denote an estimated uncertainty). The result (2.19) for  $\sigma_{\hat{\theta}}$  can also be derived using the linear mapping from the measurements  $x_i$  to the estimator  $\hat{\theta}$ , as provided by (2.18), and applying simple error propagation. A related point that is easily shown is that the variance  $\sigma_{\hat{\theta}}^2$  reaches the *minimum-variance bound* (2.3). The uncertainty  $\sigma_{\hat{\theta}}$  is to be understood in the frequentist sense: for repeated experiments, the estimated intervals  $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$  (obtained for each experiment) will contain in 68% of all cases the true value  $\theta$ , as is evident from the symmetry between  $\theta$  and  $\hat{\theta}$  in the Gaussian likelihood (2.15).

### 2.3.4

#### Variance of the Maximum-Likelihood Estimator

In general, for small sample size the likelihood function can have any non-Gaussian shape. This is illustrated for a one-parameter case in Figure 2.3. Here,



**Figure 2.3** Exemplary  $\ln L$  curves for an exponential decay with two (a) and with eight (b) observed decays. The plots show the log-likelihood functions (solid curves) and parabolas (dotted curves) representing the local Gaussian approximation around the

maximum value. The estimated symmetric Gaussian errors  $\hat{\sigma}_{\hat{\theta}}$  and the ‘negative’ and ‘positive’ errors’  $\Delta\hat{\theta}_-$  and  $\Delta\hat{\theta}_+$  are also indicated. The latter are derived from  $\Delta \ln L = -0.5$  (2.24).

the log-likelihood function for the exponential decay, following the probability density  $f(x; \theta) = e^{-x/\theta}/\theta$ , is shown for a case with two (Figure 2.3a) and eight recorded events (Figure 2.3b). The exact log-likelihood functions, obtained from  $L(x; \theta) = \prod_{i=1}^N e^{-x_i/\theta}/\theta$ , are depicted as solid curves. The dashed lines show parabolas obtained from a Taylor expansion of the log-likelihood function around its maximum at  $\hat{\theta}$ . These parabolas represent *local Gaussian approximations* of the likelihood function. Obviously, for the case of only two events the obtained parabola provides a rather poor approximation of the exact function over the plotted range in which  $\ln L$  drops by about 0.5. However, for a similar drop, the Gaussian approximation works already much better for the case of eight events. This illustrates the convergence of the likelihood function to a Gaussian distribution with increasing number of events, which is a consequence of the *central limit theorem*.

It can be shown [2] that for any probability density  $f(x; \boldsymbol{\theta})$  and with increasing numbers of events, the likelihood function  $L$  approaches a multivariate Gaussian distribution<sup>3)</sup>,

$$L \propto e^{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}, \quad (2.20)$$

and the variance of the maximum-likelihood estimate reaches the maximum-variance bound (2.3):

$$\mathbf{V}[\hat{\boldsymbol{\theta}}] \rightarrow I(\boldsymbol{\theta})^{-1}, \quad \text{with} \quad I_{jk}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln L}{\partial \theta_j \partial \theta_k}\right) \equiv \mathbf{H}. \quad (2.21)$$

For finite statistics and assuming that the Gaussian approximation is already good enough, one can estimate the covariance matrix  $\mathbf{V}(\hat{\boldsymbol{\theta}})$  of the estimated parameter vector by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \left[ -\frac{\partial^2 \ln L(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^{-1} = \mathbf{H}^{-1}, \quad (2.22)$$

3) For the one-parameter case it is just a Gaussian distribution.

where the Hessian matrix  $\mathbf{H}$  of second derivatives of the negative log-likelihood function is to be determined at the estimator value. The estimated standard deviation of each single parameter  $\hat{\theta}_j$  is given by

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{\hat{V}_{jj}(\hat{\theta})}. \quad (2.23)$$

Another way to estimate uncertainties, which is numerically equivalent for the Gaussian case (see Section 2.3.3.1) is to use the contour given by values  $\boldsymbol{\theta}'$  such that

$$\Delta \ln L \equiv \ln L(\boldsymbol{\theta}') - \ln L_{\max} = -1/2, \quad (2.24)$$

where the extreme limits on this contour on the  $\theta_j$  axis define an approximate  $1\sigma$  (i.e. 68%)<sup>4)</sup> confidence interval for  $\theta_j$ . Similarly,  $s \cdot \sigma$  confidence intervals can be obtained from the

$$\Delta \ln L = -s^2/2 \quad (2.25)$$

contour.

The two methods to determine  $1\sigma$  uncertainties, using (2.23) or (2.24), are illustrated in Figure 2.3. For the latter one has to find the two points  $\theta_{\text{low}} = \hat{\theta} - \Delta\hat{\theta}_-$  and  $\theta_{\text{up}} = \hat{\theta} + \Delta\hat{\theta}_+$  for which  $\Delta \ln L = -0.5$ . These define a 68% confidence interval  $[\theta_{\text{low}}, \theta_{\text{up}}]$  for  $\theta$ , which is in general asymmetric around  $\hat{\theta}$ . The values  $\Delta\hat{\theta}_-$  and  $\Delta\hat{\theta}_+$  are often referred to as the *negative* and *positive uncertainties*. As one can see in Figure 2.3, they can be rather different from the symmetrical uncertainty  $\hat{\sigma}_{\hat{\theta}}$  based on the local Gaussian approximation formula (2.23).

The usual notation which we follow in this chapter is to quote measurement values and their uncertainties as

$$\theta = \hat{\theta}_{-\Delta\hat{\theta}_-}^{+\Delta\hat{\theta}_+} \quad (2.26)$$

when using the  $\Delta \ln L = -0.5$  method to obtain uncertainties, and

$$\theta = \hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}, \quad (2.27)$$

when using the local Gaussian approximation. If the Gaussian approximation of the likelihood function is not yet good in the region where  $\ln L$  drops by about 1/2, the 68% confidence interval  $[\hat{\theta} - \hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \hat{\sigma}_{\hat{\theta}}]$  can become very inaccurate. However, the interval based on  $\Delta \ln L = -0.5$  has a better *coverage* (see Chapter 4), thanks to the invariance of the likelihood function for any monotonic transformation of the parameter  $\theta \rightarrow \theta'$ . In the case of a non-Gaussian likelihood, it is always possible to find a transformation  $\theta \rightarrow \theta'$ , which makes  $L(\theta')$  Gaussian. Thus, the values of  $\theta'$  for which  $\ln L$  drops by 1/2 define the  $1\sigma$  confidence region for  $\theta'$ , and the corresponding values for  $\theta$  define the  $1\sigma$  region for this parameter. Thus, in practice one can forget about  $\theta'$  and determine the uncertainties directly in the parameter  $\theta$ .

4) More accurately, a  $1\sigma$  interval has a 68.27% confidence level and a  $2\sigma$  interval has a 95.45% confidence level which we abbreviate as 68% and 95%, respectively, throughout the book.

**Example 2.2 Exponential decay**

The maximum-likelihood fit of the exponential (radioactive) decay was already introduced above – here it is discussed in a complete form. The probability density function to observe a decay at time  $t$  is given by

$$f(t; \tau) = \frac{1}{\tau} e^{-t/\tau}. \quad (2.28)$$

The parameter to be estimated is  $\tau$ , the lifetime parameter of the decay. With  $N$  measurements  $t_i$ , the maximum-likelihood fit consists in finding the maximum of

$$\ln L = \sum_{i=1}^N \ln f(t_i; \tau) = \sum_{i=1}^N \left( -\ln \tau - \frac{t_i}{\tau} \right). \quad (2.29)$$

This is achieved by finding the  $\tau$  value for which

$$\frac{\partial \ln L}{\partial \tau} = -\frac{N}{\tau} + \sum_{i=1}^N \frac{t_i}{\tau^2} = 0, \quad (2.30)$$

leading to the result

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N t_i. \quad (2.31)$$

In this particular case the maximum-likelihood estimator coincides with the sample mean. It can also be shown that  $\hat{\tau}$  is an unbiased estimator of  $\tau$ . The estimated uncertainty of  $\hat{\tau}$ , based on the local Gaussian approximation of the likelihood function (see (2.23)), is

$$\begin{aligned} \hat{\sigma}_{\hat{\tau}} &= \sqrt{\hat{V}(\hat{\tau})} = \left( -\frac{\partial^2 \ln L}{\partial \tau^2} \Big|_{\tau=\hat{\tau}} \right)^{-1/2} \\ &= \left( \frac{2}{\hat{\tau}^3} \sum_{i=1}^N t_i - \frac{N}{\hat{\tau}^2} \right)^{-1/2} = \frac{\hat{\tau}}{\sqrt{N}}. \end{aligned} \quad (2.32)$$

The following table lists the uncertainty values  $\hat{\sigma}_{\hat{\tau}}$  for exemplary cases with two and eight observed decays for a hypothetical value  $\hat{\tau} = 1$ .

$N$	$\hat{\sigma}_{\hat{\tau}}$	$\Delta\hat{\tau}_-$	$\Delta\hat{\tau}_+$
2	0.70	0.47	1.30
8	0.35	0.32	0.47

Also listed are the negative and positive uncertainties  $\Delta\hat{\tau}_-$  and  $\Delta\hat{\tau}_+$ , determined from the two points where  $\ln L$  drops by  $1/2$  from its maximum. As discussed above, they can be used to define a more reliable 68% confidence level (CL) interval

$[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$  for  $\tau$  than the one based on  $\sigma_{\hat{\tau}}$ . The different uncertainties are also compared in Figure 2.3, where the lifetime parameter was generically denoted as  $\theta$  instead of  $\tau$ . While the negative and positive uncertainties are very different from  $\hat{\sigma}_{\hat{\tau}}$  when only two events are observed, there is already a better agreement when measuring eight events.

#### 2.3.4.1 Confidence Region Evaluation with $\chi^2$ Function Quantiles

For the evaluation of the confidence regions defined by (2.25), it is practical to use the likelihood ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\mathbf{x}; \boldsymbol{\theta})}{L(\mathbf{x}; \hat{\boldsymbol{\theta}})} . \quad (2.33)$$

In the large-sample limit, where the likelihood approaches a Gaussian,  $-2 \ln \lambda(\boldsymbol{\theta})$  follows a  $\chi^2$  distribution with  $m$  degrees of freedom (Wilks' theorem). One can then use the quantiles  $\chi^2_{1-\alpha}$  of the  $\chi^2$  distribution to evaluate  $1 - \alpha$  confidence regions. These quantiles define the rise in  $-2 \ln \lambda(\boldsymbol{\theta})$  corresponding to the points of  $\boldsymbol{\theta}$  on the border of the confidence region. The value of the quantile is obtained from  $F_{\chi^2}^{-1}(1 - \alpha, m)$ , the inverse of the cumulative function  $F_{\chi^2}$  of a  $\chi^2$  distribution with  $m$  degrees of freedom. It is implemented in the ROOT framework [7] as the function `ROOT::Math::chisquared_quantile(p,ndf)`. For example, in the case of one parameter, the  $1\sigma$  interval (i.e.  $1 - \alpha = 68\%$ ) is obtained as

$$\begin{aligned} -\ln \lambda(\theta_{\text{low}} \equiv \hat{\theta} - \Delta\hat{\theta}_-) &= -\ln \lambda(\theta_{\text{up}} \equiv \hat{\theta} + \Delta\hat{\theta}_+) \\ &= \frac{1}{2} F_{\chi^2}^{-1}(0.68, 1) = 0.5 . \end{aligned} \quad (2.34)$$

In the case of two parameters, one can obtain contours in  $\theta_1$  and  $\theta_2$  enclosing a two-dimensional  $(\theta_1, \theta_2)$  confidence region at a given confidence level by finding the corresponding values in the  $-\ln \lambda(\theta_1, \theta_2)$  function. The 68% contour is obtained by finding the set of values for  $\theta_1$  and  $\theta_2$  which have

$$-\ln \lambda(\theta_1, \theta_2) = \frac{1}{2} F_{\chi^2}^{-1}(0.68, 2) = 1.15 . \quad (2.35)$$

Multi-dimensional intervals based on the variance estimated using the Hessian matrix can also be obtained but, as in the one-parameter case, this approximation will produce only symmetric intervals around  $\hat{\boldsymbol{\theta}}$  which might be less accurate.

#### 2.3.4.2 Profile Likelihood

In the case of a likelihood function depending on many parameters, but where one is interested in only one parameter  $\mu$  and its uncertainty, one can use a *profile likelihood ratio* defined as

$$\lambda(\mu) = \frac{L(\mathbf{x}; \mu, \hat{\boldsymbol{\theta}})}{L(\mathbf{x}; \hat{\mu}, \hat{\boldsymbol{\theta}})} . \quad (2.36)$$

In the numerator, the parameters  $\boldsymbol{\theta}$  are fitted to their MLE,  $\hat{\boldsymbol{\theta}}$ , for a given value of the parameter  $\mu$ . In the denominator,  $\mu$  is also estimated – the values  $\hat{\mu}$  and  $\hat{\theta}$  define the global maximum of the likelihood  $L$ . What has been described before still applies: the asymptotic distribution of  $-2 \ln \lambda(\mu)$  follows a  $\chi^2$  distribution, and the intervals in  $\mu$  are obtained using (2.25). However, it is now more complicated to evaluate the likelihood ratio, since each evaluation at a given value of  $\mu$  requires a minimisation in all the other parameters  $\boldsymbol{\theta}$ . This method of profiling the likelihood is very popular for estimating uncertainties from a maximum-likelihood fit; in high energy physics it is known as the *Minos* method of the MINUIT program [5]. For finding the confidence region in more than one parameter, it is still possible to construct a multi-dimensional likelihood ratio,  $\ln \lambda(\boldsymbol{\theta})$ , where all the parameters one is not interested in are profiled. The asymptotic distribution of  $-2 \ln \lambda$  is in this case a  $\chi^2(m)$ , where  $m$  is the number of parameters of interest. The confidence intervals obtained with the profile likelihood method are usually reliable, although in most cases exact coverage is only reached in the asymptotic limit of large sample size (see also the corresponding discussion in Section 4.3.4)

### 2.3.5

#### Minimum-Variance Bound and Experiment Design

As already mentioned, with increasing number  $N$  of events, the MLE asymptotically reaches the theoretically minimum-variance bound (MVB) which is given by the inverse of the information (see (2.3)). The MVB scales with  $1/N$ , which can be exploited to determine how many events an experiment needs to record until a desired statistical precision can be reached. The calculation can be done before the experiment is carried out and thus can be very useful for the experimental design; it only requires knowledge of the relevant pdf. The following example is taken from [8].

#### Example 2.3 Information evaluation for weak decay of a muon

For the weak decay of a muon  $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$ , the probability density of a positron being emitted at an angle  $\alpha$  to the muon spin direction is given by

$$f(x; \theta) = \frac{1 + \theta x}{2}, \quad \text{with } x = \cos \alpha.$$

The goal is to determine the polarisation parameter  $\theta$ . With  $\partial f / \partial \theta = x/2$  the information (see (2.3)) for one event is found to be

$$\begin{aligned} I(\theta) &= \int_{-1}^1 \frac{1}{f} \left( \frac{\partial f}{\partial \theta} \right)^2 dx \\ &= \int_{-1}^1 \frac{x^2}{2(1 + \theta x)} dx = \frac{1}{2\theta^3} \left[ \ln \left( \frac{1 + \theta}{1 - \theta} \right) - 2\theta \right]. \end{aligned}$$

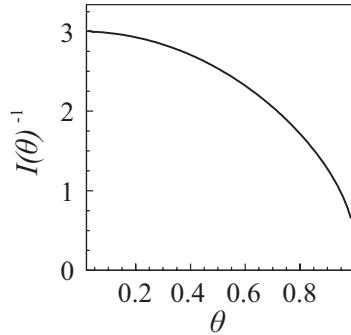
Hence, the minimum-variance bound for  $N$  events is

$$V(\hat{\theta}) \geq I(\theta)^{-1} = \frac{1}{N} \frac{2\theta^3}{\ln\left(\frac{1+\theta}{1-\theta}\right) - 2\theta}. \quad (2.37)$$

Assuming a true value of  $\theta = 1/3$  we obtain

$$V(\hat{\theta}) \geq \frac{2.8}{N}.$$

Thus, if we want to obtain a value with 1% relative statistical precision corresponding to  $V[\hat{\theta}] = 1/300^2$  and assuming that the MVB is reached, at least  $N = 2.52 \cdot 10^5$  events have to be recorded. Of course, in general, also systematic uncertainties have to be taken into account. If these can also be estimated beforehand, they can be added in quadrature to the statistical uncertainty to obtain the estimated total uncertainty.



**Figure 2.4** Inverse information  $I(\theta)^{-1}$  for the example of the weak decay of muons, as a function of the polarisation parameter  $\theta$ , for one recorded event.

In order to evaluate the information, one must have some idea of the value of the true parameter, for example from a theory prediction or from previous measurements. If neither is the case, one can plot the information as a function of the true parameter, in order to get an idea of the sensitivity of the experiment for different parameter values. For our example, Figure 2.4 shows the variance lower limit  $I(\theta)^{-1}$  (2.37) as a function of  $\theta$ . The number of events is chosen as  $N = 1$ . The variance has its largest value at  $\theta = 0$  and then decreases with increasing  $\theta$ .

Measurements in high energy physics experiments are often very complex, and the probability density function cannot be calculated analytically. In this situation, one useful alternative is to employ so-called *ensemble tests* which are explained in detail in Sections 4.3.5 and 10.5.

These tests are based on ensembles of simulated events and allow uncertainties to be estimated for almost every experimental situation. In the case that the proba-

bility density is known, ensemble tests can be used to determine also how far (if at all) the MLE is from reaching the MVB, for any number of events  $N$ .

## 2.4

### The Method of Least Squares

A very popular method is that of least squares. Since its introduction by Gauss in 1795 it has been the most widely used tool to fit model parameters to experimental observations. In its simplest form it is used if one has two variables  $x$  and  $y$  and a set of values  $x_i$  (with  $i = 1, \dots, N$ ) with no uncertainties plus a corresponding set of measurements  $y_i$  with uncertainties  $\sigma_i$ . The relation between  $x$  and  $y$  is given by the function  $f(x; \boldsymbol{\theta})$  whose form is known. The function  $f$  depends on the values of the unknown parameter vector  $\boldsymbol{\theta}$  (with dimension  $m$ ) that one wants to estimate. An example is the determination of the trajectory of a particle, where one measures the vertical position  $y_i$  in detectors at known horizontal positions  $x_i$  in order to determine the particle's flight direction. The variable  $y$  could also be a vector, for example the three-dimensional position of a particle measured at various times (so here  $x$  would be the time).

The least-squares ansatz for determining the parameter values is to minimise the sum  $\chi^2$  of the squares of the residuals (see also (1.30) in Section 1.3.4.6):

$$\chi^2 = \sum_{i=1}^N \left[ \frac{y_i - f(x_i; \boldsymbol{\theta})}{\sigma_i} \right]^2. \quad (2.38)$$

The estimated parameters  $\hat{\boldsymbol{\theta}}$  are then the values which minimise  $\chi^2$ . This is equivalent to solving the system of  $m$  equations:

$$\frac{\partial \chi^2}{\partial \theta_j} = 0, \quad (2.39)$$

which, using (2.38), results in:

$$-2 \sum_{i=1}^N \frac{1}{\sigma_i^2} \frac{\partial f(x_i; \boldsymbol{\theta})}{\partial \theta_j} [y_i - f(x_i; \boldsymbol{\theta})] = 0. \quad (2.40)$$

When the uncertainty  $\sigma_i$  is the same for all the measurements  $y_i$ , it can be taken out of the sum as a common scale factor which can be removed for the determination of  $\hat{\boldsymbol{\theta}}$  (but not for the determination of its variance). It can be shown (see for example Section 7.2.4 in [2]) that one of the roots of the above equations (2.39) is a consistent estimator of  $\boldsymbol{\theta}$ .

In the most general case the measurements are correlated and the generalised least-squares ansatz is given by

$$\chi^2 = [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})]^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})] \quad (2.41)$$

with  $\gamma = [y_1, y_2, \dots, y_N]$ ,  $f(\boldsymbol{\theta}) = (f(x_1; \boldsymbol{\theta}), f(x_2; \boldsymbol{\theta}), \dots, f(x_N; \boldsymbol{\theta}))$  and  $\mathbf{V}$  denoting the covariance matrix of  $\gamma$ . The least-squares solution is again found by solving the system of  $m$  equations (2.39).

The least-squares ansatz coincides with the maximum-likelihood method for the case of measurements with known Gaussian uncertainties. This can be seen immediately, for example for the case of uncorrelated measurements from

$$L = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{[y_i - f(x_i; \boldsymbol{\theta})]^2}{2\sigma_i^2}} = c e^{-\frac{1}{2} \sum_{i=1}^N \frac{[y_i - f(x_i; \boldsymbol{\theta})]^2}{\sigma_i^2}} = c e^{-\frac{\chi^2}{2}} \quad (2.42)$$

with

$$c = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i}. \quad (2.43)$$

We hence obtain the relation

$$\chi^2 = -2 \ln L + 2 \ln c. \quad (2.44)$$

In case the likelihood is a multivariate Gaussian function, that is for correlated measurements, an equivalent relation holds. The estimated parameters will obviously be exactly the same when maximising  $\ln L$  or minimising  $\chi^2$ . From the factor  $-2$  in (2.44) and from (2.22) it follows that the variance of  $\hat{\boldsymbol{\theta}}$  can be estimated from

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \frac{1}{2} \left[ \frac{\partial^2 \chi^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^{-1} = \mathbf{H}^{-1}. \quad (2.45)$$

The uncertainties of the estimated parameters can also be estimated (see (2.24)) from the contour given by the vector  $\boldsymbol{\theta}'$  such that

$$\chi^2(\boldsymbol{\theta}') = \chi^2_{\min} + 1. \quad (2.46)$$

The extreme limits of this contour on the  $\theta_j$  axis define an approximate  $1\sigma$  confidence interval for  $\theta_j$ . Similarly,  $s \cdot \sigma$  confidence intervals can be derived from the

$$\chi^2(\boldsymbol{\theta}') = \chi^2_{\min} + s^2 \quad (2.47)$$

contour (see also (2.25)).

In general, following the discussion of the MLE variance in Section 2.3.4, the confidence intervals based on the  $\chi^2_{\min} + 1$  variation are more reliable than those based on the Hessian matrix  $\mathbf{H}$ . In fact in high energy physics, the  $\chi^2_{\min} + 1$  method is the most frequently used one because, for an arbitrary likelihood function, the method gives the proper uncertainty estimate when using (as is often done) a generalised  $\tilde{\chi}^2 \equiv -2 \ln L$ . If the measurement resolutions  $\sigma_i$  are either not known or vary with the other parameters  $\boldsymbol{\theta}$ , then the standard  $\chi^2$  differs from  $-2 \ln L$  by the non-constant term  $2 \ln c$  (see (2.43) and (2.44)). This might lead to a bias in the  $\chi^2$ -minimisation result and it could be better to use the generalised  $\tilde{\chi}^2 = -2 \ln L$ .

## 2.4.1

**Linear Least-Squares Method**

In the case where  $f(x; \boldsymbol{\theta})$  is a linear function of the parameters  $\boldsymbol{\theta}$ ,

$$f(x; \boldsymbol{\theta}) = \sum_{j=1}^N a_j(x) \theta_j , \quad (2.48)$$

the  $\chi^2$  formula (2.41) can be written as

$$\chi^2 = (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) . \quad (2.49)$$

Here,  $\mathbf{A}$  is the so-called *design matrix*

$$\mathbf{A}: A_{i,j} = a_j(x_i) . \quad (2.50)$$

For the  $\chi^2$  minimisation one needs to solve the so-called *normal equations*

$$\frac{\partial \chi^2}{\partial \boldsymbol{\theta}} = -2 (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{y} - \mathbf{A}^T \mathbf{V}^{-1} \mathbf{A} \boldsymbol{\theta}) = 0 . \quad (2.51)$$

By inverting the matrix  $\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A}$ , the analytical solution is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}^{-1} \mathbf{y} . \quad (2.52)$$

It follows that  $\hat{\boldsymbol{\theta}}$  is a linear estimator,  $\hat{\boldsymbol{\theta}} = \mathbf{L}\mathbf{y}$ , with  $\mathbf{L} = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}^{-1}$ . Since we have a linear relation  $\hat{\boldsymbol{\theta}} = \mathbf{L}\mathbf{y}$ , the variance of the estimator can be directly obtained using *error propagation* from the covariance matrix  $\mathbf{V}$  of the data  $\mathbf{y}$ :

$$\begin{aligned} \mathbf{V}(\hat{\boldsymbol{\theta}}) &= \mathbf{L} \mathbf{V} \mathbf{L}^T = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}^{-1} \mathbf{V} \mathbf{V}^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \\ &= (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} = \mathbf{H}^{-1} \end{aligned} \quad (2.53)$$

with the Hessian matrix

$$\mathbf{H} = \frac{1}{2} \frac{\partial^2 \chi^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \mathbf{A}^T \mathbf{V}^{-1} \mathbf{A} . \quad (2.54)$$

The covariance matrix  $\mathbf{V}(\hat{\boldsymbol{\theta}})$  is exact and not only an estimate, as it is for the general least-squares case (2.45). However, the assumption must hold that the measurements  $\mathbf{y}$  have a Gaussian variance  $\mathbf{V}$ .

A special property can be derived for the case that the measurements have a covariance matrix of the form  $\mathbf{V} = \sigma^2 \cdot \mathbf{I}$ , where  $\mathbf{I}$  denotes an  $m$ -dimensional unit matrix; in other words the measurements are uncorrelated and have the same variance  $\sigma^2$ . Then, according to the *Gauss–Markov theorem*, the linear least square estimator (2.52) is a ‘best linear unbiased estimator’ (BLUE), that is it is unbiased and has the smallest possible variance. A proof of the theorem can be found for example in [2].

**Examples of linear least-squares fits** There is a huge variety of linear least-squares fit problems. Examples are

- the constant function  $y = \theta$ ,
- the straight line  $y = \theta_0 + \theta_1 x$ ,
- the parabola  $y = \theta_0 + \theta_1 x + \theta_2 x^2$ ,
- polynomials of any order (including the above examples),
- and functions like  $y = \theta \sin x$  or  $y = \theta e^{-x}$ , where  $\theta$  represents a normalisation factor.

Linear least-squares fits are often confused with the straight-line fit. However, the linearity refers solely to the fit parameters  $\theta$ , while the  $y(x)$  dependence can be anything, linear or non-linear.

#### 2.4.1.1 Averaging of Measurements with Gaussian Errors

We now return to the example of averaging several measurements with Gaussian errors (see Section 2.3.3.1). In the least-squares fit approach, the problem corresponds to the fitting of a constant function  $y = \theta$ , the most simple linear fit problem. The  $\chi^2$  for this problem is given by

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - \theta)^2}{\sigma_i^2}. \quad (2.55)$$

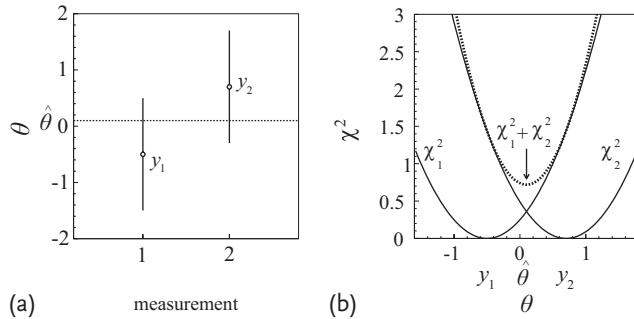
The topic is discussed here again, since it is very important and also illustrative for  $\chi^2$  fits. A very simple case – which is depicted in Figure 2.5 – is the averaging of just two measurements. Figure 2.5a shows the two data points with their uncertainties and the fitted average value (dashed line). Figure 2.5b shows a graphical representation of the  $\chi^2$  function versus the parameter  $\theta$ . It demonstrates that one can perform the  $\chi^2$  minimisation using a very simple *graphical method*. One first draws the individual  $\chi^2$  contributions  $\chi_i^2 = (y_i - \theta)^2 / \sigma_i^2$  from the two data points  $i = 1, 2$  (full curves) and then adds their values to obtain the total  $\chi^2$  function (dashed curve). Then one reads off the value of  $\hat{\theta}$  at the minimum of this curve. The standard deviation  $\sigma_{\hat{\theta}}$  of  $\hat{\theta}$  can also be read off as the horizontal distance between this point and the points  $\theta'$  where  $\chi^2(\theta') = \chi^2_{\min} + 1$ . Obviously, the parabola of the total  $\chi^2$  function is steeper than the two parabolas of the input measurements, indicating that the uncertainty on  $\hat{\theta}$  is reduced.

A simple calculation,<sup>5)</sup> using (2.18) and (2.19), shows that the  $\chi^2$  function can be reformulated as

$$\chi^2 = \frac{(y_1 - y_2)^2}{\sigma_1^2 + \sigma_2^2} + \frac{(\theta - \hat{\theta})^2}{\sigma_{\hat{\theta}}^2}. \quad (2.56)$$

Thus, the original Gaussian probability densities of the two measurements around the true value  $\theta$  have been transformed into two other Gaussian densities repre-

5) The calculation is left to the reader as an exercise at the end of the chapter.



**Figure 2.5** Weighted averaging of two measurements. (a) The two data points  $y_1$  and  $y_2$  with error bars indicating their uncertainties. The dotted line shows the estimated average

$\hat{\theta}$ . (b) The total  $\chi^2$  function (dotted curve) as a function of  $\theta$  and the contributions  $\chi_1^2$  and  $\chi_2^2$  (full curves) from the two measurements.

sented by the two parabolic terms. The second one represents the density for the estimator value  $\hat{\theta}$  to lie around a true value  $\theta$ . In the *Bayesian* approach to statistics (see Section 1.4.4) this can be interpreted as a *probability density* for the true value to lie around the observed estimator value, assuming a flat prior for the parameter.

The first term in (2.56) represents the value of  $\chi_{\min}^2$  and is given by the squared difference of the two measurements normalised by the sum of the squared uncertainties. Thus, for repeated experiments, the  $\chi_{\min}^2$  should follow the  $\chi^2$  distribution with one degree of freedom. This is the simplest example for the  $\chi_{\min}^2$  *goodness-of-fit test*, which is discussed extensively in Section 3.8. The two measurements agree reasonably well with each other if  $\chi_{\min}^2$  is of order one, while for instance a value of ten would clearly indicate inconsistent values.

The formula (2.56) can be easily generalised for fitting a constant to  $N$  measurements, with  $N \geq 2$ . The  $\chi^2$  function can be split into a parabolic term for  $\theta$  around  $\hat{\theta}$  and an independent term for  $\chi_{\min}^2$ . The latter term quantifies the agreement between the data points. It should follow a  $\chi^2$  distribution with  $N - 1$  degrees of freedom, since one degree is ‘sacrificed’ in order to determine the average.

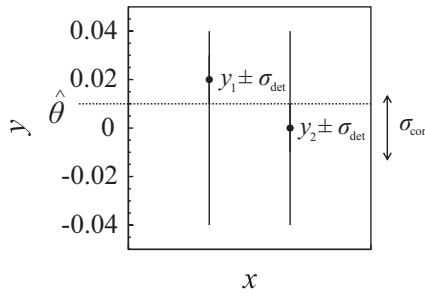
The case of averaging measurements with apparently inconsistent values is an important problem in high energy physics and elsewhere; it is discussed in Section 2.5.5.

#### 2.4.1.2 Averaging Correlated Measurements

Often measurements to be averaged are correlated. One typical example are measurements performed in different datasets obtained by the same experiment but applying the same detector calibrations causing a common systematic uncertainty. The  $\chi^2$  is given by

$$\chi^2 = (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}), \quad (2.57)$$

with the ‘parameter vector’  $\boldsymbol{\theta} = (\theta)$ ; the design matrix  $\mathbf{A}^T = (1, \dots, 1)$  is a unit



**Figure 2.6** Example of the weighted averaging of two measurements  $y_1 \pm \sigma_{\text{det}}$  and  $y_2 \pm \sigma_{\text{det}}$  with an additional fully correlated uncertainty  $\sigma_{\text{cor}}$ . The dotted horizontal line indicates the estimated average  $\hat{\theta}$ .

vector of  $N$  elements. Figure 2.6 illustrates the example of two correlated measurements being averaged. Here, the vertical position  $y$  of a particle flying horizontally, that is along the  $x$  direction, is measured in two detector planes located at different  $x$ . The detector planes are indicated in the figure by two thin vertical lines. The two measurements are  $y_1 \pm \sigma_{\text{det}}$  and  $y_2 \pm \sigma_{\text{det}}$ , where  $\sigma_{\text{det}}$  denotes the detector resolution. The whole detector has a global vertical (alignment) position uncertainty  $\sigma_{\text{cor}}$ . This has to be added in quadrature to both the diagonal and the off-diagonal elements of the covariance matrix of the two measurements, resulting in

$$\mathbf{V} = \begin{pmatrix} \sigma_{\text{det}}^2 + \sigma_{\text{cor}}^2 & \sigma_{\text{cor}}^2 \\ \sigma_{\text{cor}}^2 & \sigma_{\text{det}}^2 + \sigma_{\text{cor}}^2 \end{pmatrix}. \quad (2.58)$$

Inserting the measurements and their covariance into (2.52) and (2.53) one finds as a result for the weighted average

$$\hat{\theta} = \frac{y_1 + y_2}{2} \pm \sqrt{\sigma_{\text{det}}^2/2 + \sigma_{\text{cor}}^2}. \quad (2.59)$$

The same result would have been obtained by first averaging the two measurements without the common systematics and then adding the global position uncertainty in quadrature to the uncertainty of the weighted average.

However, a given source of systematics can affect the measurements differently, and the general form of the covariance matrix is given by

$$\mathbf{V} = \begin{pmatrix} \sigma_{\text{det}}^2 + \sigma_{\text{cor}_1}^2 & \sigma_{\text{cor}_1} \sigma_{\text{cor}_2} \\ \sigma_{\text{cor}_1} \sigma_{\text{cor}_2} & \sigma_{\text{det}}^2 + \sigma_{\text{cor}_2}^2 \end{pmatrix}. \quad (2.60)$$

Here,  $\sigma_{\text{cor}_1}$  and  $\sigma_{\text{cor}_2}$  are the signed shifts of the measurements obtained by varying the source of a systematic effect by  $1\sigma$  in a specific direction. For instance, if the detector position is fixed in the middle between the two measurement planes (because it is mounted there on a pole) but can rotate around this point, then a sys-

tematic uncertainty with negative correlation,  $\sigma_{\text{cor}_1} = -\sigma_{\text{cor}_2}$ , is obtained. In this case, the precision of the weighted average that takes the correlation into account is better than that obtained when adding a systematic uncertainty to the individual measurement uncertainties without taking correlations into account.

Another possibility is that the detector position is fixed at a point close to the first measurement plane but that again a rotational uncertainty is present. In this case, the second measurement at the position further away from the fixing point will have a larger uncertainty due to the larger lever arm ( $|\sigma_{\text{cor}_1}| < |\sigma_{\text{cor}_2}|$ ).

A further elaboration of the averaging of two measurements based on the  $\chi^2$  of (2.57) can be found in [9].

#### 2.4.1.3 Straight-Line Fit

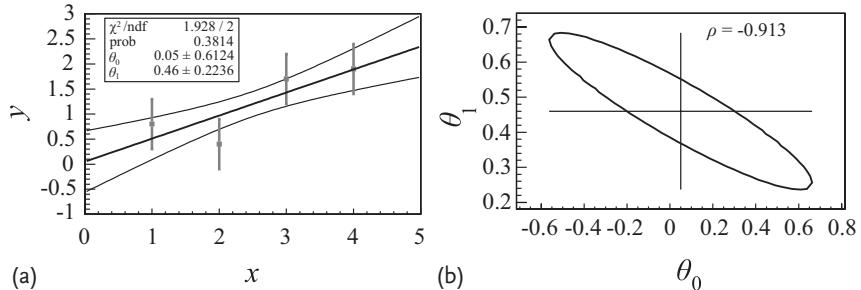
The straight-line fit is a classical linear least-squares fit example. One important application is a *trajectory fit* of a particle produced at a source at  $x = 0$  and measured in  $N$  detector layers. An illustrative example with four layers is shown in Figure 2.7.

The goal of the fit is to determine the unknown  $y$  position  $\theta_0$  at the source<sup>6)</sup> and the trajectory slope  $dy/dx = \theta_1$ . Assuming the measurements in the detector layers to be independent and with uniform Gaussian resolution  $\sigma$ , the  $\chi^2$  is given by

$$\chi^2 = \sum_{i=1}^N \frac{(\gamma_i - \theta_0 - x_i \theta_1)^2}{\sigma^2}, \quad (2.61)$$

or, equivalently,

$$\chi^2 = (\boldsymbol{\gamma} - \mathbf{A}\boldsymbol{\theta})^T \mathbf{V}^{-1} (\boldsymbol{\gamma} - \mathbf{A}\boldsymbol{\theta}), \quad (2.62)$$



**Figure 2.7** Straight-line particle-trajectory fit to position measurements in four detector layers. (a) The four position measurements, the fitted straight line and the two (thin) curves defining the borders of a 68% confidence interval for the trajectory. (b) The fitted pa-

rameters  $\hat{\theta}_0$  and  $\hat{\theta}_1$  with their  $1\sigma$  uncertainties (horizontal and vertical error bars) and the corresponding covariance ellipse given by the contour  $\chi^2 = \chi^2_{\min} + 1$ . The correlation coefficient  $\rho = V_{01}/\sqrt{V_{00}V_{11}}$  of the two parameters is  $\rho = -0.913$ .

6) For experiments with colliding beams, the determination of  $\theta_0$  may serve to determine whether an observed particle originated from the primary interaction point or from a secondary decay.

with fit parameter vector  $\boldsymbol{\theta}$ , design matrix  $\mathbf{A}$  and the covariance matrix  $\mathbf{V}$ :

$$\begin{aligned}\boldsymbol{\theta} &= \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}; \quad \mathbf{A} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}; \quad \mathbf{A}^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{pmatrix}; \\ \mathbf{V} &= \begin{pmatrix} \sigma^2 & 0 \\ \ddots & \ddots \\ 0 & \sigma^2 \end{pmatrix}.\end{aligned}\tag{2.63}$$

Using (2.52) the solution for the fit parameter vector is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}^{-1} \mathbf{y} = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1} \frac{1}{\sigma^2} \mathbf{A}^T \mathbf{y} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y} \\ &= \left( \frac{\sum_i 1}{\sum_i x_i} \quad \frac{\sum_i x_i}{\sum_i x_i^2} \right)^{-1} \left( \frac{\sum_i y_i}{\sum_i x_i y_i} \right) = \left( \frac{N}{N\bar{x}} \quad \frac{N\bar{x}}{N\bar{x}^2} \right)^{-1} \left( \frac{N\bar{y}}{N\bar{x}\bar{y}} \right) \\ &= \left( \frac{1}{\bar{x}} \quad \frac{\bar{x}}{\bar{x}^2} \right)^{-1} \left( \frac{\bar{y}}{\bar{x}\bar{y}} \right) = \frac{1}{\bar{x}^2 - \bar{x}^2} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \left( \frac{\bar{y}}{\bar{x}\bar{y}} \right) \\ &= \frac{1}{V[x]} \begin{pmatrix} \bar{x}^2 \bar{y} - \bar{x} \bar{x} \bar{y} \\ -\bar{x} \bar{y} + \bar{x} \bar{y} \end{pmatrix}.\end{aligned}\tag{2.64}$$

Here, the horizontal line over symbols (e.g.  $\bar{x}$ ) denotes the average value of the respective quantity over the detector layers  $i$ . The quantity  $V[x] = \bar{x}^2 - \bar{x}^2$  represents the squared spread of the detector layers in  $x$ . Using (2.53) the covariance matrix of  $\hat{\boldsymbol{\theta}}$  is

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = (\mathbf{A}^T \mathbf{V}^{-1} \mathbf{A})^{-1} = \frac{\sigma^2}{N V[x]} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.\tag{2.65}$$

The fit results are of remarkable simplicity. It is interesting to note that the standard deviation of the slope,

$$\sigma_{\theta_1} = \sqrt{V_{11}} = \frac{\sigma}{\sqrt{N V[x]}},$$

is reduced by a factor of two when halving the detector resolution  $\sigma$  or when doubling the spread  $\sqrt{V[x]}$  of the detector layers. In contrast, doubling the number of detector layers over the same spread will only result in a decrease by a factor  $1/\sqrt{2}$ . So for the slope measurement it is beneficial to have a large spread of detector layers, which is intuitive since the spread defines the lever arm of the detector.

Figure 2.7a shows the data points, the central straight-line fit defining the position estimate  $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$ , and the two lines  $\hat{y} \pm \sigma_{\hat{y}}$  with the standard deviation  $\sigma_{\hat{y}}$  obtained from error propagation:

$$\begin{aligned}\sigma_{\hat{y}} &= \sqrt{\left( \frac{\partial \hat{y}}{\partial \theta_0} \right)^2 V_{00} + \left( \frac{\partial \hat{y}}{\partial \theta_1} \right)^2 V_{11} + 2 \frac{\partial \hat{y}}{\partial \theta_0} \frac{\partial \hat{y}}{\partial \theta_1} V_{01}} \\ &= \sqrt{V_{00} + x^2 V_{11} + 2x V_{01}}.\end{aligned}\tag{2.66}$$

These two lines delimit a 68% confidence region of the particle trajectory which one can call the *trajectory uncertainty band*. Outside the  $x$  range of the detector the fit provides an extrapolation, and the uncertainty band increases linearly with the distance to the detector.

Figure 2.7b shows the covariance ellipse for the two fit parameters. They are largely anticorrelated as indicated by the orientation of the ellipse. The correlation coefficient<sup>7)</sup>  $\rho = V_{01}/\sqrt{V_{00}V_{11}}$  of the two parameters is  $\rho = -0.913$ . This anticorrelation can be completely avoided by moving the coordinate system to the middle of the detector such that in the new system  $\bar{x} = 0$ . Then the off-diagonal terms in the covariance matrix (2.65) vanish. However, this is not useful for trajectory measurements since one can measure the particle only in one direction (the flight direction) from the source at which one wants to determine the position  $\theta_0$ .

#### 2.4.2

##### Non-linear Least-Squares Fits

The dependence of the fit function  $f(x; \boldsymbol{\theta})$  on the fit parameters  $\boldsymbol{\theta}$  can be highly non-linear, for instance for the one-parameter function  $y = \theta e^{-\theta x}$ . In such cases the minimisation of the  $\chi^2$  function is usually done in an iterative way, using, for example, the *Newton–Raphson* method to determine the zero point  $\hat{\theta}$  of the gradient of the  $\chi^2$  function. For a one-parameter fit function one can define

$$g = \frac{\partial \chi^2}{\partial \theta} \quad \text{and} \quad G = \frac{\partial^2 \chi^2}{\partial \theta^2} = \frac{\partial g}{\partial \theta}. \quad (2.67)$$

If one has found (guessed) a sensible starting point  $\theta^{(0)}$  such that between this point and  $\hat{\theta}$ , where the function  $g$  crosses zero,  $g$  is approximately a straight line, then the Newton step

$$\delta \theta = -\frac{g(\theta^{(0)})}{G(\theta^{(0)})} \quad (2.68)$$

will lead to a first iteration point  $\theta^{(1)} = \theta^{(0)} + \delta \theta$  which is already close to the solution  $\hat{\theta}$ . An illustration of the Newton step is provided in Example 2.4.

For multi-parameter fits it is straightforward to generalise the Newton-step procedure by introducing the vector

$$\mathbf{g} = \frac{\partial \chi^2}{\partial \boldsymbol{\theta}^T} \quad (2.69)$$

and the matrix

$$G_{jk} = \frac{\partial^2 \chi^2}{\partial \theta_j \partial \theta_k}. \quad (2.70)$$

7) The correlation coefficient  $\rho$ , introduced in (1.14), specifies the following: if one parameter, e.g.  $\theta_0$  is varied (shifted) from  $\hat{\theta}_0$  to  $\hat{\theta}_0 + \Delta$ , then the other parameter  $\theta_1$  has to be shifted from  $\hat{\theta}_1$  to  $\hat{\theta}_1 + \rho \sigma_{\theta_1} (\Delta / \sigma_{\theta_0})$ , in order to keep the increase of the  $\chi^2$  minimal.

This defines a new measurement value of  $\theta_1$  for any value  $\theta_0$ , which can be useful, e.g. if  $\theta_0$  is fixed by an external constraint. In this case the uncertainty of  $\theta_1$  will be reduced by a factor  $\sqrt{1 - \rho^2}$  (to be shown in Exercise 2.5).

The first Newton step is then

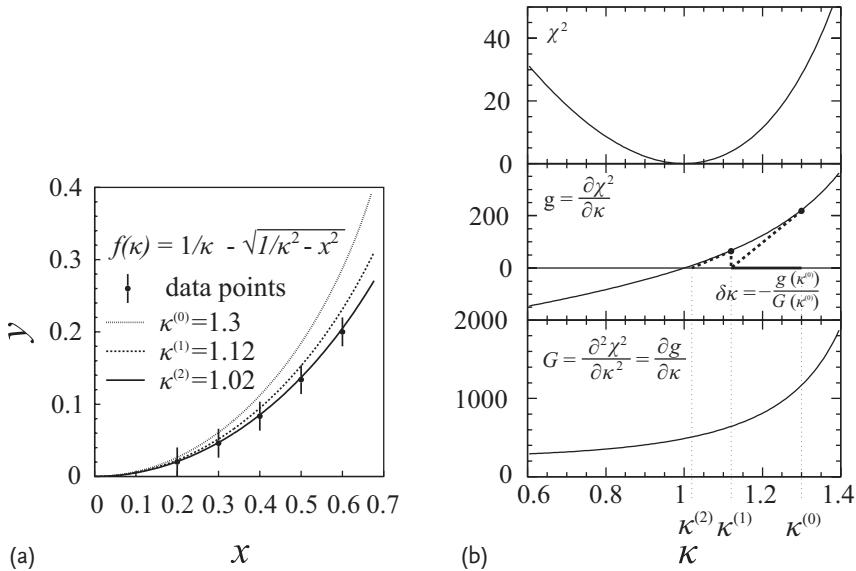
$$\delta\theta = -G^{-1}(\theta^{(0)})g(\theta^{(0)}). \quad (2.71)$$

There are many practical details involved in numerical minimisations such as the one presented here. As already mentioned above, it is advisable to use standard software tools like the MINUIT program [5] and not to attempt to develop new minimisation code.

#### Example 2.4 Non-linear least squares: circle trajectory fit

An example of a numerical minimisation by Newton steps, applied to a non-linear least-squares fit problem, is shown in Figure 2.8.

A charged particle is guided along a beamline and enters a spectrometer at  $(x, y) = (0, 0)$  with its direction along the  $x$  axis. The spectrometer is immersed in a magnetic field parallel to the  $z$  direction which forces the particle onto a circular trajectory in the  $x-y$  plane. The particle is ‘tracked’ in five detector layers, providing measurements of the  $y$  position at fixed  $x$  positions, with a resolution  $\sigma_y = 0.02$ . The goal is to determine the radius  $R$  of the circle trajectory from which one can directly infer the particle transverse momentum since both are proportional to each other. Instead of  $R$  we fit the curvature  $\kappa = 1/R$ . This provides better numerical stability for particles with high transverse momenta where the radius goes to infinity.



**Figure 2.8** Non-linear least-squares fit to a circle trajectory. (a) The measurements of the particle vertical position  $y$  in five detector layers at fixed  $x$ , indicated as points. Three circle curves are also shown for different values of

$\kappa$ . (b) The  $\chi^2$  function and its first and second derivatives with respect to  $\kappa$ . The first Newton step  $\delta\kappa$  is also indicated, leading from the start value  $\kappa^{(0)}$  to  $\kappa^{(1)}$ , while the second step leads to  $\kappa^{(2)}$ .

The least-squares function  $\chi^2$  for this problem is given by

$$\chi^2 = \sum_{i=1}^5 \left( \frac{y_i - f_i}{\sigma_y} \right)^2, \quad \text{with} \quad f_i = \frac{1}{\kappa} - \sqrt{\frac{1}{\kappa^2} - x_i^2}. \quad (2.72)$$

In Figure 2.8a the fit function is shown for three values of  $\kappa$  which correspond to an (arbitrary) initial starting value  $\kappa^{(0)} = 1.3$ , and the two values  $\kappa^{(1)} = 1.12$  and  $\kappa^{(2)} = 1.02$  obtained after the first and second Newton steps. Figure 2.8b shows, as a function of  $\kappa$ , the  $\chi^2$  curve (top) and the first derivative and second derivative functions  $g$  and  $G$  (middle and bottom). For the curve corresponding to the function  $g$ , the first and second Newton steps are indicated, based on the extrapolation of the linearised  $g$  to the point where  $g = 0$ . It can be seen that the convergence is good and that, with a third step, one would come very close to the final solution  $\hat{\kappa} = 1.0$ .

Please take note that the circle trajectory fit was presented here in a most simplified way, mainly for illustration purposes. The real problem is more involved, see for instance [10], where a fast method is presented which has been used in numerous high energy physics experiments.

#### 2.4.2.1 Non-linear Least Squares: Mass-Peak Fit (Signal Position)

Another example of a non-linear least-squares fit problem is shown in Figure 2.9. Pairs of muons have been recorded in a detector, and the spectrum of their invariant mass  $m_{\mu^+\mu^-}$  is analysed. The spectrum contains a signal of a resonance with unknown mass  $M$  to be determined. The data are fitted with the function

$$f(m_{\mu^+\mu^-}; M) = B + S \cdot e^{-\frac{(m_{\mu^+\mu^-} - M)^2}{2\sigma^2}}. \quad (2.73)$$

Here,  $B$  represents the known flat background component. The signal is parameterised by a Gaussian function with known width  $\sigma$  representing the detector resolution, the known signal normalisation  $S$ , fixed from a prediction, and the unknown mass  $M$ .<sup>8)</sup> The fit can be performed with the least-squares method using the Neyman  $\chi^2$  function<sup>9)</sup> [11]

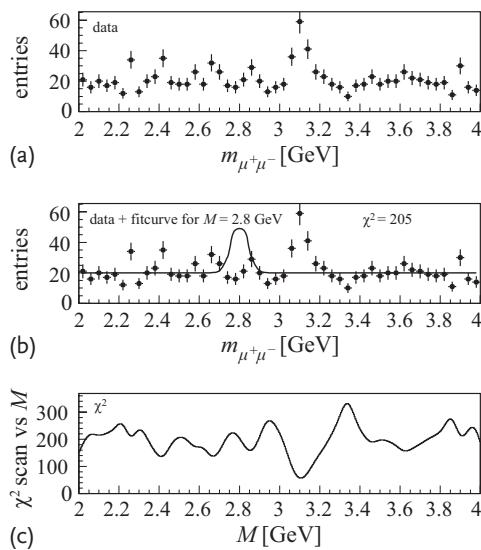
$$\chi^2 = \sum_{\text{bin } i} \frac{[k_i - f_i(m_{\mu^+\mu^-}; M)]^2}{k_i}. \quad (2.74)$$

Here  $k_i$  denotes the number of observed events in bin  $i$  and  $f_i$  the expected number of events obtained by integrating the function  $f$  over  $m_{\mu^+\mu^-}$  within the boundaries of bin  $i$  for a given mass hypothesis  $M$ .

The fit is performed by finding the value of  $M$  for which  $\chi^2$  becomes minimal. Figure 2.9a shows the recorded mass spectrum. In Figure 2.9b the fit function is shown in addition, for a mass hypothesis of  $M = 2.8$  GeV. Figure 2.9c shows the

8) The natural width of the resonance is assumed to be negligible.

9) The Neyman  $\chi^2$  is discussed in detail in Section 2.5.4.



**Figure 2.9** Fit of a resonance peak on a flat background. (a) The observed distribution of reconstructed dimuon masses  $m_{\mu^+\mu^-}$ . (b) The fit curve, consisting of a Gaussian signal

peak plus a flat background, for an exemplary central peak position  $M = 2.8$  GeV, on top of the data. (c) The  $\chi^2$  of the fit versus  $M$ .

scan of the  $\chi^2$  function versus  $M$ . A global minimum  $\chi^2_{\min} \sim 59$  is observed at a mass value  $M = 3.12^{+0.01}_{-0.01}$  GeV, where the uncertainties have been obtained using the  $\chi^2 = \chi^2_{\min} + 1$  method (see (2.46)). A striking feature revealed by the scan is the existence of additional (local) minima of the  $\chi^2$  function, for instance at mass values of  $M = 2.4$  GeV and  $M = 3.6$  GeV. This can be easily understood, since at these values the data fluctuate upwards and mimic a signal.

If the Newton step method is applied and one starts close to one of the local minima, one is in danger of getting caught there while missing the global minimum. Performing a global scan rescues the situation.

Another criterion that strongly separates the global minimum from the local minima is the value of the respective  $\chi^2_{\min}$ . For a reasonable fit, the  $\chi^2_{\min}$  value is expected to have a similar size as the number of degrees of freedom of the fit,  $\text{ndf}$ , which is  $\text{ndf} = 59$  for our example<sup>10)</sup>. This is due to the facts that for repeated experiments the  $\chi^2_{\min}$  value should approximately follow a  $\chi^2$  distribution with  $\text{ndf}$  degrees and that the expectation value of  $\chi^2$  is equal to  $\text{ndf}$  (see Section 1.3.4.6). For the global minimum the value of  $\chi^2_{\min} \approx 59$  matches  $\text{ndf}$  and thus indicates a good fit quality. However, for the local minima the  $\chi^2_{\min}$  values are much higher, for instance about 150 for the minimum at 2.4 GeV, indicating an unsatisfactory description of the data by the fit function. A detailed discussion on the use of  $\chi^2_{\min}$

10) The number  $\text{ndf}$  of degrees of freedom is given by the number of fitted bins minus the number of free fit parameters. In our example there are 60 bins and one fitted parameter.

and other observables for testing the quality of a fit is given in Section 3.8.

Smaller  $\chi^2_{\min}$  values would have been obtained for the local minima if the signal parameter  $S$  would not have been fixed, but had also been fitted to the data. This is the case if one searches for a signal of unknown position and unknown strength. In such cases, the so-called *look-elsewhere effect* can occur, that is fake signals at random mass values. These signals are caused by upward fluctuations of the data and can lead to local  $\chi^2$  minima with reasonable  $\chi^2_{\min}$  values. For a further discussion of the look-elsewhere effect see Section 3.5.4.

## 2.5

### Maximum-Likelihood Fits: Unbinned, Binned, Standard and Extended Likelihood

We now return to the maximum-likelihood estimate (MLE) method, as introduced in Section 2.3, and discuss different applications in high energy physics.

Maximum-likelihood fits can be applied in an *unbinned* or *binned* way:

- The standard MLE method is the *unbinned MLE*, where each event enters separately into the likelihood function (2.5). An already discussed example is the unbinned fit of the exponential decay (see Example 2.2). The unbinned MLE is *statistically optimal*. However, its CPU consumption increases linearly with the number of events.
- In high energy physics, the *binned MLE* is very popular: here, the events are grouped in bins of an observable  $x$ , and the counted event numbers in the bins are used in the likelihood function. This method is further discussed in Section 2.5.3.

Different types of parameters can be fitted with the MLE method:

- *Shape parameters* of the underlying probability density function, as we have seen already for the exponential decay where we fitted the decay time parameter  $\tau$  (see Example 2.2). A further numerical example is discussed below (see Example 2.5).
- *Fractions of certain event classes* which contribute to the data sample, for example signal and background, as illustrated in Section 2.5.1.1.
- If one is not only interested in the fractions but in the *total normalisations of processes*, one should use the *extended MLE*, as described in Section 2.5.2.

#### 2.5.1

##### Unbinned Maximum-Likelihood Fits

We discuss in the following example a classical application of the unbinned MLE, the fit of data distributed according to a straight line pdf.

**Example 2.5 Unbinned straight-line fit**

For the already introduced weak decay of a muon (see Example 2.3) we perform now an unbinned maximum-likelihood fit. The probability density function of the angle  $\alpha$  of the emitted positron's flight direction with respect to the muon spin direction is given by

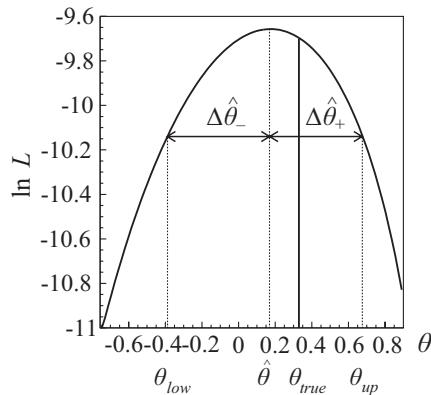
$$f(x; \theta) = \frac{1 + \theta x}{2}, \quad \text{with } x = \cos(\alpha), \quad (2.75)$$

and the goal is to determine the unknown polarisation parameter  $\theta$ . Figure 2.5 and Table 2.5 illustrate an example with 14 events simulated according to the above pdf with a true value  $\theta = 1/3$ . Table 2.2 lists the observed  $x$  values which are used to construct the likelihood function for a given  $\theta$ :

$$L = \prod_{i=1}^{14} \frac{1 + \theta x_i}{2}. \quad (2.76)$$

**Table 2.2** The 14  $x$  positions of the data points used in an unbinned MLE of a straight line.

Event	$x$	Event	$x$
1	0.251	8	-0.020
2	-0.581	9	0.595
3	0.554	10	0.008
4	-0.365	11	-0.475
5	0.230	12	0.592
6	0.623	13	0.017
7	-0.019	14	-0.876



**Figure 2.10** Unbinned maximum-likelihood fit of a straight line  $1/2(1 + \theta x)$  to 14 simulated data points. Shown are the log-likelihood curve as a function of the slope parameter  $\theta$ , the maximum-likelihood estimate  $\hat{\theta}$ , the points  $\theta_{\text{low}} = \hat{\theta} - \Delta\hat{\theta}_-$  and  $\theta_{\text{up}} = \hat{\theta} + \Delta\hat{\theta}_+$  defining a 68% CL interval for  $\theta$ , and the true value  $\theta_{\text{true}} = 1/3$ .

Figure 2.10 shows the resulting log-likelihood function as a function of  $\theta$ . As before, the position at the maximum defines the best estimate  $\hat{\theta}$ , and the two points  $\theta_{\text{low}} = \hat{\theta} - \Delta\hat{\theta}_-$  and  $\theta_{\text{up}} = \hat{\theta} + \Delta\hat{\theta}_+$ , where  $\ln L$  drops by 1/2, specify a 68% CL region  $[\theta_{\text{low}}, \theta_{\text{up}}]$  for  $\theta$ . Using the shorthand notation of (2.26), the obtained result is  $\theta = 0.17^{+0.40}_{-0.56}$ . The true value  $\theta_{\text{true}} = 1/3$  is well covered within the 68% CL interval. However, it is obvious that this measurement of  $\theta$  is very inaccurate and that many more events would be needed for a precise determination.

### 2.5.1.1 Unbinned MLE: Fitting Fractions of Processes

In high energy physics one is often interested in the *fractions* of different event classes contributing to a selected data sample. These classes could be, for instance, different resonances contributing to a recorded mass or energy spectrum. The relative production fractions of the resonances could be related to underlying theory parameters of interest, for example to the quantum numbers of the resonances.

In the following, for simplicity, we discuss the case of two event classes contributing to the data (e.g. two signals or one signal and a background). The two processes differ in the shape of the distribution of a suitable variable  $x$ , which could represent an invariant mass or the output of an MVA classifier (see Chapter 5). The shape difference allows the two fractions to be determined using the MLE method and fitting the function

$$f(x; f_s) = f_1 p_1(x) + (1 - f_1) p_2(x). \quad (2.77)$$

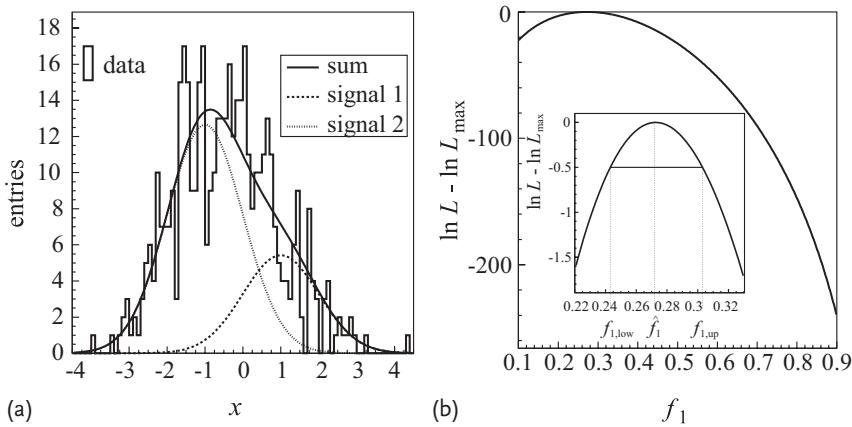
Here  $f_1$  and  $1 - f_1$  denote the unknown fractions of the first and second process and  $p_1$  and  $p_2$  their probability density functions, which are assumed to be known. Because of the normalisation constraint there is only one free parameter  $f_1$ , and the two fractions are completely anticorrelated.

An example is illustrated in Figure 2.11. Both signals are assumed to follow Gaussian distributions in the observable  $x$ , with the same width equal to unity and different mean values  $x = 1$  and  $x = -1$ , respectively. A sample of 453 simulated events is shown<sup>11)</sup> in Figure 2.11a, generated according to a true fraction  $f_1 = 0.3$ . The unbinned likelihood function is given by

$$L = \prod_{i=1}^{453} \left[ f_1 e^{-(x_i - 1)^2/2} + (1 - f_1) e^{-(x_i + 1)^2/2} \right], \quad (2.78)$$

where the common constant normalisation terms  $1/\sqrt{2\pi}$  of the two Gaussians have been dropped. Figure 2.11b shows the log-likelihood function, after subtracting its observed maximum value  $\ln L_{\text{max}} = -765.7$ , as a function of  $f_1$ . It is to very good approximation parabolic. As one can read off from the curve, the fit yields the result  $f_1 = 0.273 \pm 0.030$ . The fitted function is also shown in Figure 2.11a; it describes the data well. Despite the large overlap of the two signals which makes them

<sup>11)</sup> For illustration purposes, the data are shown in a binned histogram, but the fit is performed unbinned.



**Figure 2.11** Unbinned maximum-likelihood fit of the fractions of two Gaussian signals. (a) Simulated data (binned histogram) and the two fitted Gaussian functions. (b) Unbinned

log-likelihood function, after subtracting the maximum value, as a function of the fraction  $f_1$  of the first signal. The small insert shows a zoom into the maximum region of  $\ln L$ .

look like one signal, a quite precise determination of their fractions is obtained. The situation would be much more difficult if the peak positions and widths of the two signals were not known but would also have to be determined from the fit to the data.

### 2.5.2 Extended Maximum Likelihood

Another frequently encountered task in high energy physics is the determination of *absolute rates* (or *normalisations*) of physics processes, for instance for Higgs production at the LHC. The rates correspond to production cross sections which can be predicted by theory. When repeating an experiment with identical conditions, the observed rate  $N$  of a process will fluctuate according to a Poisson distribution around the expected (true) value  $\nu$ . The Poisson term can be incorporated as a multiplicative term in the likelihood function, yielding the *extended likelihood*:

$$L(\mathbf{x}; \nu, \boldsymbol{\theta}) = e^{-\nu} \frac{\nu^N}{N!} \prod_{i=1}^N f(x_i; \boldsymbol{\theta}). \quad (2.79)$$

Here, the product of the functions  $f$  is the standard unbinned likelihood (2.5). Dropping terms which do not depend on  $\boldsymbol{\theta}$  or  $\nu$ , the log-likelihood function can then be expressed as

$$\ln L(\mathbf{x}; \nu, \boldsymbol{\theta}) = \sum_{i=1}^N \ln f(x_i; \boldsymbol{\theta}) + N \ln \nu - \nu + \text{constant}. \quad (2.80)$$

When the expected number of events  $\nu$  is independent of the parameters  $\boldsymbol{\theta}$ , the estimated value is equal to the observed value,  $\hat{\nu} = N$ . This can be easily derived

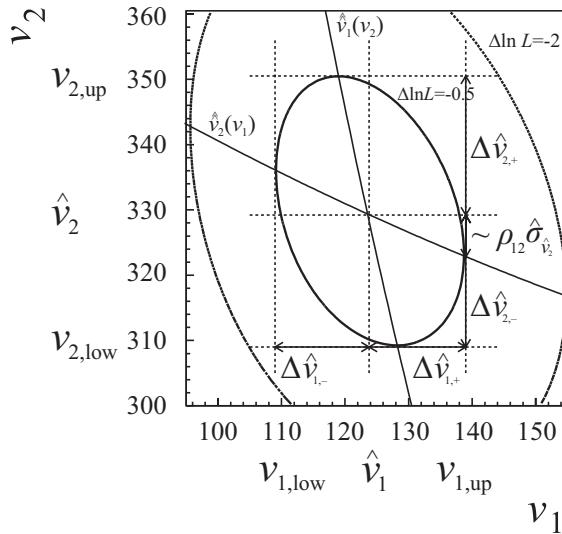
from (2.80), by solving  $\partial \ln L / \partial \nu = 0$  for  $\nu$ . The estimate for the other parameters will also take the same values as in a non-extended fit. If, however, the expected number of events  $\nu$  is a function of the other parameters  $\boldsymbol{\theta}$ , this dependence can be exploited to obtain a better (more efficient) estimate of both  $\boldsymbol{\theta}$  and  $\nu$  as will be shown in Section 2.5.2.2. For a detailed discussion of the extended MLE method see also [12].

### 2.5.2.1 Unbinned Extended MLE: Fitting Rates of Processes

We now come back to the example of a fit of two signals (see Section 2.5.1.1). While previously we were interested in the fractions of the two signals, we now want to fit the expected rates  $\nu_1$  and  $\nu_2$ . The extended likelihood function is given by

$$\begin{aligned} L &= e^{-\nu} \nu^{453} \prod_{i=1}^{453} \left[ f_1 e^{-(x_i-1)^2/2} + (1-f_1) e^{-(x_i+1)^2/2} \right] \\ &= e^{-\nu_1 - \nu_2} \prod_{i=1}^{453} \left[ \nu_1 e^{-(x_i-1)^2/2} + \nu_2 e^{-(x_i+1)^2/2} \right], \end{aligned} \quad (2.81)$$

where we have used the equivalence  $\nu_1 = f_1 \nu$  and  $\nu_2 = (1-f_1) \nu$  and where multiplicative terms of constant value have been dropped. Figure 2.12 indicates the position of the maximum of the log-likelihood function at  $(\hat{\nu}_1, \hat{\nu}_2) = (124, 329)$ . Also shown are the two contours where  $\ln L$  has dropped by  $1/2$  and  $2$  from its max-



**Figure 2.12** Extended maximum-likelihood fit of the rates  $\nu_1$  and  $\nu_2$  of the two signals contributing to the data illustrated in Figure 2.11. The values  $\hat{\nu}_1, \hat{\nu}_2$  where the maximum likelihood is reached are indicated as well as the

contours  $\Delta \ln L = -1/2$  and  $\Delta \ln L = -2$ . Also shown are the profiled curves  $\hat{\nu}_2(\nu_1)$  and  $\hat{\nu}_1(\nu_2)$  and the positive and negative uncertainties  $\Delta \nu_{i,\pm}$ , with  $i = 1, 2$ .

imum value. These contours delimit two-dimensional  $(\theta_1, \theta_2)$  confidence regions at 38% and 86% confidence level, respectively (see discussion in Section 2.3.4.1). Furthermore, the profiled curves  $\hat{\nu}_2(\nu_1)$  and  $\hat{\nu}_1(\nu_2)$  are indicated in the figure, that is the points in  $\nu_2$  where  $\ln L$  has a maximum for given fixed  $\nu_1$  and vice versa. According to the profile likelihood method (Section 2.3.4.2), the two points where  $\hat{\nu}_2(\nu_1)$  crosses the  $\Delta \ln L = -0.5$  contour,  $\nu_{1,\text{low}} = \hat{\nu}_1 - \Delta \hat{\nu}_{1,-}$  and  $\nu_{1,\text{up}} = \hat{\nu}_1 + \Delta \hat{\nu}_{1,+}$ , define a 68% CL interval for  $\nu_1$ . By construction, these two points are extreme values of  $\nu_1$  on the contour. Similarly one can construct a 68% CL interval  $[\nu_{2,\text{low}}, \nu_{2,\text{up}}]$  for  $\nu_2$  from the  $\hat{\nu}_1(\nu_2)$  curve, which is also shown in Figure 2.12. The results, in the usual shorthand notation, are  $\nu_1 = 124^{+15}_{-15}$  and  $\nu_2 = 329^{+21}_{-21}$ .

The negative slopes of the two profile curves illustrate the anticorrelation between the two signal strengths. This is easy to understand since the signals overlap (see Figure 2.11), and an increase in one signal can be partially compensated by a decrease in the other. In the Gaussian approximation of the likelihood function (i.e. the  $\ln L$  constant-value contours are ellipses and  $\Delta \hat{\nu}_{i,-} = \Delta \hat{\nu}_{i,+} = \hat{\sigma}_{\hat{\nu}_i}$  with  $i = 1, 2$ ) the variation of  $\nu_2$  along the profiled curve, corresponding to a  $1\sigma$  step in  $\nu_1$ , is equal to  $\rho_{12} \hat{\sigma}_{\hat{\nu}_2}$ . This relation can be used to determine the correlation coefficient  $\rho_{12}$ , as illustrated in Figure 2.12. In our example the  $1\sigma$  step for  $\nu_1$  is from 124 to 139, and  $\nu_2$  varies by  $\sim -7$  so that  $\rho_{12} \approx -0.33$ .

### 2.5.2.2 Unbinned Extended MLE: Fitting a Signal with Position-Dependent Normalisation

Let us now discuss an example in which the use of the extended MLE significantly improves the precision of the fitted parameters compared to the standard MLE.

The data sample is assumed to consist of events from a single signal process with an expected rate  $\nu$  that is a known function of the signal position  $\mu$  ( $\mu$  could be the mass of the produced particle). Actually, in collider experiments this is often the case – for example in the production of the top quark at the LHC: the larger the top mass is, the smaller the total production cross section will be, mainly due to the decreasing kinematical phase space.

A simulated example is shown in Figure 2.13. The events were generated according to Poisson statistics with arbitrary true expectation value  $\nu = 9$ . In the position observable  $x$  they were generated according to a Gaussian (reflecting the detector resolution) around the true peak position  $\mu = 7$  and with a width of unity. The distribution of the 11 observed events is shown in Figure 2.13a. Figure 2.13b illustrates the assumed exponential dependence<sup>12)</sup> of the expected rate  $\nu$  on the true  $\mu$ :

$$\nu = 9e^{-4(\mu-7)}. \quad (2.82)$$

Obviously  $\nu$  is nine for the true value  $\mu = 7$  but increases steeply towards smaller  $\mu$  values. Inserting this dependence into (2.80) leads to the *extended*

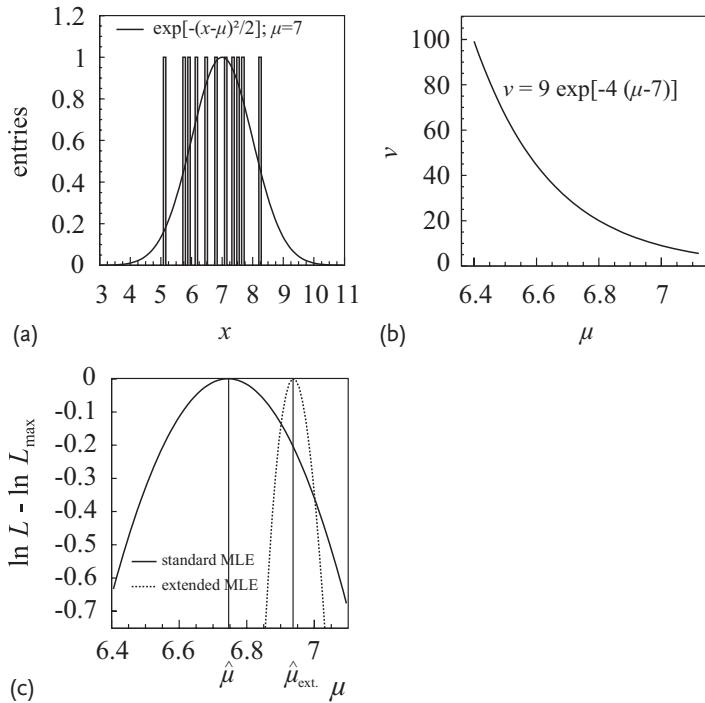
12) Both the exponential dependence and the value of  $-4$  for the exponential slope were arbitrarily chosen for this example. In reality the dependence would have to be predicted by a theory model.

log-likelihood function (omitting constant terms)

$$\ln L = \sum_{i=1}^{11} -\frac{(x_i - \mu)^2}{2} - 9e^{-4(\mu-7)} + 11 \ln(9e^{-4(\mu-7)}) . \quad (2.83)$$

Figure 2.13c shows  $\ln L$  versus the peak position  $\mu$ . It is a narrow approximate parabola and the obtained fit result is  $\mu = 6.94 \pm 0.08$ . The standard log-likelihood function, using only the first term in (2.83) (the term with the sum over the 11 events), is also shown. It is the function representing the standard weighted average and yields  $\mu = 6.7 \pm 0.3$ .

In our example the normalisation parameter  $\nu$  and the position parameter  $\mu$  are completely *dependent* on each other. Instead of maximising the extended log-likelihood function (2.83) with respect to  $\mu$ , one can replace  $\mu$  by  $\nu$ , inverting the relation (2.82), and maximise the function with respect to  $\nu$ . This will lead to a more accurate estimate of  $\nu$  than  $\hat{\nu} = 11$ , because now not only the Poisson terms in (2.83) (the last two terms) are used to constrain  $\nu$ , but also the first term with



**Figure 2.13** Extended unbinned maximum likelihood fit of the peak position  $\mu$  of a Gaussian signal. (a) The 11 simulated data points as a function of the observed position  $x$  and the Gaussian pdf (drawn with arbitrary normalisation) with true peak position  $\mu = 7$ . (b) Expected signal event rate  $\nu$  as a function of

the peak position  $\mu$ , in a small region around  $\mu = 7$ . (c) Log-likelihood function, after subtracting the maximum value, as a function of  $\mu$ , for the extended MLE (dotted curve) and the standard MLE (solid curve). The estimated values  $\hat{\mu}_{\text{ext}}$  and  $\hat{\mu}$  are also indicated.

the sum over the 11 events. For our example the gain is very small, since the  $\nu(\mu)$  dependence (2.82) is very steep, which leads to a good constraint on  $\mu$  from the number of observed events but to a weak constraint of  $\nu$  from the observed broad mass distribution. However, for a shallow  $\nu(\mu)$  dependence and a narrow mass peak distribution the situation can be reversed.

### 2.5.3

#### Binned Maximum-Likelihood Fits

When fitting a data sample with large size (large  $N$ ), it is common use to bin the data in order to be more efficient in the computation of the likelihood function. This is not harmful as long as the information on the parameter vector  $\boldsymbol{\theta}$  from the variation of the probability density function  $f(x; \boldsymbol{\theta})$  within a bin is insignificant compared to the one from the variation over all the bins.

If the total number of events  $N$  is fixed, then the probability distribution for the bins is multinomial. The likelihood function is then

$$L = N! \prod_{i=1}^B \frac{P_i(\boldsymbol{\theta})^{n_i}}{n_i!}, \quad (2.84)$$

and the log-likelihood function to be maximised is

$$\ln L = \sum_{i=1}^B n_i \ln P_i(\boldsymbol{\theta}) + \text{constant}. \quad (2.85)$$

Here  $B$  is the number of bins and  $n_i$  the observed number of entries for bin  $i$ . The quantity  $P_i$  denotes the expected probability for an event to appear in bin  $i$  and is given by

$$P_i(\boldsymbol{\theta}) = \int_{x_i^{\text{low}}}^{x_i^{\text{up}}} f(x; \boldsymbol{\theta}) dx, \quad (2.86)$$

where  $x_i^{\text{low}}$  and  $x_i^{\text{up}}$  are the bin limits.<sup>13)</sup> If the bin size is small enough, the approximation that the probability density function is constant or linear within a bin is often used, and the integral is replaced by the evaluation of the function at the bin centre  $x_i^c$ :

$$P_i(\boldsymbol{\theta}) = \Delta x_i f(x_i^c; \boldsymbol{\theta}), \quad (2.87)$$

with  $\Delta x_i = x_i^{\text{up}} - x_i^{\text{low}}$  denoting the bin width. It is trivial to see that in the limit of zero bin width the binned likelihood becomes exactly the same as the unbinned likelihood discussed before (besides multiplicative terms which do not depend on  $\boldsymbol{\theta}$ ).

<sup>13)</sup> It is assumed that the bins cover the total  $x$  range, such that  $\sum_{i=1}^B P_i(\boldsymbol{\theta}) = 1$ , otherwise the  $P_i(\boldsymbol{\theta})$  have to be renormalised:  $P_i(\boldsymbol{\theta}) \rightarrow P_i(\boldsymbol{\theta}) / \sum_{i=1}^B P_i(\boldsymbol{\theta})$ .

In the extended-likelihood case, the likelihood (2.84) is multiplied by a Poisson term for  $N$  observed events when  $\nu$  events are expected,

$$L = e^{-\nu} \frac{\nu^N}{N!} N! \prod_{i=1}^B \frac{P_i^{n_i}}{n_i!} = \prod_{i=1}^B e^{-\nu_i} \frac{\nu_i^{n_i}}{n_i!}, \quad (2.88)$$

where for the last transformation we have used  $N = \sum n_i$  and introduced  $\nu_i = P_i \nu$  such that  $\sum \nu_i = \nu$ . Thus, the likelihood is the product of the bin-wise Poisson probabilities of observing  $n_i$  events in bin  $i$  when  $\nu_i$  events are expected. The log-likelihood function is given by

$$\ln L(\mathbf{n}; \nu, \boldsymbol{\theta}) = \sum_{i=1}^B n_i \ln \nu_i(\nu, \boldsymbol{\theta}) - \nu + \text{constant}. \quad (2.89)$$

The main field of applications of binned extended-likelihood fits are the same as for the unbinned case discussed in Section 2.5.2: fits where one is interested in determining absolute rates of processes (e.g. cross sections) and fits with parameters that both influence the shape and normalisation of the fit function. The latter allows an improved determination of these parameters compared to the standard maximum-likelihood fit.

#### 2.5.4

##### Least-Squares Fit to a Histogram

If in each bin  $i$  the expected number of events  $\nu_i$  is not too small, one can use the Gaussian approximation of the Poisson distribution and apply a least-squares fit to estimate the parameters  $\boldsymbol{\theta}$ , by minimising the  $\chi^2$  defined by Pearson [13]:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^B \frac{[n_i - \nu_i(\boldsymbol{\theta})]^2}{\nu_i(\boldsymbol{\theta})}. \quad (2.90)$$

A commonly used approach is to approximate the exact variance  $\sigma_i^2 = \nu_i$  of the number of entries by the observed number of entries,  $\sigma_i^2 = n_i$ . This leads to the  $\chi^2$  defined by Neyman [11]:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^B \frac{[n_i - \nu_i(\boldsymbol{\theta})]^2}{n_i}. \quad (2.91)$$

The latter method is easier and faster to compute. In particular, in the case of linear functions  $\nu_i(\boldsymbol{\theta})$ , it corresponds to a *linear least-squares estimator*, which can be solved analytically using (2.52) and (2.53). However, problems arise if there are bins that have few entries (typically less than 5). For such bins the variance of the number of bin entries is poorly estimated. Bins with zero entries lead to an estimated variance of zero and cannot be used in (2.91). One typically omits these bins in the  $\chi^2$  minimisation. This will, however, lead to biased fit results. Also for the Pearson  $\chi^2$ , formula (2.90), bins become problematic if  $\nu_i(\boldsymbol{\theta}) \rightarrow 0$ .

The  $\chi^2$  fits to binned histograms are typically *extended fits*, that is one fits also the expected number of events:  $\nu_i(\boldsymbol{\theta}) = \nu p_i(\boldsymbol{\theta})$ , with  $\nu$  as an extra fit parameter and  $p_i$  denoting the probability to find an event in bin  $i$ . For the Pearson  $\chi^2$  one finds, by solving  $\partial\chi^2/\partial\nu = 0$ , as the best estimate

$$\hat{\nu}^{\text{Pearson}} = N + \frac{\chi_{\min}^2}{2} \quad (2.92)$$

and for the Neyman  $\chi^2$

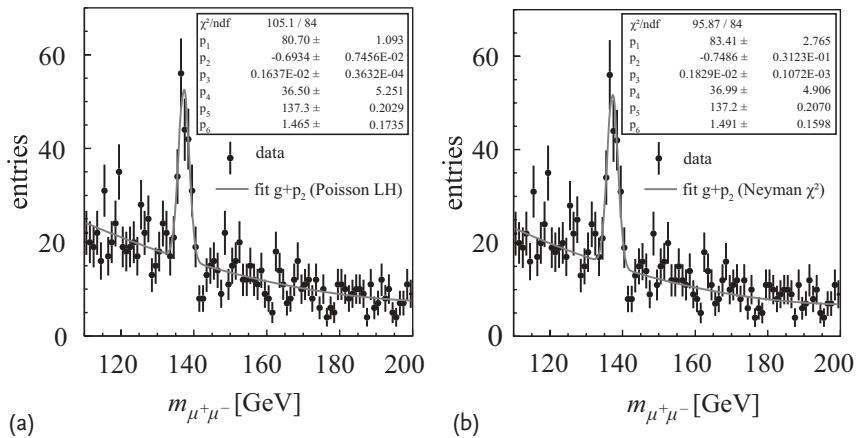
$$\hat{\nu}^{\text{Neyman}} = N - \chi_{\min}^2. \quad (2.93)$$

Both estimates are biased. Since the  $\chi^2$  function approximately follows a  $\chi^2(B-m)$  distribution, and  $E(\chi^2) = B-m$  (where  $B$  is the number of bins and  $m$  the number of parameters), the bias is negligible only if the observed number of events  $N \gg B-m$ . In contrast to this, the extended maximum-likelihood method for binned data (that is using (2.89)) does not suffer from this bias and from the problem of bins with low statistics; it is therefore the recommended method to use for fitting binned data.

#### 2.5.4.1 Binned Mass-Peak Fit: Practical Considerations

The fit of a signal peak on top of background is a classical data analysis task. A simulated example is illustrated in Figure 2.14. Pairs of muon candidates are selected and their invariant mass distribution is plotted. A clear peak is visible, indicating the presence of a particle resonance decaying into two muons. The goal is to determine the mass of the resonance as well as its production rate. The following issues may arise for this fit task:

- *Parametrisation:* So far it was always assumed that the principal shape of the pdf as a function of  $x$  (in our example  $x$  is the muon-pair invariant mass) is known.



**Figure 2.14** Fit of a signal resonance on top of background for an invariant dimuon mass distribution. (a) Fit using the Poisson likelihood. (b) Fit using the Neyman  $\chi^2$ .

However, for mass-peak fits this is often not the case. For small statistics, and if the detector resolution effects dominate over the intrinsic width of the resonance, it is often sufficient to parameterise the signal by a single Gaussian function. However, there are many cases where – in order to model the detector resolution correctly – one needs an admixture of several Gaussians with different widths or other functions with larger tails than a Gaussian. A more accurate modelling of the shape can be obtained from Monte Carlo simulations if they describe the detector performance well enough.

The background under the peak is often dominated by a so-called *combinatorial background* which consists of particles not originating from the decay of a single resonance and falsely taken as the resonance decay products. This background usually produces a smooth mass spectrum. However, Monte Carlo simulations often fail to model its normalisation and shape<sup>14)</sup>. For this reason, it is common usage to describe the background using *phenomenological parameterisations* such as polynomials or exponential functions with enough flexibility to model the background shape reasonably well.

- *Fit range:* Usually the range over which a mass spectrum (containing signal and background) is fitted can be chosen arbitrarily, at least to some extent. Ideally the range should encompass the full signal-peak region as well as lower and upper *sidebands* so that both the signal and the background contributions can be well determined. The sidebands should be at least a few times larger than the peak region since only then will the statistical uncertainty of the estimated background contribution in the peak range become small. However, for most cases, there are also good reasons to choose a not too wide fit range. A simple background parameterisation might only describe the background over a limited region but could fail outside. One could compensate this by adding further parameters to the function (e.g. by increasing the order of the fit polynomial). However, this is often not worthwhile since the increased flexibility of the fit function might help to describe the background far away from the peak region but does not improve the fit in the peak region.
- *Binned versus unbinned fits:* For our example we want to use a binned fit. The bin size should be chosen such that no relevant information on the signal is lost. One is usually on the safe side if it is about or less than half the detector resolution, as is the case for the binning chosen for the histograms in Figure 2.14.
- *Choosing the estimator – Poisson MLE versus least-squares:* As discussed above the ideal estimator for fitting signal rates is the Poisson maximum-likelihood estimate based on (2.89). For our example we use for comparison also the least-squares method based on the Neyman  $\chi^2$  (see (2.91)).

The data shown in Figure 2.14 were simulated using Poisson statistics for each bin of the  $m_{\mu^+\mu^-}$  distribution with expectation values following a Gaussian distribution for the signal and a second-order polynomial for the background. The same shape functions were also used for the fits which are shown as smooth curves in

<sup>14)</sup> Monte Carlo simulations are usually tuned to model the signal processes.

the figure. The fits were performed using the `MINUIT` [5] program. In Figure 2.14a the Poisson likelihood is used, and in Figure 2.14b the Neyman  $\chi^2$  (the default option in `MINUIT`). The fitted curve using the  $\chi^2$  method is slightly lower, which, as we know from (2.93), is a biased result. Otherwise, the fitted parameters and their estimated uncertainties, which are also listed in the plots, are very similar. The uncertainties are based on (2.22) (evaluated by calling ‘HESSE’ in `MINUIT`). Further practical considerations for the mass-peak fit with `MINUIT` (or other programs) are:

- *Choosing proper starting values:* A mass-peak fit is highly non-linear in the fit parameters, in particular for the mass and the width of the signal. The fit might therefore not converge to the correct solution if improper start values are provided (this is a purely numerical problem). Use can also be made of available a priori information on the values of the resonance mass and/or the peak width: this prior knowledge can be used to define starting values or to fix certain parameters. It can also be helpful to first only fit the background parameters to the sidebands, then fix them and fit only the signal parameters and finally fit all parameters together.
- *Checking the quality of the fit results:* The `MINUIT` program also provides the value of the Neyman  $\chi^2$ . If the statistics in the bins is not too small, the ratio of this  $\chi^2$  over the number of degrees of freedom (given by the number of bins minus the number of fit parameters) should be about unity for a good fit. In any case one should check by eye the agreement of the fit with the data for the whole fit range. In the case where these checks give a negative result one can try to optimise the parameterisation by making it more flexible – for example by adding a second, broader Gaussian for the signal to describe tails of the detector resolution and/or by increasing the order of the fit polynomial. However, one should stop adding further parameters if the minimum  $\chi^2$  does not improve significantly, that is a  $\chi^2$  decrease by about 1 for adding one parameter indicates that this parameter is not really needed, while a decrease by 10, which corresponds to a  $3\sigma$  effect, is significant.

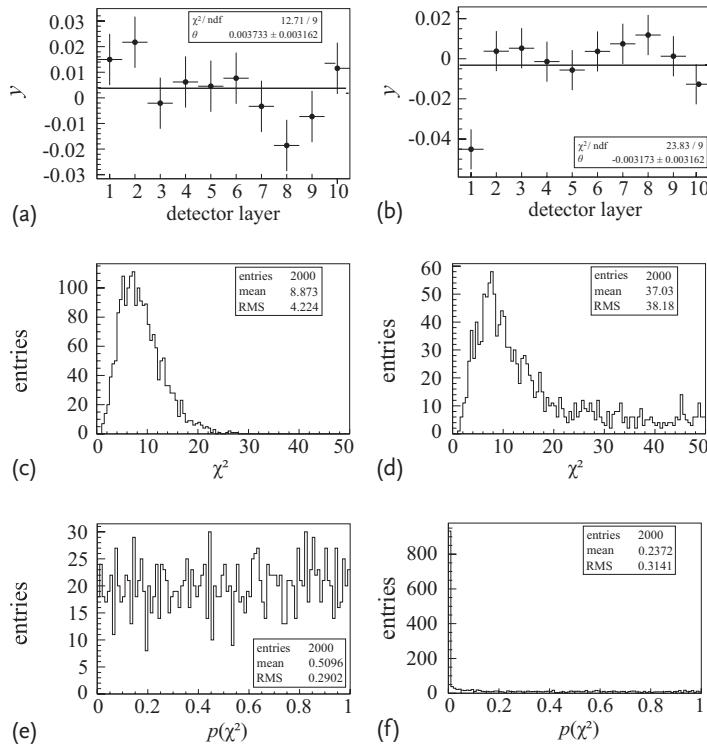
### 2.5.5

#### **Special Topic: Averaging Data with Inconsistencies**

This section addresses one fit problem which is frequently encountered in high energy physics: the averaging of different measurements with some apparent inconsistencies. For this problem (and also for the case of a bad agreement of the data and the fit function) the  $\chi^2_{\text{min}}$  value can be a very powerful tool to spot inconsistencies. This is illustrated in the following with the example of fitting a constant function to ten data points. A real-life example would be the fit of the average vertical position of a particle flying horizontally and traversing ten detector layers (neglecting any scattering in the layers). Ideally the position measurements (called ‘hits’ in the following) follow Gaussian distributions around the true track position.

Figure 2.15a shows a fit to simulated data for this analysis. In this example all hits agree within one or two standard deviations with the fitted position, and the obtained  $\chi^2_{\min}$  per number of degrees of freedom  $\chi^2_{\min}/\text{ndf} = 12.7/9$  is reasonable. In Figure 2.15c, a  $\chi^2_{\min}$  distribution is shown which was obtained by repeating 2000 times the simulation of ten data points and the subsequent fit. It follows very closely the expected  $\chi^2$  distribution function for nine degrees of freedom. In Figure 2.15e the distribution of  $P(\chi^2_{\min})$  of the track fits is plotted. The  $P(\chi^2_{\min})$  is defined as the integral of the  $\chi^2$  distribution function from the observed value  $\chi^2_{\min}$  to infinity, that is it is the probability to observe a  $\chi^2_{\min}$  value at least as large as the current one and is also often called the  $\chi^2_{\min}$  fit probability (see (3.21)). A flat distribution is observed. This follows trivially from the fact that the cumulative function of any probability density function is flatly distributed in the interval [0, 1].

Real-life detectors typically suffer from inefficiencies and noise that can create fake hits at random positions. Figure 2.15b shows a fit to simulated data where signal hits are replaced, with a 10% probability, by noise hits. The latter are generated according to a Gaussian distribution around the true track position with ten times

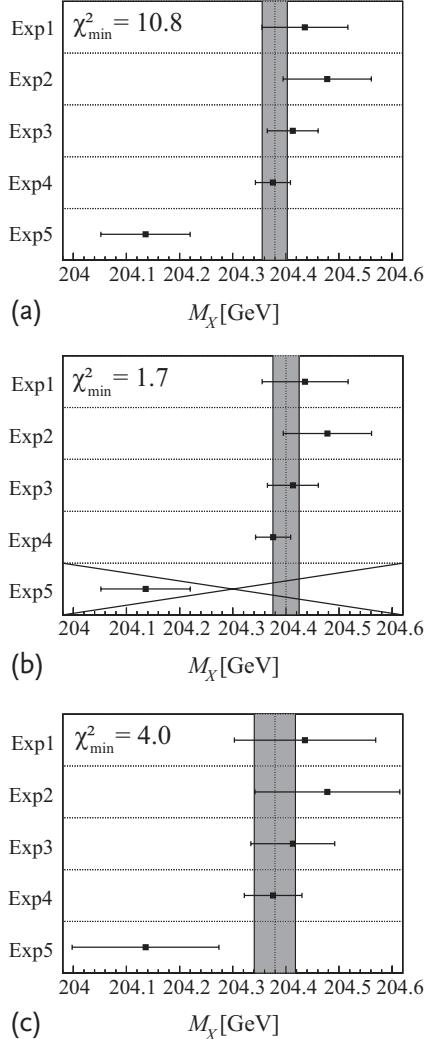


**Figure 2.15** Track fits of the average vertical position of a horizontally flying particle to ten detector hits. (a,c,e) All hits are good signal hits. (b,d,f) With a 10% random fraction of noise hits. (a,b) One exemplary track fit. (c,d)  $\chi^2_{\min}$  and (e,f)  $P(\chi^2_{\min})$  distributions, obtained from 2000 repeated simulations of ten detector hits and subsequent track fits.

larger standard deviation than for a signal hit. In the fit example shown, the first hit is obviously such a noise hit. This *outlier hit* leads to a large  $\chi^2_{\min}/\text{ndf} = 23.7/9$ . The  $\chi^2_{\min}$  distribution in Figure 2.15d, again obtained from 2000 simulations and fits, exhibits a tail to much larger values than expected for a genuine  $\chi^2$  distribution with nine degrees of freedom. However, the most striking difference compared to the case of good track fits is the  $P(\chi^2_{\min})$  distribution shown in Figure 2.15f. It exhibits a large spike at zero caused by simulated data with at least one noise hit. In this case the  $\chi^2_{\min}$  is typically so large that the probability to observe a larger value of the  $\chi^2$  distribution function is very close to zero. The distribution also features a flat tail which is caused by simulations where – by chance – all ten hits are good signal hits. The noise hits not only lead to large values of  $\chi^2_{\min}$ , but also to a significant deterioration of the true variance of the estimated track parameter (not shown). In this context it is important that the *squared residuals* of data points and fit function enter the  $\chi^2$ , and thus any outlier hit will have a large effect on the fitted curve.

In the case of a small pollution from noise hits, one suitable remedy is to simply reject track fits where the  $\chi^2_{\min}$  probability is below some given small value, for example 1%. The resulting acceptance loss might be tolerable, and the retained sample can be of good quality. Another possibility is to actively reject outlier hits in the fit. This can be done in an iterative way. First, one fits to all the hits, then one looks for the hit with the largest contribution to  $\chi^2_{\min}$  and repeats the fit without this hit. One iteratively removes further hits until a reasonable  $\chi^2_{\min}$  value is obtained. This method works well if there are enough good signal hits available to unambiguously identify comparatively few outlier hits. It exploits the redundancy of the information from the good hits. However, the method does not work if the noise hit fraction exceeds some critical value and/or there are too few detector layers. Outlier rejection belongs to the general topic of *pattern recognition*; an extensive discussion can be found for example in [14].

**World average values** The determination of world-average values, as for example performed by the Particle Data Group (PDG) [15], is an important application of averaging measurements with potential inconsistencies. In Figure 2.16, the hypothetical case of the averaging of five independent mass measurements of a particle X is shown. In Figure 2.16a the five measurements are shown with their uncertainties. The vertical line and the surrounding shaded band represent the obtained standard weighted average (see (2.18)) and the  $\pm 1\sigma$  uncertainty regions. The first four measurements are obviously in good agreement with each other, while the fifth result is about two standard deviations away. This is reflected by the rather large  $\chi^2_{\min} = 10.8$  for four degrees of freedom, corresponding to a  $\chi^2_{\min}$  fit probability  $P(\chi^2_{\min}) = 0.029$ . In this situation it is tempting to disregard the fifth measurement. The results of the corresponding fit to the remaining four data points is shown in Figure 2.16b. The shaded band now provides a better description of the first four points, and both the  $\chi^2_{\min}$  value of 1.7 and the fit probability of 0.64 indicate a good fit quality. However, unless there is a clearly identified source of problem for the fifth measurement this procedure is not at all justifiable. It could be as well that the other



**Figure 2.16** World-average value determination of the mass of a particle  $X$  from five measurements. The vertical lines and shaded bands indicate the estimated average values and the  $\pm 1\sigma$  regions. (a) The standard

weighted average. (b) Disregarding the fifth measurement. (c) Using the PDG prescription to upscale the input measurement uncertainties so that  $\chi^2_{\min} = \text{ndf}$ .

four experiments have some bias or that the observed spread of results is just due to some unlucky statistical fluctuation.

The PDG has adopted a set of prescriptions [15] for the standard weighted averaging procedure, two of which are reported here. If  $\chi^2_{\min}/\text{ndf}$  is less than or equal to 1, the standard weighted average is taken as the result. If  $\chi^2_{\min}/\text{ndf}$  is greater than 1, but not greatly so, all the uncertainties of the input measurements are scaled up

by a factor  $s = \sqrt{\chi^2_{\min}/\text{ndf}}$ , and the weighted average is re-determined. This yields, by construction, a reasonable  $\chi^2_{\min} = \text{ndf}$ . The central average result remains unchanged but its standard deviation is increased by the factor  $s$ . The method is applied for the results shown in Figure 2.16c. In the example the scale factor is  $s = 1.64$ .

The justification for this method is simple: it relies on the assumption that a large  $\chi^2_{\min}$  value indicates underestimated measurement uncertainties<sup>15)</sup> and that this should be attributed in a democratic way to all the contributing input measurements. The method is not at all statistically rigorous, but it provides rather conservative uncertainties.

## 2.6 Bayesian Parameter Estimation

Bayesian parameter estimation can be applied to all the examples discussed in this chapter. The obtained results will be similar or even identical to the ones from the frequentist methods that were used so far, however their interpretation is different. In Bayesian statistics, the posterior probability density function for  $\boldsymbol{\theta}$  given  $\mathbf{x}$  encodes the knowledge of the parameters  $\boldsymbol{\theta}$  given the observation  $\mathbf{x}$ . The posterior density function can be written using Bayes' theorem as

$$p(\boldsymbol{\theta}; \mathbf{x}) = \frac{L(\mathbf{x}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{x}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad (2.94)$$

where  $L(\mathbf{x}; \boldsymbol{\theta})$  is the likelihood function and  $\pi(\boldsymbol{\theta})$  is the prior probability density function. The prior represents the current state of knowledge or belief about the parameters  $\boldsymbol{\theta}$  before we have analysed the data. The denominator is a constant required to normalise the posterior probability density.

When one is interested in only one parameter  $\theta_j$  or in a subset of parameters, the posterior function is marginalised by integrating out all the other parameters  $\theta_{j \neq i}$ :

$$p(\theta_j; \mathbf{x}) = \int p(\boldsymbol{\theta}; \mathbf{x})d\boldsymbol{\theta}_{k \neq j} = \frac{\int L(\mathbf{x}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}_{k \neq j}}{\int L(\mathbf{x}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (2.95)$$

The posterior density contains all inference information about the parameter  $\boldsymbol{\theta}$ . It is often convenient and practical to summarise the posterior by reporting a best estimate for the parameter  $\boldsymbol{\theta}$  with its uncertainty or to report a credible interval. A popular choice for best estimate of the parameter is to use the *mode* of the posterior, that is the position where it has its maximum value. This has the advantage that it coincides with the maximum-likelihood estimate when the prior function  $\pi(\boldsymbol{\theta})$  is uniform in  $\boldsymbol{\theta}$ , that is when it is a constant function. Alternatively one can use as best estimate the mean or the median of the posterior function. The median has

15) Since for most measurements the statistical uncertainties are well known, the main suspects are undetected or underestimated systematic effects.

the advantage that it is normally a more robust estimate, less sensitive to the tails of the distribution and also invariant under certain parameter transformations.

For reporting the uncertainty on the parameter, a possibility is to quote the standard deviation (or variance) of the posterior. As has been done for the likelihood function in the case of the MLE method, one can report an approximate interval by expanding the logarithm of the posterior function around its maximum<sup>16)</sup> and by using the fact that  $\partial p(\boldsymbol{\theta}; \mathbf{x})/\partial \boldsymbol{\theta} = 0$  for  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ , since  $\hat{\boldsymbol{\theta}}$  is the maximum (mode) of the posterior function. The posterior function can then be approximated around its maximum as a Gaussian function

$$p(\boldsymbol{\theta}; \mathbf{x}) \approx A \cdot \exp \left[ \frac{1}{2} \left( \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right)^T \left. \frac{\partial^2 \ln p}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \left( \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right) \right], \quad (2.96)$$

where  $A$  is a constant. One can then quote the uncertainty on the estimated parameter by using the second derivatives of the logarithm of the posterior function. In the one-dimensional case one has:

$$\sigma = \left[ - \left. \frac{\partial^2 \ln p(\theta; \mathbf{x})}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right]^{-1/2}. \quad (2.97)$$

However, it is in general better to report a Bayesian confidence interval (credible interval) for the parameter  $\theta$  or a credible region in the multi-dimensional parameter space. This holds true particularly if the posterior distribution is highly non-Gaussian or even has several local maxima of comparable size. For the one-parameter case, a credible interval at the level  $1 - \alpha$  (e.g. 68% for a  $1\sigma$  interval) is the interval defined as

$$P(\theta_{\text{low}} < \theta < \theta_{\text{up}}) = \int_{\theta_{\text{low}}}^{\theta_{\text{up}}} p(\theta; \mathbf{x}) d\theta = 1 - \alpha. \quad (2.98)$$

This definition is clearly not unique, and many possible intervals are possible, depending on how  $\theta_{\text{low}}$  or  $\theta_{\text{up}}$  are chosen. The most common type of interval used is the shortest interval, that is the one which is obtained using the additional condition that the distance  $\theta_{\text{up}} - \theta_{\text{low}}$  is the smallest of all possible intervals. The shortest interval, in the case of unimodal (i.e. single-maximum) posterior functions, will always contain the mode of the posterior. For this reason, it is common to use the shortest interval if the mode is quoted as the best estimate. Alternatively, if we report the median of the posterior as the best estimate, it is more consistent to use a central interval, namely the one for which  $P(-\infty < \theta < \theta_{\text{low}}) = P(\theta_{\text{up}} < \theta < \infty) = \alpha/2$ .

Using the median as best estimate or quoting a central interval makes sense only when the posterior function is one-dimensional, that is for analyses where only one parameter of interest is present. In the multi-dimensional case, contour

<sup>16)</sup> This expansion is *not* applicable if the maximum of the posterior function lies on the boundary of the set of allowed parameter values.

regions which have the smallest hypervolume (area in the case of two dimensions) are typically used to summarise the posterior function.

In the case where a parameter has a limited domain, for example for parameters representing cross sections or signal yields that cannot become negative, the mode of the posterior can coincide with the lower (or upper) boundary of a parameter. In this case, one typically reports an upper (or lower) limit. In Bayesian statistics, a parameter with a limited range is naturally described by restricting the integration domain of the posterior function, which is equivalent to assuming that the prior function is zero outside the parameter domain.

A common criticism of Bayesian statistics is its subjectivity or, more precisely, the fact that there are different possible choices for choosing the prior density function for the parameters. For an extensive review of the selection of priors see for example [16]. One should choose a distribution which best represents the current belief about the unknown parameter. For example, if a previous measurement of the parameter has been performed, it is natural to use as prior function the posterior obtained from the previous measurement. When nothing is known about the parameter, a common choice is to use a uniform prior. However, uniform priors have several problems. First of all they are improper: if the range of the parameter is infinite, they cannot be normalised. However, this is not a problem if the integral of the denominator of (2.94) converges after multiplication by the likelihood. A more difficult problem with uniform prior functions is caused by parameter transformations. If one uses, for example, a prior uniform in  $\theta$ , a prior defined in  $\psi = g(\theta)$  is no longer uniform in  $\psi$ . Since  $\pi(\psi)$  is not constant, the mode of the posterior  $p_\psi(\psi; \mathbf{x})$  will not occur at  $\psi = g(\hat{\theta})$ , where  $\hat{\theta}$  denotes the mode of the posterior  $p_\theta(\theta; \mathbf{x})$ . This destroys the invariance of the Bayesian estimate of the parameter  $\theta$  under a parameter transformation. In contrast, the maximum-likelihood estimate is invariant. To solve this problem, in cases where nothing is known about a parameter, Jeffreys [17] has proposed the use of a special class of prior functions, called *objective priors*, which are obtained directly from the Fisher information matrix (see Section 1.5.3.2 and (4.63)).

Bayesian statistics is discussed in more detail in Section 4.4 in terms of confidence-interval estimation. An example for Bayesian fitting is also given in Section 10.4. Useful practical tools for Bayesian parameter estimation are the `roostats` package [18] of `ROOT` or the `BAT` program [19].

## 2.7 Exercises

### Exercise 2.1 Efficiencies and averages

A radioactive source is completely surrounded by a spherical detector. The detector consists of two hemispherical counters, each covering half ( $2\pi$ ) of the solid angle. The first counter measures with 99% efficiency and the other with 4% efficiency.

The observed numbers of counts in these detectors in one minute are  $99 \pm 9$  and  $4 \pm 2$ . What is the total decay rate of the source in one minute and its error?

- Correct the observed rates in the counters for their respective efficiencies and add them up.
- Estimate the total decay rate separately in the two counters by correcting also for their geometrical acceptance (assuming isotropic decay) and determine the weighted average of the two estimates.

### Exercise 2.2 Weighted average and $\chi^2$

Show the following equivalence (2.56) for the  $\chi^2$  of the weighted average of *two* measurements:

$$\chi^2 = \frac{(\gamma_1 - \theta)^2}{\sigma_1^2} + \frac{(\gamma_2 - \theta)^2}{\sigma_2^2} = \frac{(\gamma_1 - \gamma_2)^2}{\sigma_1^2 + \sigma_2^2} + \frac{(\theta - \hat{\theta})^2}{\sigma_{\hat{\theta}}^2}, \quad (2.99)$$

where  $\hat{\theta}$  denotes the minimum  $\chi^2$  estimate and  $\sigma_{\hat{\theta}}$  its uncertainty.

### Exercise 2.3 Unbinned fits 1

Ten events of the type  $e^+ e^- \rightarrow \mu^+ \mu^-$  are observed. The measured values of  $\cos \theta$  (where  $\theta$  is the scattering angle) are  $-0.5, -0.25, -0.1, -0.05, 0.0, 0.04, 0.11, 0.14, 0.24, 0.6$ .

- Assuming the scattering angle distribution to be  $(1 + \lambda \cos \theta)$ , obtain  $\lambda$  and its error using the maximum-likelihood method.
- Is the assumed theoretical distribution compatible with the measurement?
- Answer the same questions when the measured values are  $0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95$ .

### Exercise 2.4 Unbinned fits 2

A total of  $N$  radioactive decays are observed at times  $t_1, t_2, \dots, t_n$ .

- Assuming that the radioactive decay rate follows  $\exp(-\lambda t)$ , determine the decay constant  $\lambda$  and its error using the maximum-likelihood method.
- If the detector efficiency behaves like  $\exp(-\nu t)$ , that is it is 1 at  $t = 0$  and then decreases with time, how should we obtain  $\lambda$  and its error?
- The resolution of the measured decay times is given by a Gaussian function with a standard deviation  $\sigma$ . How does the estimated error on  $\lambda$  change?

**Exercise 2.5 Correlated fit parameters**

Suppose that a likelihood function with two variables is given by a Gaussian distribution

$$G(\mathbf{x}) = \frac{\sqrt{\det(\mathbf{C}^{-1})}}{2\pi} \cdot \exp(-\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x}),$$

where  $\mathbf{x}$  is a column vector

$$\mathbf{x} = \begin{pmatrix} x_1 - \lambda_1 \\ x_2 - \lambda_2 \end{pmatrix},$$

$\mathbf{C}$  corresponds to a covariance matrix

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and  $\det(\mathbf{C})$  denotes the determinant of  $\mathbf{C}$ . Show that the one standard deviation error of the expectation value of  $x_1$  without any restriction on  $x_2$  is given by  $\sigma_1$ . Determine the expectation value of  $x_1$  and its standard deviation when  $x_2$  is fixed to  $x_2 = \lambda'_2$ .

**Exercise 2.6 Evaluation of the information**

The decay time distributions for  $B^0$  and  $\bar{B}^0 \rightarrow J/\psi K_S^0$  are given by

$$\begin{aligned} N_{B^0 \rightarrow J/\psi K_S^0}(t) &\propto e^{-t}[1 + \lambda \sin(0.7t)], \\ N_{\bar{B}^0 \rightarrow J/\psi K_S^0}(t) &\propto e^{-t}[1 - \lambda \sin(0.7t)], \end{aligned}$$

where  $\lambda$  is the parameter we would like to measure. Determine the expected standard deviation of  $\lambda$  using the information (see (2.3) and (2.4)). Assume a value  $\lambda = 0.3$  and consider the following three cases:

- a) 500 decays  $B^0 \rightarrow J/\psi K_S^0$ .
- b) 500 decays  $\bar{B}^0 \rightarrow J/\psi K_S^0$ .
- c) 250 decays  $B^0 \rightarrow J/\psi K_S^0$  and 250 decays  $\bar{B}^0 \rightarrow J/\psi K_S^0$ .
- d) Do all three calculations give the same result? Why – or why not?

**Exercise 2.7 Unbinned maximum-likelihood estimate versus least squares**

A set of 40 radioactive decays were observed at various times  $t$  given in the following list:

{4.99, 4.87, 2.59, 3.04, 3.39, 6.20, 10.61, 7.64, 3.92, 5.33, 4.85, 2.39, 4.16, 6.74, 3.53, 5.86, 5.41, 26.25, 4.40, 10.79, 7.08, 2.86, 33.92, 3.03, 0.98, 5.63, 4.89, 2.26, 10.49, 6.51, 7.36, 2.13, 6.45, 2.29, 21.15, 4.07, 4.34, 5.38, 7.69, 4.93}.

- a) Assuming an exponential decay, that is  $N(t) \propto \exp(-\lambda t)$ , obtain the decay constant  $\lambda$  and its error using the maximum-likelihood method. Is the hypothesis of an exponential decay compatible with the data?
- b) The data can be grouped as in the following table.

Decay time interval [min]	Number of decays
0 to 5	21
5 to 10	13
10 to 15	3

Determine the decay constant  $\lambda$  and its error using the least-squares method. Compare the results with the values from the first part of the exercise.

### Exercise 2.8 Best-fit parameters

We obtained the following set of coordinate measurements:  $(x, y) = (1., 1.0 \pm 0.1), (2., 1.3 \pm 0.1), (3., 0.9 \pm 0.3), (4., 1.8 \pm 0.1), (5., 1.2 \pm 0.5), (6., 2.9 \pm 0.2)$ . By using the least-squares method, obtain the values and errors of the free parameters for the following three assumptions on the fit function  $y = f(x)$ :

- a)  $y = a$  (free parameter  $a$ ).
- b)  $y = ax + b$  (free parameters  $a, b$ ).
- c)  $y = ax^2 + bx + c$  (free parameters  $a, b, c$ ).

Which assumption fits the measurements best?

### References

- 1 Barlow, R.J. (1989) *Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons.
- 2 James, F. (2006) *Statistical Methods in Experimental Physics*, World Scientific.
- 3 Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Messenger Math.*, **41**, 155.
- 4 Aldrich, J. (1997) R.A. Fisher and the making of maximum likelihood 1912–1922. *Stat. Sci.*, **12**(3), 162.
- 5 James, F. and Roos, M. (1975) Minuit – A system for function minimization and analysis of the parameter errors and correlations. *Comput. Phys. Commun.*, **10**, 343.
- 6 Nocedal, J. and Wright, S.J. (2006) *Numerical Optimization*, 2nd edn, Springer Series in Operations Research and Financial Engineering, Springer.
- 7 Antcheva, I. et al. (2009) ROOT – A C++ framework for petabyte data storage, statistical analysis and visualization, *Comput. Phys. Commun.*, **180**, 2499.
- 8 Orear, J. (1958) Notes on statistics for physicists. <http://ned.ipac.caltech.edu/level5/Sept01/Orear/frames.html> (last accessed 16 Feb. 2013).
- 9 Lyons, L., Gibaut, D., and Clifford, P. (1988) How to combine correlated estimates of a single physical quantity. *Nucl. Instrum. Methods A*, **270**, 110.

- 10** Karimaki, V. (1991) Effective circle fitting for particle trajectories. *Nucl. Instrum. Methods A*, **305**, 187.
- 11** Neyman, J. (1949) Contribution to the theory of the  $\chi^2$  test. *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, p. 29.
- 12** Barlow, R.J. (1990) Extended maximum likelihood. *Nucl. Instrum. Methods A*, **297**, 496.
- 13** Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.*, **50**, 157.
- 14** Bock, R.K., Grote, H., Notz, D., and Regele, M. (2000) Data analysis techniques for high-energy physics experiments. *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.*, **11**.
- 15** Nakamura, K. et al. (2010) Review of particle physics. *J. Phys. G*, **37**, 075021.
- 16** Kass, R.E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.*, **91**, 1343.
- 17** Jeffreys, H. (1966) *Theory of Probability*, Oxford University Press.
- 18** Moneta, L., et al. (2010) The RooStats Project, Proc. ACAT2010 workshop. arXiv:1009.1003.
- 19** Caldwell, A., Kollar, D., and Kröninger, K. (2009) BAT: The Bayesian Analysis Toolkit. *Comput. Phys. Commun.*, **180**, 2197.

### 3

## Hypothesis Testing

*Grégory Schott*

There are two important classes of statistical inference performed in high energy physics: estimating and constraining parameters (a topic covered in Chapters 2 and 4) and testing one or multiple hypotheses. *Hypothesis testing* is a tool used for decision making and for drawing conclusions based on an acquired set of measurements. We are often aiming at determining if a dataset is consistent with a given hypothesis in order to check the validity of this hypothesis and to eventually disprove it; for example, one could test the hypothesis that the observed data originate only from Standard Model background processes when searching for indications of new physics. Some of the statistical concepts introduced in this chapter are related to those applied in classification as discussed in Chapter 5.

Hypothesis tests require a clear definition of the hypothesis (or hypotheses) to be tested and a number of other key ingredients that will be introduced in Section 3.1. A standard procedure to follow will be given in Section 3.1.5. In Sections 3.2 through 3.6, the statistical elements entering the tests will be further discussed, and an overview of the Bayesian approach to hypothesis testing will be given in Section 3.7. *Goodness-of-fit tests* are a branch of hypothesis tests where one wants to quantify how well a set of measurements agrees with a given hypothesis; this topic will be presented in Section 3.8.

### 3.1

#### Basic Concepts

##### 3.1.1

#### Statistical Hypotheses

A hypothesis test starts with requiring an unambiguous formulation of the hypothesis being tested. The *null hypothesis*, denoted  $H_0$ , is the hypothesis subject to the test and usually corresponds to the hypothesis considered to be true by default; in common language the null hypothesis could be defined as ‘the plain boring stuff that, by default, we expect to be true’. In the case of searches for new physics, this

could for example be the Standard Model of particle physics.

While one tests the null hypothesis and draws conclusions on it, it might be helpful to also define a complementary *alternative hypothesis*  $H_1$  (or multiple ones) that differs from  $H_0$ . Alternative hypotheses serve as useful comparisons to  $H_0$  and can be used to optimise hypothesis tests as will be discussed in Sections 3.2 and 3.3. Depending on the statement one wishes to test, the role of null and alternative hypotheses may get switched.

Some examples of hypotheses are:

- a) Studying a reconstructed object in an event, for example a particle, hypothesise that it is a muon. Here, more than one alternative hypothesis may be considered – for example that it is instead an electron or a proton.
- b) Studying the measured mass distribution of a set of events, suppose that these observations are the outcome of Standard Model processes exclusively.
- c) Studying the same measured mass distribution, conjecture the additional presence of an exotic particle not predicted by the Standard Model but by an alternative theory.
- d) Having two measurements, assert they are not correlated with each other.

In high energy physics, one frequently uses as null hypothesis  $H_0$  the assumption that only Standard Model processes contribute to the measurements (as for the second instance in the list above). This hypothesis is also referred to as the *background-only hypothesis*. It is complemented with an alternative hypothesis  $H_1$  in which additional new physics signal processes contribute (as for the third instance in the list). This is referred to as the *signal-plus-background hypothesis*.

Hypotheses can be either simple or composite: in a *simple hypothesis*, the expected distribution of the data can be entirely determined, that is there is no free parameter. In contrast to this, a *composite hypothesis* is based on an ensemble, or family, of simple hypotheses – which may be related by a continuous transformation of a parameter in the model. In the third example given above, the hypothesis is composite if the cross section for the production of the hypothesised exotic particle is not fixed and the simple hypotheses in the ensemble correspond to different cross-section values.

Once  $H_0$  and, eventually,  $H_1$  are defined, one would like to see how they manifest themselves in the measurements. Then the hypothesis testing could address the following questions:

- Can the null hypothesis be rejected based on the experimental measurements?
- Are the measurements compatible with the null hypothesis?

### 3.1.2

#### Test Statistic

One key ingredient in *frequentist hypothesis testing* is the definition of a *test statistic*. Let us assume the data are a set of  $N$  measurements  $\mathbf{x} = (x_1, \dots, x_N)$ . In this

chapter the  $x_i$  are assumed to be scalar quantities, but they could also be vectors. A test statistic  $t(\mathbf{x})$  is a variable constructed from the measurements alone. In hypothesis testing one defines and uses a test statistic in order to determine the level of agreement of a hypothesis with the observation. There is a relative freedom in the choice of the function  $t(\mathbf{x})$ , and a discussion of different choices can be found in Section 3.2.

While the observables may be described by their joint probability density function (pdf) under hypothesis  $H$ ,  $f(\mathbf{x}|H)$ , there is also a pdf for the test statistic,  $g(t|H)$ . In order to perform the hypothesis test, it will be necessary to determine  $g(t|H_0)$  (see Section 3.4).

Although we will restrain  $t$  to being a scalar function, it may also be a vector – in which case its pdf and the critical region (defined in Section 3.1.3 below) will also be multi-dimensional. However, the hypothesis-test formalism described below for the one-dimensional case would still apply.

### Example 3.1 Counting analysis

Let us consider a simple example: assume that in a counting analysis, an optimised selection is applied to the data so that the number of expected background events  $\nu_b = 1.3$  is reasonably small compared to the expected number of signal events  $\nu_s = 2$ . The number  $N$  of events passing the selection criteria is then obtained, and this variable is used as test statistic  $t = N$ . Under the background-only hypothesis,  $H_0$ , the function  $g(N|H_0)$  follows a Poisson distribution with mean  $\nu_b = 1.3$ , and under the signal-plus-background hypothesis,  $H_1$ ,  $g(N|H_1)$  follows a Poisson distribution with mean  $\nu_s + \nu_b = 3.3$ . The corresponding distributions are shown in Figure 3.1.

#### 3.1.3 Critical Region

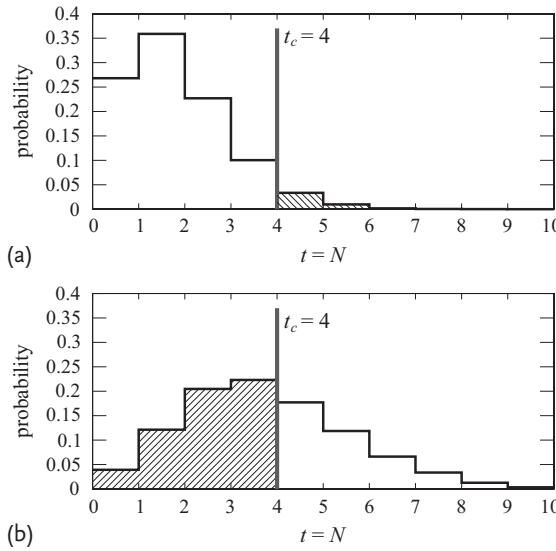
The decision on the hypothesis is based on the observed value of  $t$  and on the definition of a *critical region*. If the measured value of  $t$  lies within the critical region the hypothesis  $H_0$  is rejected; if  $t$  lies in the *acceptance region*, we failed to reject  $H_0$ .<sup>1)</sup>

For a one-tailed test the critical region is defined by a cut value  $t_c$  of the test statistic. If, under an alternative hypothesis  $H_1$ ,  $t$  tends to have larger values than under  $H_0$ , the critical region corresponds to the domain with  $t \geq t_c$  (see Figure 3.1). Conversely, for two-tailed tests, the critical region is disjoint on both sides of the test statistic. Here, we will restrict ourselves to one-tailed tests.

The probability for  $H_0$  to be rejected while  $H_0$  is true is

$$\int_{t_c}^{+\infty} g(t|H_0) dt = \alpha . \quad (3.1)$$

1) This is in general not the same as saying: ‘We accept the  $H_0$  hypothesis.’



**Figure 3.1** Probability distribution for a null (a) and an alternative hypothesis (b), which both follow Poisson distributions with mean values of 1.3 and 3.3, respectively. The number  $N$  of observed events is used as the test

statistic. The critical region  $t \geq t_c$  is defined by the boundary  $t_c = 4$ . The hashed histogram areas represent in (a) the size  $\alpha$  and in (b) the quantity  $\beta$  which are both explained in Section 3.1.3.

The quantity  $\alpha$  is called the *size of the test*, or *significance level*; it depends on the value of  $t_c$ . Note that, although related, it is conceptually different from the confidence level of a frequentist interval (see Section 4.3) or an observed significance (see Section 3.5). The latter corresponds to a *p-value* or a *Z-value*, and is determined from the observation *after* the experiment has been performed, while the size  $\alpha$  of the test is decided beforehand and independently of the observations.

If an alternative hypothesis  $H_1$  is specified, the probability to reject it although it is true is

$$\int_{-\infty}^{t_c} g(t|H_1) dt = \beta . \quad (3.2)$$

The quantity  $1 - \beta$  is called the *power of the test* (relative to the alternative hypothesis  $H_1$ ). The usefulness of the test resides in its ability to discriminate the null hypothesis against alternative hypotheses: the more  $H_0$  and  $H_1$  are separated, the smaller is  $\beta$ . The critical region needs to be adjusted depending on the analysis and the type of test we want to perform; this aspect is discussed in Section 3.3.

### Example 3.1 Counting analysis (continued)

Let us return to the counting-analysis example above. The size of the test  $\alpha$  needs to be decided before looking at the data: here, we choose  $\alpha = 0.05$ . Since the test

statistic distribution is discrete, this exact test size cannot be obtained; choosing  $t_c = 4$  results in a more conservative test size  $\alpha = 0.046$ . The background-only hypothesis is then rejected if at least four events are observed; otherwise it cannot be rejected.

Another consideration based on this example relates to the power of the test. The default event selection criteria result in a power  $1 - \beta = 0.412$ . However, a looser event selection may have yielded  $v'_b = 2.52$  and  $v'_s = 2.3$ . Then, choosing  $t_c = 6$  results in a size  $\alpha = 0.043$ , a similar value as above, but the power of the test drops to  $1 - \beta = 0.328$ ; this is clearly a less powerful test.

Although the test and its conclusions usually only concern the null hypothesis, the examination of possible alternative hypotheses is also important for example for the choice of the critical region or the choice of a test statistic that provides a large power  $1 - \beta$ .

### 3.1.4

#### Type I and Type II Errors

There are two kinds of incorrect conclusion one can draw. They are commonly called *type I* and *type II errors*:

- *Type I error* (also known as *error of the first kind*): The null hypothesis has been rejected while it was actually true; this will, by construction, happen with a probability equal to  $\alpha$ , the size of the test as defined in (3.1).
- *Type II error* (also known as *error of the second kind*): One has failed to reject the null hypothesis while it was actually false. If a single alternative hypothesis is provided, and assuming that only one of the two hypotheses is true, the probability of that type of error is  $\beta$ , as defined in (3.2).

Table 3.1 summarises the different possible scenarios and their probabilities of occurrence. The rates of both types of error are correlated since they depend on the choice made for the definition of the critical region. Reducing one type of error usually increases the other type, and a balance needs to be found (see Section 3.3).

**Table 3.1** Type I and type II errors and their occurrence probabilities.

	$H_0$ is true	$H_0$ is false
Rejected $H_0$	Type I error ( $\alpha$ )	Correct decision ( $1 - \beta$ )
Did not reject $H_0$	Correct decision ( $1 - \alpha$ )	Type II error ( $\beta$ )

## 3.1.5

**Summary: the Testing Process**

The testing process may be decomposed as follows:

1. Define the null hypothesis  $H_0$  and possibly also some alternative hypotheses.
2. Select a test statistic  $t$ . The appropriate or optimal choice of a test statistic may depend on the specifics of the analysis.
3. Determine the expected distribution of  $t$  for the null hypothesis,  $g(t|H_0)$ .
4. Define the size  $\alpha$  of the test taking into account the cost of both type I and type II errors and obtain the critical region.
5. Determine the observed value of  $t$  from the measured data sample.
6. Check if the observed value of  $t$  lies in the critical region and make a decision:
  - if  $t$  is within the critical region, reject the null hypothesis;
  - otherwise, conclude that there is not enough evidence to reject the null hypothesis.

In order to avoid introducing a bias in the statistical inference procedure, it is important to determine both the test statistic and the critical region before performing the measurement, that is before the data are looked at.

It is also important to note that hypothesis tests proceed through *proof by contradiction*. While we may be able to reject a hypothesis, it may not otherwise be concluded that this hypothesis ‘was accepted’ by the test and that it is considered true since there may be other hypotheses that could have led to the same data. What is valid instead is the statement that we failed to reject the hypothesis. Furthermore, the rejection of the null hypothesis is not a proof of the truth of the alternative hypothesis (especially if we do not know the whole set of possible alternative hypotheses). Finally, a rejected hypothesis is only rejected at a given confidence level and may still be valid with a probability  $\alpha$  (type I error). Furthermore, if some source of uncertainty was not accounted for in the hypothesis test (see Section 3.5.2) or other mistakes were made, the conclusion – to reject a hypothesis or not – may be wrong.

## 3.2

**Choosing the Test Statistic**

A good test statistic results in a clear separation of the distributions of  $t$  for the null and for the alternative hypotheses. For this reason one needs to have a clear idea about which alternative hypotheses can be expected. A test statistic is called *sufficient* if, for a given hypothesis, there exists no other test statistic that provides additional relevant information on the model (or on parameters of the model).

In *counting analyses*,  $t$  is chosen as the number of events. Also event observables may be taken as the test statistic, for example the reconstructed mass or the trans-

verse momentum of a particle; another common case corresponds to using a  $\chi^2$  variable for goodness-of-fit tests (see Section 3.8.1).

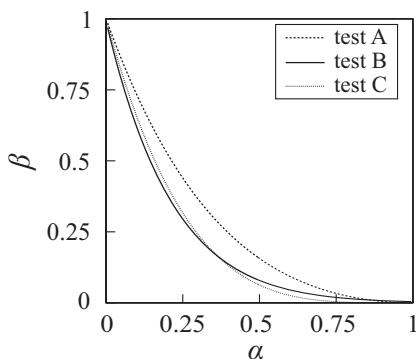
When choosing a test statistic one would like to avoid having a *biased test* which happens if  $1 - \beta \leq \alpha$ , that is if the probability to accept  $H_0$  is greater when  $H_1$  is true than when  $H_0$  is true. Among the other desirable features of a test is *consistency*, that is that the power tends to unity as the number of observations tends to infinity. An ideal test statistic obeys these conditions and for a given  $\alpha$  provides the best power, that is the smallest  $\beta$ :

- In the case of simple hypotheses, the *Neyman–Pearson lemma* states that the likelihood ratio is an optimal choice:

$$Q(\mathbf{x}) = \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)}. \quad (3.3)$$

The best critical region consists of points  $\mathbf{x}$  satisfying  $Q(\mathbf{x}) > c_\alpha$ , where  $c_\alpha$  is a constant adjusted to reach the size  $\alpha$ . In case the likelihood ratio is not numerically accessible, multivariate classifiers such as neural networks or Fisher discriminants (see Chapter 5) can be used to define the test statistic.

- For composite hypotheses, assuming that the set of alternative hypotheses belongs to one continuous family, we may write this set of alternative hypotheses as a function of a parameter  $\theta$ ; the set of alternative hypotheses is denoted as  $H_1(\theta)$ . The parameter  $\theta$  could for example be the production cross section of a new particle. In that case, the power of the test,  $\beta(\theta)$ , will also be a function of  $\theta$  and the best test statistic to use (the one maximising the power function) may also depend on  $\theta$ . The choice of  $t$  may not be clear if there is not one single test statistic which is *uniformly the most powerful* for all values of  $\theta$ . A reasonable choice may be one that provides a small  $\beta(\theta)$  in a reasonable range of  $\theta$  values.



**Figure 3.2** Exemplary curves of  $\beta$  versus  $\alpha$  for three tests. Test A (dashed line) is clearly the worst one since for a given size  $\alpha$  it has a lower power  $1 - \beta$  while test B (solid line) and test C (dotted line) lie close to each other. Neither

of the two tests can be considered better than the other: while test B is more powerful for small  $\alpha$ , test C is more powerful for large  $\alpha$ . As  $\alpha$  is usually chosen to be small one would prefer test B in most cases.

Another approach to identify the test statistic to use is to study the  $\alpha$ -versus- $\beta$  curves which retain some of the symmetry between  $\alpha$  and  $\beta$  [1]. An example is illustrated in Figure 3.2.

### 3.3

#### Choice of the Critical Region

As already mentioned, usually a trade-off between type I and type II errors has to be made. The size  $\alpha$  controls the rate of occurrence of errors of the first kind. The error of the second kind depends on the separation of the pdfs for both hypotheses. Defining the appropriate critical region will depend on what *cost* is assigned to both types of error. This cost function may be quantifiable (e.g. a financial cost) or not (e.g. a cost on your reputation for falsely discovering a hypothetical signal). In some cases one may want to drastically reduce type I errors, while in some other cases one may want to keep type II errors under control.

In high energy physics, when dealing with discovery,  $\alpha$  is commonly taken to be very small for a discovery (see Section 3.5.1) and, for an exclusion, sizes of 5% or 10% are used to reach 95% or 90% confidence levels. When using hypothesis tests to define a rough selection of events (such as defining trigger filters), the critical-region criteria may be much more relaxed to avoid missing interesting signal events even if we still keep large amounts of background events.

### 3.4

#### Determining Test Statistic Distributions

In the case of large data samples (i.e. large  $N$ ) drawn randomly, the distribution of means of the observables  $\mathbf{x}$  for several such samples would follow a normal distribution even if the measured quantities  $\mathbf{x}$  are not normally distributed. This is a consequence of the *central limit theorem*. Thus, in the case of large  $N$ , the pdf for the test statistic  $t$  under a hypothesis  $H$ ,  $g(t|H)$ , can be easily determined. In the general case, however, the function  $g(t|H)$  is non-trivial and numerical methods are used to determine it, for example, through Monte Carlo simulations. In that case, the set of observables  $\mathbf{x}$  is generated using the pdf function  $f(\mathbf{x}|H)$  introduced in Section 3.1.2, and the value of the test statistic  $t(\mathbf{x})$  is computed for the set. These steps are repeated to accumulate statistics for the  $g(t|H)$  distribution until a sufficiently large number of random Monte Carlo experiments have been drawn.

For computing very small  $p$ -values with reasonable precision, a large number of MC iterations is required. In that case the tail of the distribution of  $g(t|H)$  is most important. The procedure above may be improved by resorting to techniques such as *importance sampling* which concentrates on generating Monte Carlo datasets that lie in those tails.

Using the likelihood-ratio test statistic (see (3.3)) has another advantage: namely, that we can more effectively resort to *asymptotic distributions* (i.e. distributions with large statistics) as an approximation of the true distribution. It can be shown that  $-2 \ln Q$  follows a central (i.e. standard)  $\chi^2(r)$  distribution if the  $H_0$  hypothesis is true and a non-central  $\chi^2(r)$  distribution<sup>2)</sup> if  $H_1$  is true, where  $r$  is the number of parameters which are fitted for in the maximisation of  $L(\mathbf{x}|H_1)$  but fixed in  $L(\mathbf{x}|H_0)$  (see [1], pp. 271–273). The usage of log-likelihood-ratio test statistics and their asymptotic behaviour is also discussed in [2].

### 3.5 *p*-Values

The *p-value* is a function that quantifies how often, if an experiment was repeated many times, one would obtain data as far away (or more) from the null hypothesis as the observed data, assuming the null hypothesis to be true. The *p-value* is a measurement of the *observed level of significance*. It is a function of the data and therefore a random variable; it should not be confused with the size of the test  $\alpha$ , which is a predefined constant. The size does not depend on  $t_{\text{obs}}$ , the observed value of the test statistic, but on  $t_c$ , the cut defining the critical region.

The *p*-values  $p_0$  and  $p_1$  for the two hypotheses  $H_0$  and  $H_1$  are given by the probabilities to find  $t$  values that are equal to or greater than the value  $t_{\text{obs}}$  observed in the present experiment:

$$p_0 = \int_{t_{\text{obs}}}^{+\infty} g(t|H_0) dt , \quad (3.4)$$

$$p_1 = \int_{t_{\text{obs}}}^{+\infty} g(t|H_1) dt . \quad (3.5)$$

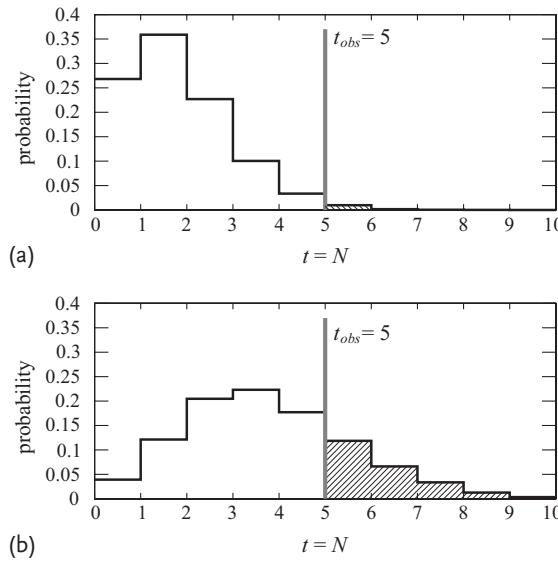
#### Example 3.1 Counting analysis (continued)

The observed number of events in the example given above is  $t_{\text{obs}} = 5$ , as indicated in Figure 3.3. The resulting *p*-values for the two hypotheses are  $p_0 = 0.012$  and  $p_1 = 0.235$ .

Although *p*-values are not confidence levels, they are sometimes abusively replaced by the notations  $\text{CL}_b = 1 - p_0$  and  $\text{CL}_{s+b} = p_1$  in analyses where the null hypothesis corresponds to the presence of background processes only and the alternative hypothesis to the presence of both signal and background processes.

It is important to note that a *p*-value cannot be regarded as the ‘probability of the hypothesis to be true given the data’, as this would require Bayes’ theorem

2) A non-central  $\chi^2$  distribution is obtained by summing the squares of independent Gaussian random variables that have unit variance but non-zero means.



**Figure 3.3** Probability distribution for (a) a null and (b) an alternative hypothesis, which both follow Poisson distributions with mean values of 1.3 and 3.3, respectively. The number  $N$  of observed events is used as the test statistic. The distributions determined under

both hypotheses are compared to the value observed in data,  $t_{\text{obs}}$ . The  $p$ -values are computed by integrating the distributions for values of  $t$  equal to or greater than  $t_{\text{obs}} = 5$  and shown as hashed histogram areas.

and a prior on the hypothesis; it is instead the probability to obtain data at least as incompatible with the hypothesis as the present data if the hypothesis considered is actually true.

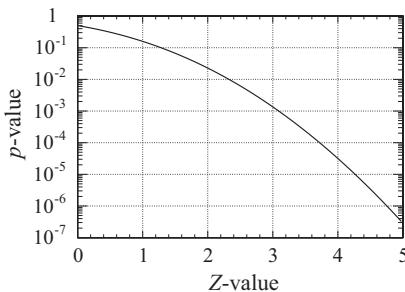
If the hypothesis for which the  $p$ -value is computed is true, then for a continuous test statistic the  $p$ -value is uniformly distributed between 0 and 1. A small  $p$ -value is an indication for an inconsistency with the hypothesis. The null hypothesis is rejected in hypothesis tests if  $p_0$  is smaller than a given confidence level  $\alpha$ . Typically  $p$ -values can be used for hypothesis tests with a fixed size  $\alpha$ , for significance tests and for goodness-of-fit tests.

### 3.5.1 Significance Levels

One way to express  $p$ -values is to redefine them as *significance levels*, also referred to as *Z-values*:

$$Z = \Phi^{-1}(1 - p), \quad (3.6)$$

where  $\Phi^{-1}$  is the cumulative distribution function of the unit Gaussian. The  $Z$ -value corresponds to the one-sided tail probability of a Gaussian distribution. For example, a  $p$ -value of 0.05 corresponds to  $Z = 1.64$ . Conversely, the  $p$ -value may be



**Figure 3.4** One-sided tail probability (*p*-value) as a function of the number of standard deviations (*Z*-value) for a unit Gaussian probability density function.

obtained from a *Z*-value via

$$p = \frac{1 - \text{erf}(Z/\sqrt{2})}{2}, \quad (3.7)$$

where  $\text{erf}(x)$  is the error function given in (1.17). Figure 3.4 presents the *p*-value as function of the *Z*-value.

When using a likelihood ratio  $Q = L_{SB}/L_B$  as the test statistic and assuming that it follows a Gaussian distribution, the significance can be estimated as

$$Z = \sqrt{-2 \ln Q}. \quad (3.8)$$

For counting experiments following Poisson statistics, the significance is sometimes approximated by  $Z = S/\sqrt{S + B}$ , that is the ratio of the signal strength over the uncertainty of the total number of events. In this equation,  $S$  and  $B$  are the observed or expected signal excess and background level depending on which of the observed significance or expected significance is wanted. An alternative is given by  $Z = 2(\sqrt{S + B} - \sqrt{B})$ .<sup>3)</sup> However, a more accurate approach is to insert the Poisson likelihood terms in (3.8) to obtain

$$Z = \sqrt{2(S + B) \ln(1 + S/B) - 2S}. \quad (3.9)$$

In a counting analysis, when a simple significance estimator is needed – for example in order to optimise an event selection – the use of (3.9) is recommended. To obtain an exact significance one should rather evaluate the *p*-value from the exact test statistic distribution.

If a *p*-value reaches  $2.87 \cdot 10^{-7}$ , the *Z*-value is 5 and one says that the significance level is ‘ $5\sigma$ ’.<sup>4)</sup> A significance level of  $5\sigma$  is conventionally used in high energy physics to claim a discovery, and  $3\sigma$  for an evidence when testing against the null hypothesis.

- 3) When adding a Gaussian uncertainty  $\Delta B$  on the background estimate  $B$ , the significance can be approximated as  $2(\sqrt{S + B} - \sqrt{B}) \cdot B/(B + \Delta B^2)$ ; some alternatives are discussed in [3]. A study of how the different approximations compare is given in appendix A of [4].
- 4) Some authors use a different convention where a  $5\sigma$  significance corresponds to  $p = 5.7 \cdot 10^{-7}$  (two-sided tail probability).

## 3.5.2

**Inclusion of Systematic Uncertainties**

In practice, the likelihood function or a likelihood ratio is often used as test statistic  $t$ . Systematic effects often lead to a larger overlap of test statistic distributions of the null and alternative hypotheses; the hypotheses become less separated and therefore the power of the test is decreased, corresponding to larger  $p$ -values on average.

A suitable approach to account for systematic uncertainties is to incorporate them in the likelihood function as *nuisance parameters*. For example, in a counting experiment the number of expected background events  $\nu_b$  can be such a nuisance parameter in a likelihood function  $L(N; \nu_s + \nu_b)$ . In the following we are staying with this example and describe two common approaches to include systematic effects: the Bayesian-inspired marginalisation approach [5] and the method of profiling the likelihood function.

**Bayesian-inspired approach** Here a *prior probability* distribution  $\pi(\nu_b)$  is assumed for the nuisance parameter  $\nu_b$ . The parameter  $\nu_b$  is eliminated by integrating the product of the pdf and the prior over the nuisance parameter space, leading to the marginalised likelihood:

$$L_m(N; \nu_s) = \int L(N; \nu_s + \nu_b) \pi(\nu_b) d\nu_b . \quad (3.10)$$

While this integration may in principle be done analytically, in practice the relevant functions are more complex and the distribution of  $L_m$  is determined with a Monte Carlo method: the nuisance parameter is sampled randomly according to its prior distribution. For each sampling point, the corresponding pdf term is calculated. Since the hypothesis-testing framework is a frequentist one but accounting for nuisance parameters follows a Bayesian ansatz, this is often called a ‘hybrid’ approach. A commonly used test statistic in this case is obtained from the ratio of marginalised likelihoods:

$$-2 \ln \frac{L_m(\mathbf{x}|H_1)}{L_m(\mathbf{x}|H_0)} = -2 \ln \frac{L_m(N; \nu_s)}{L_m(N; \nu_s = 0)} . \quad (3.11)$$

**Profile likelihood approach** Here the likelihood function is profiled, with nuisance parameters treated as free fit parameters. However, one needs to have either another measurement or an external constraint on the nuisance parameters. In the counting-experiment example, one possibility is to constrain the parameter  $\nu_b$  from a sideband measurement of  $N'$  events (called *auxiliary measurement*):

$$L(N'; \nu_b) = \frac{(\tau \nu_b)^{N'}}{N'!} e^{-\tau \nu_b} , \quad (3.12)$$

where  $\tau$  is a scale factor assumed to be known. The full model is written as the product

$$L(N, N'; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^N}{N!} e^{-(\nu_s + \nu_b)} \frac{(\tau \nu_b)^{N'}}{N'!} e^{-\tau \nu_b}. \quad (3.13)$$

This likelihood function is then profiled with respect to the nuisance parameter  $\nu_b$ . The profiled likelihood function  $L_p(N, N'; \nu_s, \hat{\nu}_b)$  is obtained for each value of  $\nu_s$  by maximising the likelihood (3.13) against the parameter  $\nu_b$  (which converges to  $\hat{\nu}_b$ ), thereby removing it from the problem.

It is also possible to incorporate an uncertainty on the parameter  $\tau$  by multiplying its prior pdf to (3.13). For example, assuming it is estimated to be  $\tau_0$  with a Gaussian uncertainty  $\sigma_\tau$ , the factor to multiply (3.13) with is  $1/(\sqrt{2\pi}\sigma_\tau)e^{-(\tau-\tau_0)^2/(2\sigma_\tau^2)}$ . The resulting likelihood function has now to be profiled with respect to *both*  $\nu_b$  and  $\tau$ .

In the case of the profiled approach, a common test statistic is the ratio

$$-2 \ln \frac{L_p(\mathbf{x}|H_1)}{L_p(\mathbf{x}|H_0)} = \frac{L_p(N, N'; \nu_s, \hat{\nu}_b, \hat{\tau})}{L_p(N, N'; \nu_s = 0, \hat{\nu}_b, \hat{\tau})}. \quad (3.14)$$

Note that the nuisance parameters will converge to different values in the likelihood functions in the numerator and denominator. For a discussion of some alternative profiled likelihood ratios see [2, 6].

While the two methods described above are the most widely used, other approaches for handling systematic uncertainties exist, especially at the level of *p*-values, see for example [7].

### 3.5.3 Combining Tests

In order to improve the sensitivity of an analysis, one may want to combine multiple tests into a single, more powerful one. However, this is not a trivial task and there is no straightforward answer to a question such as: ‘Having measured  $4\sigma$  and  $3\sigma$  significances, for a certain process in two channels, what is the combined significance?’ One needs to take into account the correlations between the two measurements and the exact shape of the underlying test statistic distributions.

In general, it is preferable to combine the primary measurements and to perform the full statistical inference based on a test statistic (such as a combined likelihood function) that takes into account the known correlations between the analyses. There are a number of methods that can be used to approximate the *p*-value of the combined test. A simple and intuitive approach was formulated by Fisher [8]. Let us assume two independent tests yielding *p*-values  $p_1^{\text{obs}}$  and  $p_2^{\text{obs}}$ . The possible set of outcomes  $(p_1, p_2)$  is ranked in Fisher’s method according to their product  $p_1 p_2$ . Under the condition that  $p_1$  and  $p_2$  are uniform under the null hypothesis,

the probability that  $p_1 p_2$  is smaller than  $p_1^{\text{obs}} p_2^{\text{obs}}$  is

$$\begin{aligned} P(p_1 p_2 \leq p_1^{\text{obs}} p_2^{\text{obs}}) &= \int_0^1 dp_1 \int_0^{p_1^{\text{obs}} p_2^{\text{obs}} / p_1} dp_2 p_1 p_2 \\ &= p_1^{\text{obs}} p_2^{\text{obs}} [1 - \ln(p_1^{\text{obs}} p_2^{\text{obs}})] , \end{aligned} \quad (3.15)$$

which is larger than  $p_1^{\text{obs}} p_2^{\text{obs}}$ . When  $n$  tests are considered, one computes the product of  $p$ -values

$$\gamma = \prod_{i=1}^n p_i .$$

If the null hypothesis is valid, the quantity  $-2 \ln \gamma$  follows a  $\chi^2$  distribution with  $2n$  degrees of freedom.

Note that when  $p$ -values are obtained from discrete test statistic distributions, the requirement that the  $p_i$  are uniform is not fulfilled. Even though Fisher's method might still provide a reasonable approximation, the true confidence level is effectively smaller than the one obtained from the formula above.

Many alternatives to Fisher's approach exist. They work more or less well for certain classes of problems; a short review can be found in [9]. Specifically, there are approaches that combine weighted tests [10, 11] which allow, in particular, to deal with cases where the sensitivity of a high-quality measurement may get affected when combining with a low-quality one.

### 3.5.4

#### Look-Elsewhere Effect

An instance of a *look-elsewhere effect* (LEE) occurs when searching for a signal peak with unknown location in an invariant-mass spectrum. It is important to distinguish the probability to find a fluctuation in some *particular* location from the probability to find such a fluctuation *anywhere*. The former is called the *local significance* whereas the latter is referred to as the *global significance*. The reason why these two values are different is known as the LEE. Obviously, the global significance is smaller than the largest local one. Both significances (i.e. probabilities) are related to each other by a multiplicative factor proportional to the number of *independent* search regions (i.e. the spectrum width divided by the resolution). It has been shown [12] that this factor is also a linear function of the observed local significance. The factor is best estimated using Monte Carlo methods, but asymptotic approximations were developed for practical purposes [12]. Sometimes, only this bias – introduced when searching in a spectrum for a peaking signal whose location is unknown – is considered as LEE. However, other types of look-elsewhere effects also exist. Let us discuss some examples.

Consider 20 hypothesis tests performed each with a size  $\alpha = 0.05$ . Under the null hypothesis, one expects on average one of the 20 tests to be rejected (type I

error) and 19 to pass the test. If indeed exactly one experiment fails the test and if this is the only result reported, the conclusion is biased: the confidence level for this result would not be 95% as claimed. While an LEE correction may be used to rectify this probability, it would be better to report all the tests performed and all results obtained.

Additional significance biases might be introduced in measurements by the analyst himself, for example in a combination of measurements. If one considers only the channels which yield a significant signal and drops the others, this biases the total significance towards large values. One could possibly use only the channels that are relevant for the model being tested or the most sensitive channels; the decision of which channels to use should, however, be taken *before* looking at the data. A similar situation occurs, for example, when interpreting the set of deviations between observations and Standard Model predictions obtained through global electroweak fits. And there is typically a great excitement if one of these fluctuations is larger than 2 or  $3\sigma$ .

Related types of biases may be introduced if the data are used to optimise the event selection criteria – or to choose the discriminating variables, to choose a Monte Carlo generator, and so on – based on the observed signal significance, for example. It may be difficult to assess the magnitude of such biases. Whenever possible, it is best to define the analysis procedure without any use of the data, for example using Monte Carlo samples instead. If not possible otherwise, one could use in the tuning and optimisation phase of an analysis only a small fraction of the data in order to reduce some of the possible biases.

Often, biases are unintentionally introduced and may remain unnoticed. Whenever possible, performing a *blind analysis* (see Sections 8.5.2.2, 8.5.3 and 10.6.1) is advisable to avoid the bias altogether.

### 3.6 Inversion of Hypothesis Tests

When the hypothesis under test is composite and depends on the values of free parameters, then hypothesis testing and *interval estimation* (discussed in Chapter 4) may be related.

In the following, a composite hypothesis with a single free parameter  $\theta$  is assumed. Testing in turn hypotheses with different values of  $\theta$  leads to the determination of an exclusion region. An example appears in the search for the Higgs boson where, for a given Higgs mass, the parameter  $\theta$  is taken as the signal cross section. A hypothesis can be tested for each  $\theta$  value, and the corresponding  $p$ -value is computed. Let us assume we aim for a test of size  $\alpha = 5\%$ . Each of the hypotheses that has a  $p$ -value of less than 5% can be excluded at the 95% CL. This analysis is thus answering the question of what is the region of  $\theta$  – that is of the cross section for Higgs-boson production – which can be excluded at 95% CL. The procedure is then usually repeated for a range of possible Higgs masses.

Sometimes a downward fluctuation in the number of background events can lead to a limit on the signal process in an unphysical region (e.g. a negative cross section). Then, an alternative to using  $p_1$  is to take the ratio  $CL_s = p_1/(1 - p_0)$ . This also allows one to prevent exclusion where one has little or no sensitivity to distinguish between  $H_0$  and  $H_1$ . This approach [13, 14] has been used to set limits on the Higgs mass at LEP, the Tevatron and the LHC. Other approaches are the one proposed by Feldman and Cousins [15] or the application of power-constrained limits [16].

### Example 3.2 Search for a Higgs boson

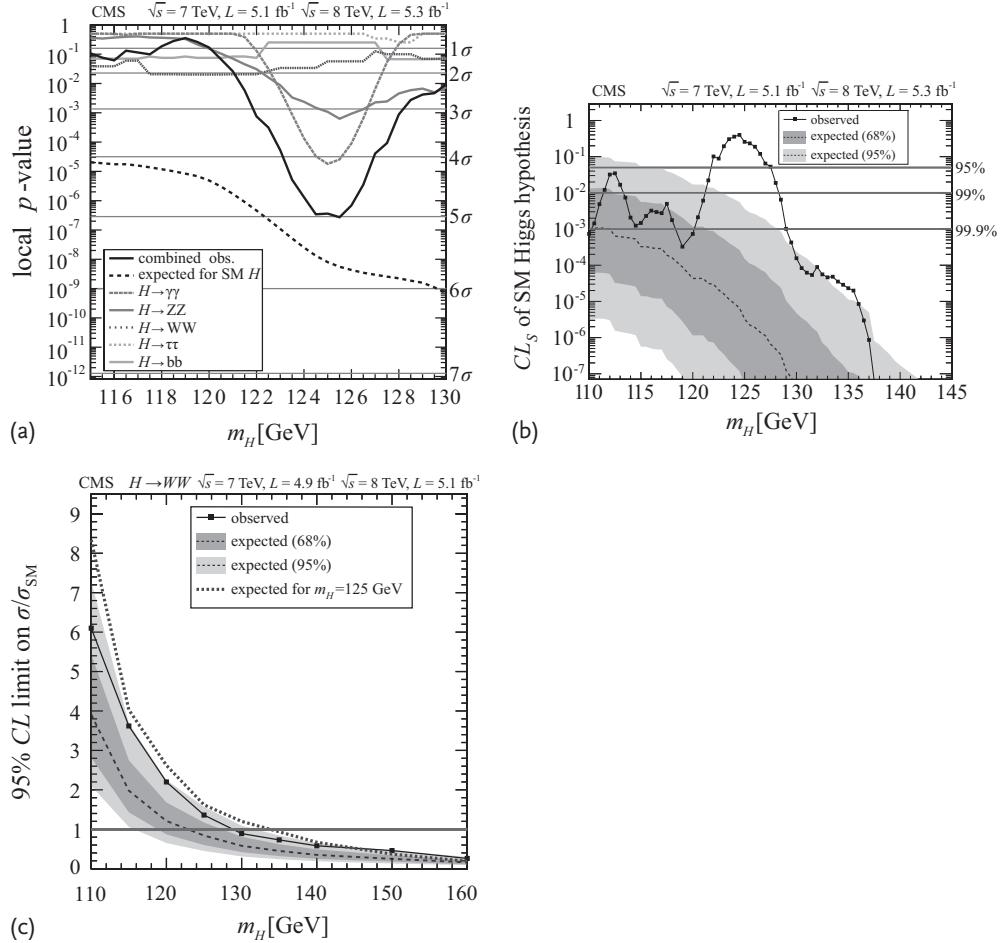
An application of hypothesis testing and test inversion is the search for the Higgs boson in the CMS experiment [17] which was performed with the statistical procedure described in [6].

Figure 3.5a shows the  $p$ -value of the background-only hypothesis (only Standard Model background processes contribute) obtained from a combination of five analyses sensitive to the presence of a Higgs boson. For comparison, the dashed curve shows the median  $p$ -value expected for the signal-plus-background hypothesis (both Standard Model background and Higgs processes contribute): pseudo-data are generated according to the signal-plus-background hypothesis and tested against the background-only hypothesis. The hypotheses are formulated for all mass values  $m_H$  and tested separately in the displayed mass range. The maximum local  $p$ -value reaches a significance of  $5.0\sigma$  around  $m_H = 125$  GeV which, after taking into account the LEE, is reduced to a global significance of  $4.5\sigma$ .

Figure 3.5b displays, for the combined analysis and as a function of the Higgs mass, the observed  $CL_s = p_1/(1 - p_0)$  values and the expected  $CL_s$  assuming the background-only hypothesis. In the  $CL_s$  approach, one excludes a hypothesis in a test of size  $\alpha = 0.05$  if the observed value of  $CL_s$  is smaller or equal to 0.05; in the present case, this happens for Higgs masses in the range 110–121.5 GeV and for masses above 128 GeV. In the intermediate mass range, where an excess with respect to the background-only hypothesis is observed, the signal-plus-background hypothesis is not excluded. Other tests of sizes  $\alpha = 0.01$  and  $\alpha = 0.001$  are also shown. The light (dark) shaded bands around the median expected  $p$ -value show intervals in which 68% (95%) of the experimental outcomes are expected if the measurements were repeated a large number of times and the background-only hypothesis was true.

Searches are sometimes used to constrain parameters, for example the ratio of the signal cross section and the predicted cross section for a Standard Model Higgs,  $\sigma/\sigma_{\text{exp}}$ . This is another example of a test inversion. Figure 3.5c shows the 95% CL upper limit obtained in the  $H \rightarrow WW$  channel. For each mass  $m_H$ , the value of the parameter  $\sigma/\sigma_{\text{exp}}$  is scanned and the corresponding hypothesis is tested until a value which results in  $CL_s = 0.05$  is found. All larger cross sections than that one are also excluded at the 95% confidence level. Cross-section ratios  $\sigma/\sigma_{\text{exp}}$  smaller than 1 are excluded for  $m_H \gtrsim 129$  GeV; this can also be understood in the sense

that the Standard Model Higgs boson (i.e.  $\sigma/\sigma_{\text{exp}} = 1$ ) is excluded at 95% CL for all hypotheses with  $m_H > 129 \text{ GeV}$ .



**Figure 3.5** Search for a Higgs boson by the CMS experiment. The plots represent as a function of the Higgs mass: (a) the  $p$ -value of the Standard Model hypothesis obtained in a combination of five analyses sensitive to the presence of a Higgs boson; (b) the observed and expected values of  $CL_s = p_1/(1 - p_0)$

for the combination of the five analyses; (c) the 95% CL upper limit on the signal cross section relative to the Higgs production cross section in the Standard Model, obtained in the  $H \rightarrow WW$  channel. Further details of (a–c) are given in the main text. (Adapted from [17].)

### 3.7

#### Bayesian Approach to Hypothesis Testing

While in the *frequentist Neyman–Pearson approach* the computed  $p$ -values cannot be interpreted as the probability of a hypothesis, it is possible in a *Bayesian approach* to assign a posterior probability  $P(H|\mathbf{x})$  to a hypothesis  $H$  given the data  $\mathbf{x}$ . On the basis of Bayes' theorem, this *posterior probability* is given as

$$P(H|\mathbf{x}) = \frac{\int L(\mathbf{x}|\theta, H)P(H)\pi(\theta)d\theta}{P(\mathbf{x})}, \quad (3.16)$$

where  $\pi(\theta)$  and  $P(H)$  are the prior density on a set of nuisance parameters  $\theta$  and the prior probability of the hypothesis<sup>5)</sup> itself, respectively, and where the denominator  $P(\mathbf{x})$  represents the sum over all possible hypotheses. On the basis of a Bayesian decision rule one can reject a model  $H$  if the posterior probability  $P(H|\mathbf{x})$  is sufficiently small.

From the ratio of the posterior probabilities of two hypotheses  $H_i$  and  $H_j$  one may construct a quantity  $B_{ij}$ ,

$$B_{ij} = \frac{\int L(\mathbf{x}|\theta_i, H_i)P(H_i)\pi(\theta_i)d\theta_i}{\int L(\mathbf{x}|\theta_j, H_j)P(H_j)\pi(\theta_j)d\theta_j}, \quad (3.17)$$

called the *Bayes factor*. In the absence of nuisance parameters and if  $P(H_i) = P(H_j)$ , the Bayes factor is simply the likelihood ratio; otherwise it is a ratio of marginalised likelihoods. More on the Bayesian approach may be found in [18].

### 3.8

#### Goodness-of-Fit Tests

*Goodness-of-fit* (GoF) tests address the question how well a data distribution is described by the functional form provided by a hypothesis. This form may be completely fixed or there can be free parameters which are fitted to the data. One difference to the hypothesis tests described above is that GoF tests only involve one hypothesis, the null hypothesis  $H_0$ . In the design of GoF tests it can, however, be useful to have an idea on the possible range of alternative hypotheses, since these might hint towards the type of discrepancy the GoF test should be particularly sensitive to.<sup>6)</sup>

The first step in GoF tests is to construct a test statistic whose value is sensitive to the level of agreement between the measured data and the null hypothesis. Let us assume in the following that a larger value indicates a worse agreement. In a second step, one computes the probability of obtaining a test statistic value at least as large as the one measured, or in other words to obtain data that are as compatible

5) Often there is no uniquely agreed prior  $P(H)$  for the hypothesis.

6) It is, however, important that the test does not become too specific and remains sensitive to a wide range of alternatives.

with the hypothesis as the measured data, or less. This probability is called the  $p$ -value, and its size specifies the agreement between data and hypothesis.

The tests presented below have the advantage of being independent of the nature of the distribution they are built upon; for example, the test is the same independently of whether the data follow a Gaussian or a linear distribution. Other tests exist that are specific to certain functional forms. In the following we discuss first tests that are applied to binned distributions and then close the chapter with a description of some unbinned tests. The latter ones are usually more powerful, but may be more complex and less general.

### 3.8.1

#### Pearson's $\chi^2$ Test

Out of all GoF tests one clearly stands out and plays an important role in the field of statistics: *Pearson's  $\chi^2$  test*. This test is usually easy to apply.<sup>7)</sup>

The method of least squares used in parameter estimation was discussed in Section 2.5.4 of this book. It is based on a  $\chi^2$  quantity that may also be used to estimate the goodness of a fit or, more generally, to compare an observed distribution with the one expected from a null hypothesis when measurements have Gaussian errors. When the measurements have Poisson errors, Pearson's  $\chi^2$  is used.

Let us assume we have a set of  $N$  measurements that have been arranged into a histogram with  $M$  bins, with the measurements in each bin being independent. One can write a sum of squared normalised deviations as

$$\chi^2 = \sum_{i=1}^M \frac{(n_i - \nu_i)^2}{V[\nu_i]} , \quad (3.18)$$

where  $n_i$  is the number of events in bin  $i$ ,  $\nu_i$  the number of events expected in this bin according to the null hypothesis, and  $V[\nu_i]$  is its variance.

In the case that the data follow a Poisson distribution<sup>8)</sup> (as is often the case in high energy physics) and the measurements  $n_i$  are not too small, then they are approximately distributed according to a Gaussian function and the statistic follows a  $\chi^2$  distribution for  $\text{ndf}$  degrees of freedom:

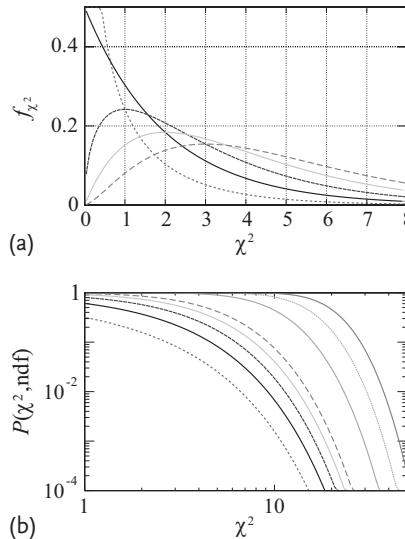
$$f_{\chi^2}(\chi^2, \text{ndf}) = \frac{1}{2^{\text{ndf}/2} \Gamma(\text{ndf}/2)} (\chi^2)^{\text{ndf}/2-1} e^{-\chi^2/2} , \quad (3.19)$$

where

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt . \quad (3.20)$$

7) Note that there is no ideal GoF test that is universally applicable and efficient; Pearson's  $\chi^2$  test also has some pitfalls to beware of and is not always the most sensitive solution. A suitable choice of GoF tests to run will depend on the specifics of a given analysis.

8) For the Poisson function  $V[\nu_i] = \nu_i$  see Section 1.3.2.



**Figure 3.6** (a)  $\chi^2$  distribution and (b)  $\chi^2$  probability (*p*-value) of the observed  $\chi^2$  for different numbers of degrees of freedom as a function of the  $\chi^2$  measured in data ((a) ndf = 1, 2, 3, 4, 5; (b) ndf = 1, 2, 3, 4, 5, 10, 15, 20).

The  $\chi^2$  distribution for ndf degrees of freedom has a mean  $E[\chi^2] = \text{ndf}$  and a variance  $V[\chi^2] = 2\text{ndf}$ ; for large ndf the function approaches a Gaussian (see Section 1.3.4.6).

The  $\chi^2$  probability measures the probability (or *p*-value) that, under the null hypothesis, a set of  $M$  measurements give a  $\chi^2$  as large as, or larger than, the one obtained:

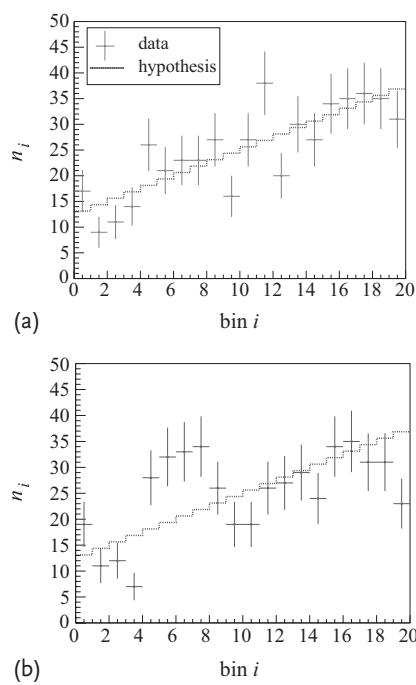
$$P(\chi^2, M) = \int_{\chi^2}^{+\infty} f_{\chi^2}(z, M) dz . \quad (3.21)$$

A low *p*-value corresponds to bad agreement of the data with the tested hypothesis. The  $\chi^2$  function (3.19) and the  $\chi^2$  probability (3.21) are represented in Figure 3.6 for various values of  $M$ .

With additional degrees of freedom in a data fit (through a minimisation of  $m$  parameters) the computed  $\chi^2$  will be smaller since it was minimised. The minimised value of the  $\chi^2$  will then follow a  $\chi^2(M - m)$  distribution.

Also, if the total number of events is fixed to the observed value  $N$ , the test only checks against the shape and not the total normalisation. The random variable to use is

$$\chi^2 = \sum_{i=1}^M \frac{(n_i - N p_i)^2}{N p_i} , \quad (3.22)$$



**Figure 3.7** Binned distribution of two exemplary datasets (data points) compared to the expected distribution from a tested hypothesis (dashed line). (a) A dataset in good agreement with the model. (b) Another dataset in bad agreement.

where  $p_i$  is the probability for an event to be measured in bin  $i$  according to the null hypothesis. In this case, the  $\chi^2$  variable is expected to follow a  $\chi^2(M - 1)$  distribution.

### Example 3.3 Binned $\chi^2$ test

Figure 3.7 shows an example distribution for  $n_i$  and  $\nu_i$ . For the 20 bins we have the constraint that  $\sum n_i = \sum \nu_i$ , and no fit is applied before the test is performed; therefore  $\text{ndf} = 19$ . Using formula (3.22), one obtains  $\chi^2 = 21.2$ , which corresponds to a  $p$ -value of 0.328 for the dataset in Figure 3.7a. Data and model are therefore in relatively good agreement according to Pearson's  $\chi^2$  test. For the dataset in Figure 3.7b, an upward fluctuation seems to appear around bins 5 to 8; one obtains  $\chi^2 = 49.0$  which corresponds to a  $p$ -value of 0.00019 – pointing to a quite poor agreement.

If the probability density function under the tested hypothesis describes the data distribution well, it should roughly agree with the measurement within the measured error.<sup>9)</sup> In this case the  $\chi^2$  is roughly equal to  $\text{ndf}$ . A too large  $\chi^2$  points

9) Actually, since these errors represent a 68% interval, it is expected that on average every third point's error bars do not include the expected value.

towards a discrepancy – something is wrong (either with the hypothesis or the data, or maybe there are unaccounted bin correlations between the measured errors). If the  $\chi^2$  is too small (i.e. too good) this can mean that the errors have been overestimated, that the data have been specially selected or that the measurements were just obtained by chance. How small a  $\chi^2$  has to be in order to be considered ‘bad’ is subjective – so it is left to you to decide. Typically  $P(\chi^2, \text{ndf}) < 0.001$  is considered bad because it is only expected in one out of a thousand cases; often  $P(\chi^2, \text{ndf}) < 0.05$  is used to define a failed test.

The reduced  $\chi^2/\text{ndf}$ , which is more easily computed than the  $p$ -value, is often used as a measure of agreement between the data and the hypothesis since the expectation value of the  $\chi^2$  distribution equals the number of degrees of freedom. One should, however, note that the  $p$ -value is more informative. Compare for example a case with  $\chi^2/\text{ndf} = 7/5$ , where the  $p$ -value is 0.22, to a case with  $\chi^2/\text{ndf} = 70/50$  where the  $p$ -value is 0.03; both have the same  $\chi^2/\text{ndf}$  ratio but the first case has a much better  $p$ -value. Together with giving this ratio one should therefore also provide either the number of degrees of freedom, ndf, or the  $p$ -value.

In order to apply the approach described in this section, the number of events in each bin needs to be large (in order to be in the regime where the Poisson distribution can be approximated with a Gaussian one); five to ten events are typically required. In case the number of events is smaller than that, one may change the total number of bins or have a variable size for different bins so that all the bins are sufficiently populated; in order to avoid biases on the test, the choice of binning to apply should be decided without any use of the measured data (e.g. based on the number of expected events with some safety margins). In cases where binning issues arise, an unbinned GoF test would usually be more powerful since some information is lost when the bin width is not small compared to the experimental resolution.

When asymptotic conditions cannot be reached, one should not use the  $\chi^2$  probability as the  $p$ -value (from (3.21)) but should compute the exact  $p$ -value. This can, for example, be done with a Monte Carlo approach that considers all possible outcomes by computing their individual probabilities before comparing them to the observed measurement. When using such a Monte Carlo approach, one may relax the requirements on the bin size.

### 3.8.2

#### Run Test

One disadvantage of Pearson’s  $\chi^2$  test is its insensitivity to the sign of the deviations. Let us consider an example with a distribution of five bins. The expected number of entries in the bins are  $(10, 10, 10, 10, 10)$  with a standard deviation of 3 for each bin. Whether the data outcomes are  $(7, 13, 13, 13, 7)$ ,  $(13, 7, 7, 7, 13)$  or  $(7, 13, 7, 13, 7)$ , the  $p$ -value computed from the  $\chi^2$  distribution for each of these three outcomes is the same. Consider now the case where over a wide region of

a spectrum the data are systematically larger than the values expected by the tested model, and smaller than expected in the other regions. In this case the  $\chi^2$  value may remain reasonable, although there is a clear indication of an incompatibility of the current model parameterisation and the data.

In general, it is useful to complement the  $\chi^2$  test with other, if possible independent GoF tests. One such test, applicable to binned distributions, is the so-called *run test* [19]. In the run test, one defines regions of consecutive bins where the data show deviations in the same direction from the expectation. Each such region is called a *run* and the number of runs is denoted as  $r$ . The number of observed runs is expected to follow a Binomial distribution. With  $N_+$  bins measured above the expectation and  $N_-$  bins below, the number of ways these bins can be arranged is  $(N_+ + N_-)!/(N_+!N_-!)$  and the average expected number of runs and the variance are

$$E[r] = 1 + \frac{2N_+N_-}{N_+ + N_-}, \quad (3.23)$$

$$V[r] = \frac{2N_+N_-(2N_+N_- - N_+ - N_-)}{(N_+ + N_-)^2(N_+ + N_- - 1)}. \quad (3.24)$$

For a distribution with more than about 20 bins, the number of runs can be reasonably approximated by a Gaussian distribution. Therefore, the  $Z$ -value can be computed as:

$$Z = \frac{r - E[r]}{\sqrt{V[r]}}. \quad (3.25)$$

#### Example 3.4 Run test

Consider the series of deviations for the distribution in Figure 3.7a:

$$(+---+++-+-+-+--++--).$$

The number of upward deviations is  $N_+ = 12$  and the number of runs is  $r = 10$ . Using the above equations we obtain  $E[r] = 10.6$  and  $V[r] = 4.35$ . Assuming the Gaussian approximation to be valid we obtain, from (3.25),  $Z = -0.288$ . The  $p$ -value for  $|Z| \geq 0.288$  is 0.773. We can conclude that the data are in very good agreement with the predictions from the tested hypothesis. For the dataset Figure 3.7b, the series of deviations is

$$(+---+++-+-+----+--).$$

The number  $r = 6$  is observed, while  $10.6 \pm 2.1$  were expected. This corresponds to a  $p$ -value of 0.0274, indicating a poor agreement of data and model expectations.

Run tests are less powerful than Pearson's  $\chi^2$  tests but still very useful. The two tests are complementary since they are weakly correlated when testing against simple hypotheses; the  $\chi^2$  test is sensitive to the size of deviations but neither to their signs nor to their ordering. In contrast, the run test is only sensitive to the signs and ordering of the deviations. The  $p$ -values of the Pearson- $\chi^2$  and the run test can be combined (e.g. using (3.15)) to form a more powerful test.

### 3.8.3

#### $\chi^2$ Test with Unbinned Measurements

It is in general possible to use one test statistic for determining the best value of the parameters and another one for measuring the discrepancy between data and prediction. If a likelihood function has been used to fit an unbinned dataset, it is still possible to perform a binned GoF test, for example a Pearson  $\chi^2$  test, and using the best fit parameters to calculate expectation values  $\hat{\nu}_i$  for the bins.

For multinomially distributed data one can construct a  $\chi^2$  variable

$$\chi_M^2 = 2 \sum_{i=1}^M n_i \ln \frac{n_i}{\hat{\nu}_i} \quad (3.26)$$

that in the large-sample limit will follow a  $\chi^2(M - m - 1)$  distribution. Here  $M$  is the number of bins and  $m$  the number of free parameters of the likelihood fit. The quantities  $n_i$  and  $\hat{\nu}_i$  denote the number of observed and expected events in bin  $i$ . The latter is obtained using the values of the fitted parameters. Note that bins with  $n_i = 0$  are dropped from the summation. Similarly, for Poisson distributed data one can define

$$\chi_p^2 = 2 \sum_{i=1}^M n_i \ln \frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i , \quad (3.27)$$

which, in the large-sample limit, is expected to follow a  $\chi^2(M - m)$  distribution.

Alternatively, one can use the test statistic from Section 3.8.1 given in (3.18) and (3.22), where the terms  $\nu_i$  and  $p_i$  are replaced by  $\hat{\nu}_i$  and  $\hat{p}_i = \hat{\nu}_i / \sum_i \hat{\nu}_i$ , respectively. These will follow (in the large-sample limit)  $\chi^2$  distributions of  $M - m$  and  $M - m - 1$  degrees of freedom, respectively.

For limited sample size a more elaborate approach is needed; for example, one may use Monte Carlo simulations to obtain the distribution of the test statistic.

### 3.8.4

#### Test Using the Maximum-Likelihood Estimate

If a *maximum-likelihood fit* with a function  $L(\mathbf{x}; \theta)$  is performed to adjust parameters of the model, the value of the likelihood function at the best fit position  $L_{\max}(\mathbf{x}; \hat{\theta})$  is sometimes used as a GoF test statistic. One may determine the distribution of  $L_{\max}$  with pseudo-experiments such as Monte Carlo simulations. For this

the parameters  $\theta$  of the null hypothesis are set to their expected true values. The obtained distribution of  $L_{\max}$  can be compared to the observed  $L_{\max}$  value from data. However, one should note that the  $L_{\max}$  distributions are often badly separated under alternative hypotheses and that therefore this approach is relatively weak. A further discussion, discouraging this method, can be found in [20].

### 3.8.5

#### Kolmogorov–Smirnov Test

In unbinned GoF tests, each event is accounted for at the exact value of the test statistic measured (and the exact probability at this point). These tests are free from the arbitrariness and information loss related to binning. Unbinned analyses also allow one to more easily accommodate scenarios where the test statistic is a vector rather than a scalar quantity. The best known unbinned GoF test is the *Kolmogorov–Smirnov test*.

The empirical cumulative distribution function  $F_N(x)$  for  $N$  measurements of an observable can be written as

$$F_N(x) = N^{-1} \sum_{i=1}^N s(x_i), \quad (3.28)$$

where  $s$  is the step function:

$$\begin{aligned} s(x_i) &= 0 \quad \text{for } x < x_i, \\ s(x_i) &= 1 \quad \text{for } x \geq x_i. \end{aligned}$$

In this test this function is compared to the cumulative distribution of the probability distribution,  $F(x)$ , under the hypothesis being tested, which is usually a smooth curve. An example is shown in Figure 3.8.

One can compute the largest absolute difference between the two curves,

$$D_N = \max |F_N(x) - F(x)| \quad \text{for all } x. \quad (3.29)$$

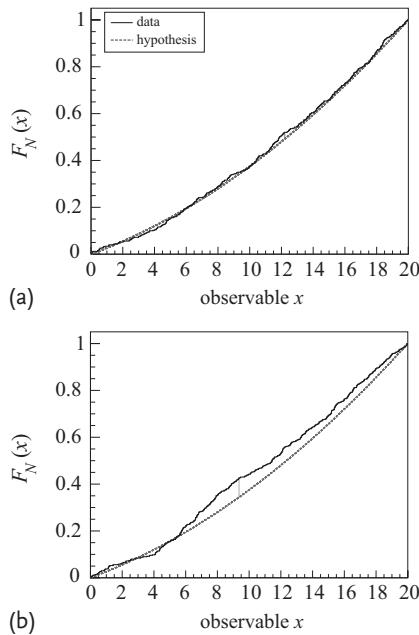
For one-sided tests, this difference can also be computed taking into account the sign of the deviations,

$$D_N^\pm = \max [\pm(F_N(x) - F(x))] \quad \text{for all } x. \quad (3.30)$$

The values of  $\sqrt{N}D_N$  and  $\sqrt{N}D_N^\pm$  that define the critical region of the Kolmogorov–Smirnov test statistic have been tabulated as a function of the  $p$ -value (see Table 3.2) for various  $N$ . For  $N \gtrsim 80$ , a  $\chi^2$  distribution with two degrees of freedom may be used to approximate the quantity  $4N(D_N^\pm)^2$ . For large  $N$  one may also use the quantity

$$2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 N D_N^2} \quad (3.31)$$

for the  $p$ -value determination from a Kolmogorov–Smirnov test.



**Figure 3.8** Cumulative functions of  $x$ . The empirical cumulative function for data (darker solid lines) is shown as well as the cumulative function from a tested hypothesis (lighter dashed lines). The same datasets as in Figure 3.7a,b are used. The data in (a) agree with the prediction while the data in (b) do not.

The largest distance  $D_N$  computed in a Kolmogorov–Smirnov test is observed in (a) at  $x = 12.2$  and in (b) at  $x = 9.4$ .

### Example 3.5 Kolmogorov–Smirnov test

Let us return to Example 3.3 (see Figure 3.7). We now use the same data, but without applying any binning. The  $F_N(x)$  and  $F_N(x)$  distributions are determined (see Figure 3.8), and  $D_N$  is calculated to be  $D_N = 0.029$  for the first dataset and  $D_N = 0.082$  for the second, corresponding to  $p$ -values (using formula (3.31)) of 0.80 and 0.0022, respectively.

**Table 3.2** Values of  $D_N$  that define the critical region in Kolmogorov–Smirnov tests as a function of the  $p$ -value and for various numbers of events  $N$ .

$p$ -value	Critical value of $D_N$			
	$N = 5$	$N = 20$	$N = 60$	$N > 100$
0.99	0.67	0.35	0.21	$1.629/\sqrt{N}$
0.98	0.63	0.33	0.19	$1.518/\sqrt{N}$
0.95	0.56	0.29	0.17	$1.358/\sqrt{N}$
0.90	0.51	0.26	0.16	$1.223/\sqrt{N}$
0.80	0.45	0.23	0.14	$1.073/\sqrt{N}$

One should note that the Kolmogorov–Smirnov test can only be applied when the distribution is fixed before the experiment is performed; if parameters of the model in the tested hypothesis are fitted to the data then this test cannot be used.

In Kolmogorov–Smirnov tests the agreement is constrained by construction to be good at the boundaries of the tested distribution, and the maximum differences  $D_N$  or  $D_N^\pm$  will most likely be located towards the center of the distribution. Therefore, this GoF test suffers from a sensitivity loss at the boundaries.

The *Anderson–Darling test* [21] is similar to the Kolmogorov–Smirnov test but it is not affected by the sensitivity loss at the boundaries. This comes, however, with a loss in generality since this test is not independent of the functional shape according to which the data are distributed.

### 3.8.6

#### **Smirnov–Cramér–von Mises Test**

The *Smirnov–Cramér–von Mises test* is an alternative to the Kolmogorov–Smirnov test. It is based on the test statistic measuring the average square difference:

$$W^2 = \int_{-\infty}^{+\infty} [F_N(x) - F(x)]^2 dF(x), \quad (3.32)$$

where  $F_N(x)$  and  $F(x)$  denote the same functions as in the previous section. Rather than using one single point where the difference  $|F_N(x) - F(x)|$  is largest, the integral of the squared difference is used.

The expectation value and variance of  $W^2$  are given by

$$\begin{aligned} E[W^2] &= \frac{1}{6N}, \\ V[W^2] &= \frac{4N-3}{180N^3}. \end{aligned}$$

Some tabulated values of the critical function  $N W^2$  as a function of the test size can be found in [22].

### 3.8.7

#### **Two-Sample Tests**

*Two-sample tests* are related to the issue of goodness-of-fit tests. They are performed when one wants to compare two samples to see how compatible they are. For example, one may want to test the hypothesis that both samples have the same mean or the same width. Since the result of such tests depends strongly on the assumptions in the hypothesis, one needs to be very clear on what is the hypothesis tested in order to interpret the result of the test.

It should be noted that there also exist general approaches similar to the two unbinned tests above; for example, replacing the  $F(x)$  term in (3.29) or in (3.32) by

the empirical cumulative distribution function of the second sample  $F_{N'}(x)$ . Also, in the binned  $\chi^2$  test, one may replace (3.18) by

$$\chi^2 = \sum_{i=1}^M \frac{(n_i - n'_i)^2}{\sigma_{n_i}^2 + \sigma_{n'_i}^2} \quad (3.33)$$

in order to compare two samples of measurements with bin entries  $n_i$  and  $n'_i$ .

### 3.9

#### Conclusion

The main concepts of hypothesis testing, typical applications in high energy physics and commonly used goodness-of-fit tests were presented in this chapter. For further reading on this topic one may consult the textbooks by Barlow [23], Cowan [24], Lyons [25] (all three at the introductory level), James [1] (more advanced) or Sivia [18] (on the Bayesian aspects).

### 3.10

#### Exercises

##### Exercise 3.1 Significance

Consider a Poisson process with  $N = 56$  observed events and an expectation value  $\nu_b = 40$ .

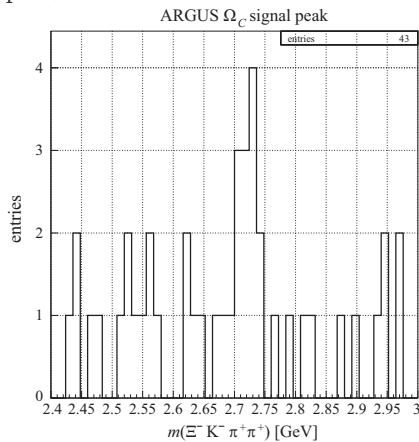
- a) Calculate the significance of the observation using the various simple significance formulae given in Section 3.5.1 and compare the results to the exact significance from the Poisson tail probability.
- b) Assume  $\nu_b$  to have a systematic Gaussian uncertainty  $\Delta\nu_b/\nu_b = 7.5\%$ ; calculate the significance taking this systematic into account and using the marginalised likelihood ratio. Compare to the formula given in footnote 3.
- c) What is the 95% CL upper-limit on  $\nu_s$  in a hypothesis test inversion without and with the systematic uncertainty? Assume the signal (we are searching for) to have an expectation value  $\nu_s = 25$ . Is it excluded with the current dataset?
- d) If the experiment accumulates additional statistics: how much more data are needed to expect a signal discovery assuming the signal theory to be valid?

##### Exercise 3.2 $\Omega_c$ peak at ARGUS

In 1992, the ARGUS  $e^+ e^-$  experiment reported the observation of the charmed and doubly strange baryon  $\Omega_c$  through its decay channel  $\Xi^- K^- \pi^+ \pi^+$  [26]. The obtained mass spectrum is shown in Figure 3.9. Try to make your own assessment of the signal and its significance.

1. *Fluctuation probability:* Under the assumption that there is only background with constant density:

- Estimate the average number of background events per mass bin (note: the histogram contains 43 entries in 50 bins);
- Define a  $\pm 2\sigma$  mass window around the peak (note: the resolution  $\sigma$  is approximately 12 MeV, the histogram bin width);
- Count the total number of candidates  $N_{\text{cand},s}$  in the  $\pm 2\sigma$  region around the peak;



**Figure 3.9** The invariant-mass spectrum used in Exercise 3.2. (Adapted from [26].)

- Estimate the number of expected background events  $\mu_b$  in this region;
- Estimate the probability for the Poisson distribution to fluctuate from  $\mu_b$  to  $N_{\text{cand},s}$  or larger values.

2. *Signal significance:* Under the signal-plus-background hypothesis try to estimate the signal and its significance:

- Estimate the number of background events per bin from the average density of events in the regions outside the peak. Estimate from this density the number of expected background events  $\mu_b$  in the  $\pm 2\sigma$  region around the peak;
- Obtain the number  $N_s = N_{\text{cand},s} - \mu_b$ , estimate an error  $\sigma_{N_s}$  and determine the signal significance  $N_s/\sigma_{N_s}$ .

### Exercise 3.3 Goodness-of-fit tests

The data measurements and expectation values for the distributions plotted in Figure 3.7 are the following:

Bin	1	2	3	4	5	6	7
Observed 1	17	9	11	14	26	21	23
Observed 2	19	11	12	7	28	32	33
Expected	13.125	14.375	15.625	16.875	18.125	19.375	20.625
Bin	8	9	10	11	12	13	14
Observed 1	23	27	16	27	38	20	30
Observed 2	34	26	19	19	26	27	29
Expected	21.875	23.125	24.375	25.625	26.875	28.125	29.375
Bin	15	16	17	18	19	20	
Observed 1	27	34	35	36	35	31	
Observed 2	24	34	35	31	31	23	
Expected	30.625	31.875	33.125	34.375	35.625	36.875	

- a) Reproduce the  $p$ -values of the Pearson  $\chi^2$  test and of the run test for each dataset.
- b) Calculate a combined  $p$ -value of both tests with use of the Fisher method.
- c) Perform a GoF test of the compatibility of the two datasets with each other, independently of the expectation values.

## References

- 1 James, F. (2006) *Statistical Methods in Experimental Physics*, 2nd edn, World Scientific.
- 2 Cowan, G. et al. (2011) Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, **71**, 1554.
- 3 Cousins, R.D., Linnemann, J.T., and Tucker, J. (2008) Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. *Nucl. Instrum. Methods A*, **595**, 480.
- 4 CMS Collab. (2007) Physics technical design report, volume II: Physics performance. *J. Phys. G*, **34**, 995.
- 5 Cousins, R.D. and Highland, V.L. (1992) Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Methods A*, **320**, 331.
- 6 ATLAS and CMS Collab., LHC Higgs Combination Group (2011) Procedure for the LHC Higgs boson search combination in Summer 2011. ATL-PHYS-PUB-2011-011; CMS NOTE-2011/005.
- 7 Demortier, L. (2007)  $p$ -values and nuisance parameters. *Proc. PHYSTAT-LHC Workshop*. CERN-2008-001.
- 8 Fisher, R.A. (1970) *Statistical Methods for Research Workers*, 14th edn, Oliver and Boyd.

- 9** Cousins, R.D. (2007) Annotated bibliography of some papers on combining significances or  $p$ -values. arXiv:0705.2209.
- 10** Good, I.J. (1955) On the weighted combination of significance tests. *J. R. Stat. Soc. Ser. B*, **17**, 264.
- 11** Janot, P. and Le Diberder, F. (1998) Optimally combined confidence limits. *Nucl. Instrum. Methods A*, **411**, 449.
- 12** Gross, E. and Vitells, O. (2010) Trial factors for the look elsewhere effect in high energy physics. *Eur. Phys. J. C*, **70**, 525.
- 13** Read, A.L. (2002) Presentation of search results: the CLs technique. *J. Phys. G: Nucl. Part. Phys.*, **28**, 2693.
- 14** Junk, T. (1999) Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Methods A*, **434**, 435.
- 15** Feldman, G.J. and Cousins, R.D. (1998) Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, **57**, 3873.
- 16** Cowan, G. *et al.* (2011) Power-constrained limits. arXiv:1105.3166.
- 17** CMS Collab. (2012) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, **716**, 30.
- 18** Sivia, D. and Skilling, J. (2006) *Data Analysis: a Bayesian Tutorial*, Oxford University Press.
- 19** Wald, A. and Wolfowitz, J. (1940) On a test whether two samples are from the same population. *Ann. Math. Stat.*, **11**, 147.
- 20** Heinrich, J. (2003) Pitfalls of goodness-of-fit from likelihood, in *Statistical Problems in Particle Physics, Astrophysics, and Cosmology* (eds L. Lyons, R. Mount, and R. Reitmeyer), *Proc. PHYSTAT-2003 Workshop*, p. 52.
- 21** Anderson, T.W. and Darling, D.A. (1952) Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *Ann. Math. Stat.*, **23**, 193.
- 22** Marshall, A.W. (1958) The small sample distribution of  $n\omega_n^2$ . *Ann. Math. Stat.*, **29**, 307.
- 23** Barlow, R.J. (1989) *Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons.
- 24** Cowan, G. (1998) *Statistical Data Analysis*, Oxford University Press.
- 25** Lyons, L. (1986) *Statistics for Nuclear and Particle Physicists*, Cambridge University Press.
- 26** ARGUS Collab., Albrecht, H. *et al.* (1992) Evidence for the production of the charmed, doubly strange baryon  $\Omega_c$  in  $e^+e^-$  annihilation. *Phys. Lett. B*, **288**, 367.

**4****Interval Estimation**

*Luc Demortier*

**4.1**  
**Introduction**

The point estimation procedures described in Chapter 2 provide very concise summaries of what the data have to say about a given parameter of interest: each summary consists of a single number, representing what is in some sense the most likely value of the parameter, given the observed data. The downside of this conciseness is that it does not come with a characterisation of the reliability of the estimate. This is what interval estimates attempt to remedy: instead of a single numerical estimate, two numerical limits are provided, plus a level of confidence about the true value of the parameter of interest lying between these limits. If there is more than one parameter of interest, intervals are replaced by multi-dimensional regions. It is also possible for an interval construction method to yield a union of disjoint intervals or regions, and this may be sensible in some contexts (see Example 4.4 below).

Not surprisingly, the correct interpretation of the confidence level of an interval or region depends strongly on the statistical paradigm one is operating in, *Bayesian* or *frequentist*. Furthermore, specification of a desired confidence level does not uniquely determine an interval construction. Other desiderata enter into play, for example interval length, behaviour under reparameterisation, effect of physical boundaries, systematic uncertainties, and so on. After briefly reviewing such interval characterisations in Section 4.2, we describe frequentist constructions in Section 4.3 and Bayesian ones in Section 4.4. Following the two sections on methodology, a graphical comparison using a problem involving a physical boundary is provided in Section 4.5. Next, the role of interval construction in search procedures is addressed in Section 4.6, in particular in relation to the issues of *coverage* and *measurement sensitivity*. Some final recommendations are given in Section 4.7.

## 4.2

### Characterisation of Interval Constructions

Confidence level is the primary characteristic of an interval construction, but its meaning is radically different in the Bayesian and frequentist approaches to statistical inference. In the Bayesian approach, the final result of a measurement is the posterior distribution of the parameter of interest, and interval estimation is one method among others for summarising the information contained in this distribution. The confidence level associated with a Bayesian interval is the integral of the posterior over that interval and is also called *credibility*. It represents the probability for the parameter of interest to lie somewhere inside the interval, given one's prior beliefs and the observed data.

Conversely, the frequentist approach does not associate probability distributions with constants of nature and therefore requires a different concept to quantify the reliability of interval estimates. This is the concept of *coverage*, which characterises how an interval construction procedure behaves over large numbers of replications of the measurement under consideration. Coverage answers the question: 'If  $N$  new datasets are collected under the same conditions as the actually observed one, and the same measurement is performed each time, what fraction of these measurements will yield a confidence interval that contains the true value of the parameter of interest, as  $N \rightarrow \infty$ ?' It should be noted that a desired coverage cannot always be achieved exactly, for example when the observable is discrete (e.g. a number of events), or when systematic uncertainties are present.

Even though the credibility and coverage interpretations of a confidence level belong to different statistical paradigms, it is often instructive to investigate the credibility of frequentist intervals and the coverage of Bayesian intervals. This is because Bayesian inferences fully condition on the observed data, whereas frequentist ones take both observed and unobserved data into account. Thus, one could question the *relevance* of a frequentist result for the data at hand, and this can be clarified by studying its posterior credibility with a well-motivated, proper Bayesian prior. Similarly, one could question the *replicability* of a Bayesian result, and this can be investigated with a well-defined ensemble of measurements (real or simulated). An interesting result in this regard is that when a proper prior is used, the prior-averaged coverage of a Bayesian interval construction equals its nominal credibility, thus guaranteeing replicability in some average sense. When a proper, *evidence-based prior* is not available, coverage may still provide useful guidance in choosing a so-called *objective prior* [1].

As already indicated, the desired confidence level of an interval estimate does not uniquely specify how to construct such an estimate. There are many possibilities, and for choosing among them it is useful to examine other interesting properties:

- *Interval length*: For a given confidence level, short intervals are more informative than long ones, at least when they cover the true value of the parameter. A frequentist concept that may be useful in this regard is that of *expected length*, which is the interval length averaged over the ensemble of all possible obser-

vations and viewed as a function of the parameter of interest. It can be shown that the expected length of a confidence interval is equal to the probability of including a false value of the parameter in the interval, integrated over all false values [2]. Since the expected length involves an ensemble average, it is not a Bayesian criterion. However, given a Bayesian posterior distribution, a popular interval construction is that known as *highest posterior density* (HPD), which yields the shortest interval of a given credibility (see Section 4.4).

- *Equivariance under parameter transformations:* When measuring a quantity such as a particle mass  $\theta$ , the result may be used by theorists to draw inferences about another quantity  $\eta = f(\theta)$ , where  $f$  can be an arbitrarily complicated function. Suppose now that we apply the *same* interval construction procedure to both parameters, obtaining  $[\theta_1, \theta_2]$  and  $[\eta_1, \eta_2]$ , respectively. It would be useful to have  $[\eta_1, \eta_2] = [f(\theta_1), f(\theta_2)]$ , but this is generally not true; for example, the shortest intervals in  $\theta$  do not necessarily map onto the shortest intervals in  $\eta$ .
- *Behaviour with respect to systematic uncertainties:* Systematic uncertainties are modelled with the help of nuisance parameters, that is, parameters that are of no direct interest to the experimenter but must be known in order to draw inferences about the parameter of interest. Examples include calibration constants, energy scales, and detection efficiencies. Nuisance parameters are constrained by auxiliary measurements or Bayesian priors, which determine the distribution of associated systematic uncertainties. Typically one expects the length of an interval for the parameter of interest to increase with the width of that distribution.
- *Effect of physical boundaries:* When the parameter space has boundaries imposed by physical constraints, some interval constructions may yield intervals that lie partially or completely in the unphysical region for some subset of observations. The physical part of these intervals is then either unreasonably narrow or empty, a highly undesirable situation. Examples where this may happen are measurements of efficiencies and acceptances, where the true value is constrained to lie between 0 and 1, and particle masses, where it must be positive. It is also possible for a parameter boundary to have special physical significance. In a search for new particles, for example, the production rate is constrained to positive values. The value zero, however, has special significance since it corresponds to the background-only hypothesis (no new particles). Whether interval estimation is the appropriate type of inference in such situations, as opposed to for example hypothesis testing, is an issue that needs to be carefully thought out.
- *Relation to point estimate:* When measuring a property of a system known to exist (e.g. the mass of the top quark), one usually reports both an interval and a point estimate, and it is desirable that the latter be contained in the former. However, intervals and point estimates provide different types of inference and there is no unique relationship between them. One can try to introduce such a relationship,<sup>1)</sup> but this does not necessarily yield optimal procedures. Conversely,

<sup>1)</sup> The Hodges–Lehmann estimator, for example, is defined as the limit of an interval construction as the confidence level goes to zero [3].

there are some natural associations, as between equal-tailed intervals and medians, and between likelihood-ratio-ordered intervals and maximum-likelihood estimates, but these associations are not exclusive. Furthermore, one should not expect an interval to be *centred* on the associated point estimate; this depends on the probability distribution of the observation(s), on the presence of physical boundaries, systematic uncertainties, and so on. Finally, it is sometimes meaningful to report an interval without a point estimate, for example when a new physics process has not been observed and one wishes to provide an upper limit on its production rate.

- *Generality:* Is the interval construction procedure general enough that it can be applied to any problem, regardless of its complexity?

Needless to say, there does not exist a single interval construction method that adequately addresses all the above characterisations in all the problems encountered in practice. It is nevertheless useful to keep these characterisations in mind, and perhaps to prioritise them when searching for an optimal method in a specific situation.

### 4.3

#### Frequentist Methods

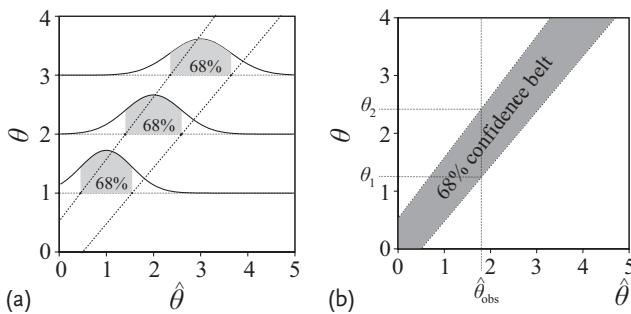
The basic frequentist interval construction is due to Neyman [4]. This procedure is very general, can be applied to multi-dimensional problems and also provides a method for the elimination of *nuisance parameters*.<sup>2)</sup> We therefore start this section with a discussion of this construction. Subsequently, Sections 4.3.2 through 4.3.4 present simpler methods. In less simple, more realistic situations, *bootstrapping methods* as described in Section 4.3.5 can be used. These are particularly well suited to particle physics, where observations ('events') are independent and identically distributed. As a matter of fact, parametric bootstrap methods are already being used by physicists every time they substitute a point estimate for a parameter in a model in order to generate so-called pseudo-data. It is therefore important to understand what can be expected from the bootstrap in terms of some of the interval properties listed in the introduction. We close this discussion of frequentist intervals with comments on the handling of nuisance parameters, together with a detailed case study, in Section 4.3.6.

##### 4.3.1

###### Neyman's Construction

The Neyman construction is illustrated in Figure 4.1 for the case of estimating a one-dimensional continuous parameter  $\theta$  from observations whose distribution

2) 'Eliminating nuisance parameters' is statistics terminology for 'incorporating the effect of systematic uncertainties.'



**Figure 4.1** (a) Neyman construction of a 68% confidence interval on a parameter  $\theta$ . (b) Example use of this construction (see text). Only the part of the construction that falls inside the first quadrant is shown.

depends only on  $\theta$ . The first step is to choose a point estimator  $\hat{\theta}$  of  $\theta$ , to make a graph of  $\theta$  versus  $\hat{\theta}$  and to plot the probability density distribution (pdf) of  $\hat{\theta}$  for several values of  $\theta$ . In Figure 4.1a this has been done for  $\theta = 1, 2$ , and 3. For each value of  $\theta$  considered in step 1, step 2 consists in selecting an interval of  $\hat{\theta}$  values that has a fixed integrated probability, for example 68%. Finally, at step 3 the interval boundaries are connected across  $\theta$  values to obtain the so-called confidence belt. Once data are collected, the observed value  $\hat{\theta}_{\text{obs}}$  of the estimator of  $\theta$  is computed, and the confidence belt is used to derive the corresponding interval  $[\theta_1, \theta_2]$  for  $\theta$  (Figure 4.1b).

To see why this procedure works, consider that if  $\theta_{\text{true}}$  is the true value of  $\theta$ , there is by construction a 68% probability for the point  $(\hat{\theta}_{\text{obs}}, \theta_{\text{true}})$  to be inside the confidence belt, and only when this happens will  $\theta_{\text{true}}$  be inside the interval  $[\theta_1, \theta_2]$  corresponding to  $\hat{\theta}_{\text{obs}}$ . There is therefore a 68% chance that the reported interval will contain  $\theta_{\text{true}}$ , and this holds regardless of the value of  $\theta_{\text{true}}$ .

The Neyman construction requires four ingredients: an estimator  $\hat{\theta}$  of the parameter of interest  $\theta$ , a reference ensemble, an ordering rule, and a confidence level. We now take a look at each of these ingredients individually.

#### 4.3.1.1 Ingredient 1: the Estimator

The estimator is the quantity plotted along the abscissa in the Neyman construction plot. Suppose for example that we collect  $n$  independent measurements  $x_i$  of the mean  $\theta$  of a Gaussian distribution with known standard deviation. Then clearly we should use the average  $\bar{x}$  of the  $x_i$  as an estimate of  $\theta$ , since  $\bar{x}$  is a sufficient statistic<sup>3)</sup> for  $\theta$ . If  $\theta$  is constrained to be positive, then it might make more sense to use  $\hat{\theta} = \max\{0, \bar{x}\}$  instead of  $\hat{\theta} = \bar{x}$ . These two estimators lead to intervals with very different properties. We will come back to this example in Section 4.5.

3) A statistic  $T(X)$  is sufficient for  $\theta$  if the conditional distribution of the sample  $X$  given the value of  $T(X)$  does not depend on  $\theta$ . In a sense,  $T(X)$  captures all the information about  $\theta$  contained in the sample.

It should be pointed out that the original formulation of Neyman's construction does not require a choice of point estimator. It proceeds directly from the distribution of the full data sample at each parameter value. Thus, if the sample contains  $n$  measurements, step 2 of the construction consists in delimiting an  $n$ -dimensional region of sample space with total integrated probability equal to the desired confidence level. This is clearly a non-trivial operation. Fortunately, in the vast majority of practical cases the reduction of the observed sample to a point estimate is a simplifying step that captures all the relevant information.

#### 4.3.1.2 Ingredient 2: the Reference Ensemble

This refers to the probability distribution of the point estimator under replication of the measurement. In order to specify these replications, one must decide which random and non-random aspects of the measurement are relevant to the inference of interest. We give two examples to illustrate this point.

#### Example 4.1 Efficiency estimation

First consider the measurement of an efficiency  $\epsilon$ . A useful point estimator of  $\epsilon$  is the ratio  $\hat{\epsilon} \equiv k/n$  of the number  $k$  of events of interest ('successes') over the total number  $n$  of events collected. However, the distribution of this estimator depends on how the data were collected. If we took data until our total sample reached a certain size, then  $k$  will follow a binomial distribution. If we took data until we found a pre-specified number of events of interest, then the total number of events collected will have a negative binomial distribution. These two data collection schemes differ by their stopping rule, a non-random aspect of the measurement that affects inferences about  $\epsilon$ . Note that these schemes also imply very different prior opinions about  $\epsilon$ : in the binomial case one leaves open the possibility that  $\epsilon$  could be zero, whereas in the negative binomial case  $\epsilon$  is a priori believed to be non-zero.

#### Example 4.2 Mass of a short-lived particle

For an example where a random aspect of the observation affects inferences, consider the measurement of the mass of a short-lived particle whose decay mode determines the measurement resolution. We only have one observation of the particle. Should we then refer our measurement to an ensemble that includes all possible decay modes, or only the decay mode actually observed? For simplicity assume that the estimator  $\hat{\theta}$  of the mass follows a Gaussian distribution with mean  $\theta$  and standard deviation  $\sigma$ , and that there is a probability  $p_h$  that the particle decays hadronically, in which case  $\sigma \equiv \sigma_h$ ; otherwise the particle decays leptonically and  $\sigma \equiv \sigma_\ell < \sigma_h$ . Thus, if we decide to condition on the observed decay mode, the distribution of  $\hat{\theta}$  is Gaussian with mean  $\theta$  and width  $\sigma_h$  or  $\sigma_\ell$ . If we don't condition, the distribution of  $\hat{\theta}$  is a mixture of two Gaussians:

$$f_\theta(\hat{\theta}) = p_h \frac{e^{-\frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma_h}\right)^2}}{\sqrt{2\pi}\sigma_h} + (1-p_h) \frac{e^{-\frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma_\ell}\right)^2}}{\sqrt{2\pi}\sigma_\ell}. \quad (4.1)$$

By ignoring the decay-mode information we can actually expect a more precise measurement. Indeed, if we report our measurement in the form  $\hat{\theta} \pm \delta$ , then  $\delta$  equals  $\sigma_h$  for hadronic decays and  $\sigma_\ell$  for leptonic decays. When the decay mode is ignored,  $\delta$  is the solution of

$$\int_{\theta-\delta}^{\theta+\delta} f_\theta(\hat{\theta}) d\hat{\theta} = p_h \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma_h}\right) + (1-p_h) \operatorname{erf}\left(\frac{\delta}{\sqrt{2}\sigma_\ell}\right) = 0.68 . \quad (4.2)$$

For a numerical example, take  $p_h = 0.5$ ,  $\sigma_h = 10$  and  $\sigma_\ell = 1$  (in arbitrary units). The expected interval length with known decay mode is then  $2[p_h\sigma_h + (1-p_h)\sigma_\ell] = 11.0$ . When the decay mode is ignored, the expected interval length is  $2\delta \approx 9.50$ , noticeably smaller. To understand this feature, imagine repeating the measurement a large number of times. In the conditional analysis the coverage of the interval is 68% both within the subensemble of hadronic decays and within the subensemble of leptonic decays. Conversely, in the unconditional analysis the coverage is  $\operatorname{erf}(\delta/(\sqrt{2}\sigma_h)) \approx 36\%$  for hadronic decays and  $\operatorname{erf}(\delta/(\sqrt{2}\sigma_\ell)) \approx 100\%$  for leptonic decays, correctly averaging to 68% over all decays combined. Qualitatively, by shifting some coverage probability from the hadronic decays to the higher-precision leptonic ones, the unconditional construction is able to reduce the expected interval length.

The above problem is an adaptation to high energy physics of a famous example in the statistics literature [5, 6], used to discuss the merits of conditioning versus power (or interval length). In spite of the loss of expected precision, most physicists would agree to condition on the observed decay mode.

#### 4.3.1.3 Ingredient 3: the Ordering Rule

The ordering rule is the rule we use to decide which  $\hat{\theta}$  values to include in the interval at step 2 of the construction. The only constraint on that interval is that it must contain 68% of the  $\hat{\theta}$  distribution (or whatever confidence level is desired for the overall construction). For example, we could start with the  $\hat{\theta}$  value that has the largest probability density and then keep adding values with lower and lower probability density until we cover 68% of the distribution. Another possibility is to start with  $\hat{\theta} = -\infty$  and add increasing values of  $\hat{\theta}$ , again until we reach 68%. Of course, in order to obtain a smooth confidence belt at the end, we should choose the ordering rule consistently from one  $\theta$  value to the next. This is what endows the resulting intervals with their inferential meaning: an ordering rule is a rule that orders parameter values according to their perceived compatibility with the observed data. Below we list the most common ordering rules, all assuming that we are interested in a  $(1 - \alpha)$ -level confidence set  $C_{1-\alpha}$  for a parameter  $\theta$ . We use a point estimator  $\hat{\theta}$  whose observed value in the data at hand is  $\hat{\theta}_{\text{obs}}$ ; the cumulative distribution of  $\hat{\theta}$  is  $F_\theta(\hat{\theta})$  and its density is  $f_\theta(\hat{\theta})$ .

- *Lower-limit ordering:*  $C_{1-\alpha} = \{\theta : F_\theta(\hat{\theta}_{\text{obs}}) \leq 1 - \alpha\}$

$C_{1-\alpha}$  is the set of  $\theta$  values for which  $\hat{\theta}_{\text{obs}}$  is smaller than or equal to the  $100(1 - \alpha)$ th percentile of  $F_\theta$ . If, as is usually the case,  $\hat{\theta}$  is stochastically increasing with  $\theta$ ,<sup>4)</sup> then the parameter value  $\theta_{\text{low}}$  with  $F_{\theta_{\text{low}}}(\hat{\theta}_{\text{obs}}) = 1 - \alpha$  is the lower limit of  $C_{1-\alpha}$ .

- *Upper-limit ordering:*  $C_{1-\alpha} = \{\theta : F_\theta(\hat{\theta}_{\text{obs}}) \geq \alpha\}$

$C_{1-\alpha}$  is the set of  $\theta$  values for which  $\hat{\theta}_{\text{obs}}$  is larger than or equal to the  $100\alpha^{\text{th}}$  percentile of  $F_\theta$ . The parameter value  $\theta_{\text{up}}$  with  $F_{\theta_{\text{up}}}(\hat{\theta}_{\text{obs}}) = \alpha$  is the upper limit of the set  $C_{1-\alpha}$ .

- *Equal-tails ordering:*  $C_{1-\alpha} = \{\theta : \frac{\alpha}{2} \leq F_\theta(\hat{\theta}_{\text{obs}}) \leq 1 - \frac{\alpha}{2}\}$

$C_{1-\alpha}$  is the set of  $\theta$  values for which  $\hat{\theta}_{\text{obs}}$  falls between the  $100(\frac{\alpha}{2})^{\text{th}}$  and  $100(1 - \frac{\alpha}{2})^{\text{th}}$  percentiles of  $F_\theta$ . The previous definitions of lower and upper limits show that equal-tailed intervals must have the form  $C_{1-\alpha} = [\theta_1, \theta_2]$ , where the boundaries are themselves confidence limits:  $]-\infty, \theta_1]$  and  $[\theta_2, +\infty[$  are both  $(\frac{\alpha}{2})$  CL intervals. Furthermore,  $\theta_1$  and  $\theta_2$  can be solved from  $F_{\theta_1}(\hat{\theta}_{\text{obs}}) = 1 - \frac{\alpha}{2}$  and  $F_{\theta_2}(\hat{\theta}_{\text{obs}}) = \frac{\alpha}{2}$ . The relationship between equal-tailed intervals and confidence limits leads to the following interpretation of the former in terms of ‘plausibility’ ([7], p. 157): values of  $\theta$  smaller than  $\theta_1$  are implausible because they result in probability less than  $\alpha/2$  of obtaining a  $\hat{\theta}$  value at least as *large* as observed, and values of  $\theta$  larger than  $\theta_2$  are implausible because they result in probability less than  $\alpha/2$  of obtaining a  $\hat{\theta}$  value at least as *small* as observed.

- *Probability-density ordering:*  $C_{1-\alpha} = \{\theta : f_\theta(\hat{\theta}_{\text{obs}}) \geq k_{1-\alpha}(\theta)\}$

$C_{1-\alpha}$  is the set of  $\theta$  values for which  $\hat{\theta}_{\text{obs}}$  falls within the  $100(1-\alpha)\%$  most probable region of  $f_\theta$ . The cutoff  $k_{1-\alpha}(\theta)$  is determined by the coverage requirement, namely that  $\int f_\theta(\hat{\theta}) d\hat{\theta} = 1 - \alpha$ , where the integral is over all  $\hat{\theta}$  values for which  $f_\theta(\hat{\theta}) \geq k_{1-\alpha}(\theta)$ ; this requirement can introduce a  $\theta$  dependence in  $k_{1-\alpha}$ , but no  $\hat{\theta}$  dependence.

- *Likelihood-ratio ordering:*  $C_{1-\alpha} = \{\theta : f_\theta(\hat{\theta}_{\text{obs}})/[\max_{\theta'} f_{\theta'}(\hat{\theta}_{\text{obs}})] \geq k'_{1-\alpha}(\theta)\}$

$C_{1-\alpha}$  is the set of  $\theta$  values for which  $\hat{\theta}_{\text{obs}}$  falls in the region of sampling probability  $1 - \alpha$  where the likelihood ratio in favour of  $\theta$  is larger than anywhere outside the region ( $k'_{1-\alpha}(\theta)$  is fixed by the coverage requirement). Note that the maximisation in the denominator of the likelihood ratio must be restricted to the physical region of  $\theta$ -space [8].

In contrast with the equal-tails, upper-limit, and lower-limit ordering rules, the probability-density and likelihood-ratio rules do not always produce simple intervals: in complex problems they may yield confidence sets that are unions of disjoint intervals.

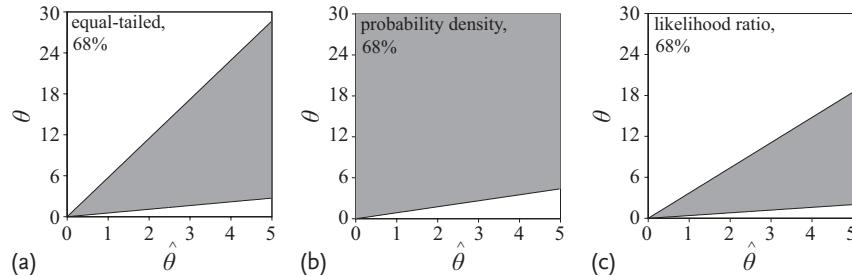
Table 4.1 summarises these ordering rules together with their defining equations, and shows the result of applying them to a measurement of the lifetime  $\theta$  of an exponential decay. In this case the probability density of the estimated lifetime  $\hat{\theta}$  is  $f_\theta(\hat{\theta}) = \exp(-\hat{\theta}/\theta)/\theta$ , and the cumulative distribution is  $F_\theta(\hat{\theta}) = 1 - \exp(-\hat{\theta}/\theta)$ .

4) The random variable  $\hat{\theta}$  is said to be stochastically increasing with the parameter  $\theta$  if  $\theta_1 < \theta_2$  implies  $F_{\theta_1}(\hat{\theta}) > F_{\theta_2}(\hat{\theta})$ . In words, the bulk of the distribution of  $\hat{\theta}$  tracks changes in  $\theta$ .

**Table 4.1** Common ordering rules used in constructing frequentist intervals with confidence level  $1 - \alpha$ . The first two columns give the name and defining equation for each rule, the third column shows the solution of the defining equation for the measurement of the lifetime  $\theta$  of an exponential decay, and the last column applies the solution to the case

$1 - \alpha = 68\%$ . The quantity  $1 - \alpha'$  at the bottom of the third column is the unique number between 0 and  $\alpha$  that satisfies the equation  $(1 - \alpha' + 1 - \alpha) \ln(1 - \alpha' + 1 - \alpha) = (1 - \alpha') \ln(1 - \alpha')$ . For  $1 - \alpha = 0.68$ ,  $1 - \alpha' \approx 0.0829$ . For the exponential-decay example,  $k_{1-\alpha}(\theta) = \alpha/\theta$  and  $k'_{1-\alpha}(\theta) = -e(1 - \alpha') \ln(1 - \alpha')$ .

Ordering rule	Defining equation	Exponential decay example	
		General solution	Case $1 - \alpha = 68\%$
Lower limit:	$F_{\theta_{\text{low}}}(\hat{\theta}_{\text{obs}}) = 1 - \alpha$	$\theta_{\text{low}} = \frac{-\hat{\theta}_{\text{obs}}}{\ln \alpha}$	$[0.88\hat{\theta}_{\text{obs}}, +\infty[$
Upper limit:	$F_{\theta_{\text{up}}}(\hat{\theta}_{\text{obs}}) = \alpha$	$\theta_{\text{up}} = \frac{-\hat{\theta}_{\text{obs}}}{\ln(1-\alpha)}$	$[0, 2.59\hat{\theta}_{\text{obs}}]$
Equal tails:	$\begin{cases} F_{\theta_1}(\hat{\theta}_{\text{obs}}) = 1 - \frac{\alpha}{2} \\ F_{\theta_2}(\hat{\theta}_{\text{obs}}) = \frac{\alpha}{2} \end{cases}$	$\begin{cases} \theta_1 = \frac{-\hat{\theta}_{\text{obs}}}{\ln(\frac{\alpha}{2})} \\ \theta_2 = \frac{-\hat{\theta}_{\text{obs}}}{\ln(1-\frac{\alpha}{2})} \end{cases}$	$[0.55\hat{\theta}_{\text{obs}}, 5.74\hat{\theta}_{\text{obs}}]$
Prob. density:	$f_{\theta}(\hat{\theta}_{\text{obs}}) \geq k_{1-\alpha}(\theta)$		Same as lower limit
Likelihood ratio:	$\frac{f_{\theta}(\hat{\theta}_{\text{obs}})}{\max_{\theta'} f_{\theta'}(\hat{\theta}_{\text{obs}})} \geq k'_{1-\alpha}(\theta)$	$\begin{cases} \theta_1 = \frac{-\hat{\theta}_{\text{obs}}}{\ln(1-\alpha')} \\ \theta_2 = \frac{-\hat{\theta}_{\text{obs}}}{\ln(1-\alpha'+1-\alpha)} \end{cases}$	$[0.40\hat{\theta}_{\text{obs}}, 3.70\hat{\theta}_{\text{obs}}]$



**Figure 4.2** Confidence belts for an exponential lifetime, with  $1 - \alpha = 68\%$  and three different ordering rules: (a) equal-tailed; (b) probability density; (c) likelihood ratio.

It is interesting to note that the boundaries of the exponential intervals increase linearly with the observation  $\hat{\theta}_{\text{obs}}$ . Since  $\hat{\theta}$  is an unbiased estimator of  $\theta$ , the expected lengths of these intervals are trivial to obtain: for  $1 - \alpha = 68\%$  they are  $5.19\theta$ ,  $2.59\theta$ , and  $3.29\theta$  for the equal-tailed, upper-limit, and likelihood-ratio intervals, respectively. Some confidence belts for this measurement are plotted in Figure 4.2.

#### 4.3.1.4 Ingredient 4: the Confidence Level

The confidence level labels a family of intervals; some conventional values are 68%, 90%, and 95%. It is very important to understand that a confidence level does *not*

characterise single intervals; it only characterises families of intervals. The following example illustrates this.

### Example 4.3 Mass of a new elementary particle

Suppose we wish to measure the mass  $\theta$  of a new elementary particle, and assume for simplicity that our measurement  $x$  of this mass has a Gaussian distribution with unit variance. Thus, even though  $\theta$  must be positive for physics reasons, measurement resolution effects can cause  $x$  to be negative. Before performing our measurement we decide that we will report a 68% CL likelihood-ratio ordered interval. However, a colleague of ours prefers to report an 84% CL upper limit. The measurement is then performed and yields  $x = 0$ , leading both of us to report the same numerical interval  $[0.0, 0.99]$ . This demonstrates that the same numerical interval can have two very different coverages (confidence levels), depending on which ensemble it is considered to belong to.

In the above example the frequentist coverages of both interval procedures agree exactly with their respective confidence levels. As mentioned in Section 4.2, this agreement is not always possible, especially when the observations are discrete or when nuisance parameters are present. In general an interval construction is considered valid from the frequentist point of view if its coverage, as a function of the true value of the parameter of interest, is everywhere equal to, or larger than, the stated confidence level. If this is not the case one says that the construction *undercovers*. One way to verify the coverage characteristics of a given interval procedure is to plot the interval boundaries as a function of the observation. This yields a confidence belt, as in Figure 4.1b. For each value of the parameter  $\theta$ , integrating the probability density of the observation between the confidence belt boundaries yields the coverage at that particular  $\theta$  value.

#### 4.3.2

#### Test Inversion

As indicated in the previous subsection, the inferential core of Neyman's construction is the ordering rule; the rest is just a geometrical embedding to enforce the coverage constraint. The ordering rule itself can be viewed as the construction of a test for each physical value of the parameter; each such test has a different acceptance region in sample space (see Chapter 3). Therefore, if we have a proper frequentist test to start with, we can dispense with the rest of Neyman's construction and proceed directly to defining the confidence interval as the set of parameter values for which the acceptance region contains the observation. This is known as the *test-inversion method*. To fix the notation, suppose we are interested in a parameter  $\theta \in \Theta$ , and that for each allowed value  $\theta_0$  of  $\theta$  we can construct a size  $\alpha$  test of

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0 . \quad (4.3)$$

Consider then the set  $C_{1-\alpha}$  of all the  $\theta_0$  values for which  $H_0$  is accepted. This set depends on the observations and is therefore random. We have:

$$P[\theta_0 \in C_{1-\alpha} | \theta = \theta_0] = P[H_0 \text{ is accepted} | H_0] = 1 - \alpha. \quad (4.4)$$

Hence  $C_{1-\alpha}$  is a  $(1 - \alpha)$  CL set for  $\theta$ . To picture the shape of this set, note that if  $H_0$  is rejected for a given  $\theta_0$ , then, because of the form of  $H_1$ , all values of  $\theta$  larger than  $\theta_0$  will also be rejected. Therefore, the set  $C_{1-\alpha}$  of accepted  $\theta$  values will have an upper boundary  $\theta_{\text{up}}$ . For the simple example of a Gaussian data point  $x$  with mean  $\theta$  and known standard deviation  $\sigma$ , one can test (4.3) with the statistic  $y \equiv \theta_0 - x$ . Under  $H_0$  this statistic has a Gaussian distribution with mean zero and standard deviation  $\sigma$ , and large observed values  $y_{\text{obs}}$  of  $y$  constitute evidence against  $H_0$  in the direction of  $H_1$ . The  $p$ -value of test (4.3) is therefore

$$\begin{aligned} p(\theta_0) &= \int_{y_{\text{obs}}}^{\infty} \frac{e^{-\frac{1}{2}(\frac{y}{\sigma})^2}}{\sqrt{2\pi}\sigma} dy = \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{y_{\text{obs}}}{\sqrt{2}\sigma} \right) \right] \\ &= \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{\theta_0 - x_{\text{obs}}}{\sqrt{2}\sigma} \right) \right]. \end{aligned} \quad (4.5)$$

Values of  $\theta$  for which  $p(\theta) \geq \alpha$  are accepted by the test and included in the confidence interval. The upper limit of the interval is the solution of  $p(\theta_{\text{up}}) = \alpha$ , which is  $\theta_{\text{up}} = x_{\text{obs}} + z_{1-\alpha}\sigma$ , where  $z_{1-\alpha} \equiv \sqrt{2}\text{erf}^{-1}(1-2\alpha)$  is the  $(1-\alpha)$ -quantile of the standard normal distribution (a Gaussian with zero mean and unit standard deviation).

If one is interested in a  $(1 - \alpha)$  CL lower limit  $\theta_{\text{low}}$  on  $\theta$ , the test to consider is

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H'_1 : \theta > \theta_0, \quad (4.6)$$

again with size  $\alpha$ . For the Gaussian example the result is  $\theta_{\text{low}} = x_{\text{obs}} - z_{1-\alpha}\sigma$ .

A  $(1 - \alpha)$  CL two-sided interval for  $\theta$  can be obtained by computing lower and upper limits at the  $(1 - \frac{\alpha}{2})$  CL, or by inverting a size  $\alpha$  two-sided test:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H''_1 : \theta \neq \theta_0. \quad (4.7)$$

In this case, an appropriate test statistic for the Gaussian example is  $y = |x - \theta_0|$ , which has a folded Gaussian distribution. By solving the appropriate  $p$ -value equation as before, one obtains  $[\theta_{\text{low}}, \theta_{\text{up}}]$  with  $\theta_{\text{low}} = x_{\text{obs}} - z_{1-\frac{\alpha}{2}}\sigma$  and  $\theta_{\text{up}} = x_{\text{obs}} + z_{1-\frac{\alpha}{2}}\sigma$ .

It should be clear from the above discussion that the construction of confidence intervals by this method requires the inversion of a *family* of tests rather than of a single test. Thus, the method will not work if one has a nice test for a special value of the parameter of interest, but the test does not generalise to other values. It may also happen that inversion of a family of tests results in a union of disjoint intervals rather than a single interval. In general one can expect the properties of a family of tests to be reflected in the properties of the resulting intervals: conservative tests lead to wide intervals, and powerful tests to narrow intervals.

## 4.3.3

**Pivoting**

A pivot is a function  $Q(\theta, \mathbf{x})$  of both the observation  $\mathbf{x}$  and the parameter  $\theta$  whose distribution does not depend on any unknown parameters (not even on  $\theta$ ). Because of this special property, it is in principle possible to find, for any  $\alpha \in [0, 1]$ , constants  $a(\alpha)$  and  $b(\alpha)$  such that  $P[a(\alpha) \leq Q(\theta, X) \leq b(\alpha)] = 1 - \alpha$  for all  $\theta$ . Therefore, the set of observations  $\mathbf{x}$  such that  $a(\alpha) \leq Q(\theta, \mathbf{x}) \leq b(\alpha)$  can be interpreted as the acceptance region of a size  $\alpha$  test of the hypothesis that  $\theta$  is the true value. Since this acceptance region is by construction valid for testing any  $\theta$ , inverting the test leads to a  $(1-\alpha)$  CL set for  $\theta$ . This is the *pivoting method* for constructing confidence sets. In general there is no guarantee that such sets will be simple intervals, or that they will be optimal in any sense, but the simplicity of the construction makes it worth trying in situations that allow it. Furthermore, the pivot concept is crucial to the development of the frequentist theory of confidence intervals beyond the setting where exact solutions can be found. This will become clear in Section 4.3.4 on asymptotic approximations and Section 4.3.5 on the bootstrap.

## 4.3.3.1 Gaussian Means and Standard Deviations

To illustrate the pivoting method, consider the example of  $n$  measurements  $x_i$  from a Gaussian distribution with mean  $\theta$  and standard deviation  $\sigma$ . This example has a particularly rich pivot structure. Suppose first that  $\theta$  is unknown and  $\sigma$  known. In this case the quantity

$$Q_1(\theta, \mathbf{x}) \equiv \frac{\bar{x} - \theta}{\sigma/\sqrt{n}}, \quad \text{where } \bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i, \quad (4.8)$$

has a standard normal distribution and is therefore a pivot. Thus, writing  $z_\gamma$  for the corresponding  $\gamma$ -quantile, we have:

$$1 - \alpha = P[z_{\frac{\alpha}{2}} \leq Q_1(\theta, \mathbf{x}) \leq z_{1-\frac{\alpha}{2}} \mid \theta, \sigma], \quad (4.9)$$

$$= P\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{x} - \theta}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \mid \theta, \sigma\right], \quad (4.10)$$

$$= P\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \mid \theta, \sigma\right], \quad (4.11)$$

$$= P\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \mid \theta, \sigma\right], \quad (4.12)$$

where in the last line we used the symmetry of the standard normal distribution to write  $z_\gamma = -z_{1-\gamma}$ . This result shows that we have obtained a symmetric confidence interval for  $\theta$ . Setting for example  $1 - \alpha = 68\%$  yields  $z_{1-\frac{\alpha}{2}} = z_{0.84} = 1$ , and the confidence interval after observing  $\bar{x} = \bar{x}_{\text{obs}}$  is simply  $\bar{x}_{\text{obs}} \pm \sigma/\sqrt{n}$ .

A pivot is not necessarily unique or optimal. Consider for instance the case where  $\theta$  is known and  $\sigma$  unknown. In principle we could use pivot (4.8) again,

this time to construct confidence intervals for the variance  $\sigma^2$ . Solving

$$z_{\frac{\alpha}{2}} \leq \frac{\bar{x}_{\text{obs}} - \theta}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \quad (4.13)$$

for  $\sigma^2$  yields a  $(1 - \alpha)$  CL lower limit

$$\sigma^2 \geq \frac{n(\bar{x}_{\text{obs}} - \theta)^2}{\chi^2_{1,1-\alpha}}, \quad (4.14)$$

where  $\chi^2_{n,1-\alpha}$  is the  $(1 - \alpha)$ -quantile of a  $\chi^2$  for  $n$  degrees of freedom, and we used the relation  $z_{\alpha/2}^2 = \chi^2_{1,1-\alpha}$ .

We can then take advantage of the fact that the interval between two lower limits, one at the  $(1 - (\alpha/2))$  CL and the other at the  $(\alpha/2)$  CL, is itself a  $(1 - \alpha)$  CL equal-tailed two-sided interval, to obtain

$$\frac{n(\bar{x}_{\text{obs}} - \theta)^2}{\chi^2_{1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{n(\bar{x}_{\text{obs}} - \theta)^2}{\chi^2_{1,\frac{\alpha}{2}}} \quad \text{with confidence } 1 - \alpha. \quad (4.15)$$

However, this interval is not optimal because it is based on a rather poor estimator of  $\sigma^2$ , namely  $n(\bar{x} - \theta)^2$ , which has a variance of  $2\sigma^4$ . In contrast, the usual estimator

$$S_{\theta,n}^2 \equiv \frac{1}{n} \sum_{i=1}^n (x_i - \theta)^2 \quad (4.16)$$

has variance  $2\sigma^4/n$ . Helpfully, the quantity  $nS_{\theta,n}^2/\sigma^2$  is pivotal with a  $\chi^2_n$  distribution and can therefore serve to construct intervals for  $\sigma$ . We have:

$$P \left[ \chi^2_{n,\frac{\alpha}{2}} \leq \sum_{i=1}^n \left( \frac{x_i - \theta}{\sigma} \right)^2 \leq \chi^2_{n,1-\frac{\alpha}{2}} \mid \theta, \sigma \right] = 1 - \alpha, \quad (4.17)$$

so that

$$\frac{\sum_{i=1}^n (x_i - \theta)^2}{\chi^2_{n,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (x_i - \theta)^2}{\chi^2_{n,\frac{\alpha}{2}}} \quad \text{with confidence } 1 - \alpha. \quad (4.18)$$

Note that for  $n = 1$  this interval coincides with that given in (4.15). However, at larger values of  $n$ , the interval (4.18) has smaller expected length.

Finally, we consider the case where both  $\theta$  and  $\sigma$  are unknown. If we are interested in  $\sigma^2$ , we can use

$$S_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.19)$$

as estimator, and construct intervals based on the fact that

$$Q_2(\sigma, \mathbf{x}) \equiv (n-1) \frac{S_n^2}{\sigma^2} \quad (4.20)$$

is a pivot following a  $\chi^2$  distribution function with  $n - 1$  degrees of freedom. Conversely, if  $\theta$  is of interest, a new pivot can be constructed by taking the ratio of  $Q_1(\theta, \mathbf{x})$  and  $\sqrt{Q_2(\sigma, \mathbf{x})/(n-1)}$ :

$$Q_3(\theta, \sigma, \mathbf{x}) = \frac{\sqrt{n}(\bar{x} - \theta)/\sigma}{\sqrt{S_n^2/\sigma^2}} = \frac{\bar{x} - \theta}{S_n/\sqrt{n}}. \quad (4.21)$$

This ratio is distributed as a central  $t$  variate<sup>5)</sup> for  $n - 1$  degrees of freedom. With  $t_{n-1,1-\alpha}$  the  $(1 - \alpha)$ -quantile of Student's  $t_{n-1}$  distribution and  $(\bar{x}_{\text{obs}}, S_n)$  the observed value of  $(\bar{x}, S_n)$ , the following is a  $(1 - \alpha)$  CL equal-tailed, symmetric interval for  $\theta$ :

$$\bar{x} - \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \theta \leq \bar{x} + \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}. \quad (4.22)$$

For  $1 - \alpha = 68\%$  one finds  $t_{1,1-\frac{\alpha}{2}} = 1.82$  and  $t_{19,1-\frac{\alpha}{2}} = 1.02$ . As  $n$  becomes large, the  $t_{n-1}$  quantiles converge to the corresponding standard normal quantiles.

#### 4.3.3.2 Exponential Lifetimes

Although the previous subsection is limited to problems involving the Gaussian distribution, the pivoting method can be applied to many other situations. In fact, a pivot that is often available is the *cumulative distribution* of the data (cdf), viewed as a function of the data and the parameters. For continuous data this pivot is uniformly distributed. We illustrate this idea with the construction of confidence intervals on the lifetime  $\tau$  associated with an exponential decay. The cdf of the measurement  $t$ , and hence the pivot, is  $Q(\tau, t) = 1 - e^{-t/\tau}$ . Since this is a uniform pivot, a  $(1 - \alpha)$  CL interval for  $\tau$  is given by

$$\frac{\alpha}{2} \leq 1 - e^{-t/\tau} \leq 1 - \frac{\alpha}{2} \quad (4.23)$$

or

$$\frac{t}{-\ln(\frac{\alpha}{2})} \leq \tau \leq \frac{t}{-\ln(1 - \frac{\alpha}{2})}. \quad (4.24)$$

For  $1 - \alpha = 68\%$  this yields  $\tau \in [0.55t, 5.74t]$ , which is the equal-tailed interval listed in Table 4.1.

#### 4.3.3.3 Binomial Efficiencies

For discrete data the cdf is no longer an exact pivot, but it can still be used to construct confidence sets. As an example we consider the measurement of an efficiency  $\epsilon$  based on the observation of  $x$  successes out of  $n$  trials. The cdf is binomial, but it will be convenient to express it in terms of a Beta cdf  $B(x; a, b)$ ,

$$\begin{aligned} P[K \leq x | \epsilon, n] &= \sum_{k=0}^x \binom{n}{k} \epsilon^k (1 - \epsilon)^{n-k} \\ &= \int_0^{1-\epsilon} \frac{t^{n-x-1} (1-t)^x}{B(x+1, n-x)} dt = B(1 - \epsilon; n - x, x + 1), \end{aligned} \quad (4.25)$$

5) A variate is a random variable.

where  $B(a, b) \equiv \Gamma(a)\Gamma(b)/\Gamma(a + b)$ . The second equality can be derived by integration by parts. Although the cdf of a binomial variable  $K$  is not an exact pivot, it can be made exact by introducing a random variable  $U$  that is uniform on the interval  $[0, 1]$ , and considering the cdf of the sum  $K + U$ . This is a continuous cdf and therefore an exact uniform pivot, so that the following inequalities define a  $(1 - \alpha)$  CL interval for  $\epsilon$ :

$$\frac{\alpha}{2} \leq P[K + U \leq x|\epsilon, n] \leq 1 - \frac{\alpha}{2}. \quad (4.26)$$

Since  $U$  is unobserved, solving for  $\epsilon$  requires a worst-case analysis, in which the random variable  $U$  is replaced by a constant such that the inequalities in formula (4.26) hold regardless of the value of  $U$ . For the lower limit this means that  $U$  should be replaced by 0:

$$\frac{\alpha}{2} \leq P[K + U \leq x|\epsilon, n] \leq P[K \leq x|\epsilon, n] = P[B_{n-x,x+1} \leq 1 - \epsilon], \quad (4.27)$$

where we used (4.25) and  $B_{a,b}$  is a random variable with a Beta( $a, b$ ) distribution. Writing  $B_{a,b,\alpha}$  for the  $\alpha$ -quantile of this distribution, the above result implies that

$$1 - \epsilon \geq B_{n-x,x+1,\frac{\alpha}{2}}, \quad (4.28)$$

which yields the upper limit  $\epsilon_{\text{up}}$  of the desired interval for  $\epsilon$ :

$$\epsilon \leq 1 - B_{n-x,x+1,\frac{\alpha}{2}} = B_{x+1,n-x,1-\frac{\alpha}{2}} \equiv \epsilon_{\text{up}}. \quad (4.29)$$

This expression is undefined for  $x = n$ , in which case we set  $\epsilon_{\text{up}} = 1$ . In order to guarantee the upper inequality in (4.26), we must replace the unobserved random variable  $U$  by the constant 1. Similar manipulations as above yield the lower limit  $\epsilon_{\text{low}}$  of the interval:

$$\epsilon \geq B_{x,n-x+1,\frac{\alpha}{2}} \equiv \epsilon_{\text{low}}. \quad (4.30)$$

For  $x = 0$  we set  $\epsilon_{\text{low}} = 0$ . The interval  $[\epsilon_{\text{low}}, \epsilon_{\text{up}}]$ , with endpoints given in (4.29) and (4.30), is known as a  $(1 - \alpha)$  CL Clopper–Pearson interval for the efficiency  $\epsilon$  [9]. This interval is easy to code into a computer program, using the incomplete beta function [10]. It can also be computed from tables of Snedecor's  $F$  distribution and the following relationship between Beta and  $F$  quantiles:

$$B_{a,b,\gamma} = \left(1 + \frac{b}{a} F_{2b,2a,1-\gamma}\right)^{-1}. \quad (4.31)$$

Because of the worst-case analysis involving the random variable  $U$ , Clopper–Pearson intervals are conservative (they overcover). Overcoverage is generally unavoidable in discrete sample spaces. An in-depth comparison with other constructions can be found in [11].

#### 4.3.3.4 Poisson Means

Another frequent application in physics is the computation of an upper limit on the expected mean  $\theta$  of the number of events of a new signal at a collider experiment. The observation is a number of events  $n$ , which is assumed to be Poisson distributed with mean  $\theta + \nu$ , where  $\nu$  is a known background contamination. The cdf is again discrete and therefore not an exact pivot, but we can make it exact by adding a uniform variate  $U$  on  $[0, 1]$  to the Poisson variate  $N$ . Thus, the set of  $\theta$  satisfying  $\alpha \leq P[N + U \leq n | \theta + \nu]$  is an exact  $(1 - \alpha)$  CL interval. Repeating the worst-case analysis argument of the previous section, we replace  $U$  by the constant 0 to obtain a conservative interval. This is a one-sided interval bounded by an upper limit, as we now show. First note that

$$\alpha \leq P[N \leq n | \theta + \nu] = \sum_{k=0}^n \frac{(\theta + \nu)^k e^{-\theta-\nu}}{k!} = \int_{\theta+\nu}^{+\infty} \frac{t^n e^{-t}}{\Gamma(n+1)} dt , \quad (4.32)$$

where the last equality can be proved by integration by parts. After substituting  $t = z/2$  in the integrand on the right, one recognises this as the cdf of a  $\chi^2$  variate  $\chi^2_{2(n+1)}$  for  $2(n+1)$  degrees of freedom. With some rearrangement the above inequality can therefore be rewritten as

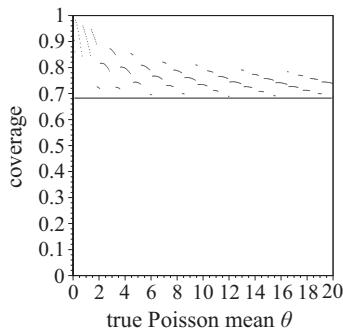
$$P\left[\chi^2_{2(n+1)} \leq 2(\theta + \nu)\right] \leq 1 - \alpha , \quad (4.33)$$

which implies that  $2(\theta + \nu)$  is smaller than the  $(1 - \alpha)$ -quantile of the variate  $\chi^2_{2(n+1)}$ . Hence the following is a  $(1 - \alpha)$  CL upper limit on  $\theta$ :

$$\theta \leq \frac{1}{2} \chi^2_{2(n+1), 1-\alpha} - \nu . \quad (4.34)$$

**Table 4.2** Frequentist interval constructions for the mean of a Poisson distribution when  $N$  events are observed: 95% CL upper limits (column 2), 68% CL equal-tailed intervals (column 3), and 95% (column 4) and 68% (column 5) CL Feldman–Cousins intervals (from [8]).

$N$	Upper limit 95% CL	Garwood		Feldman–Cousins	
		Equal-tailed interval 68% CL	95% CL	95% CL	68% CL
0	3.00	[0.00, 1.84]	[0.00, 3.09]	[0.00, 1.29]	
1	4.74	[0.17, 3.30]	[0.05, 5.14]	[0.37, 2.75]	
2	6.30	[0.71, 4.64]	[0.36, 6.72]	[0.74, 4.25]	
3	7.75	[1.37, 5.92]	[0.82, 8.25]	[1.10, 5.30]	
4	9.15	[2.09, 7.16]	[1.37, 9.76]	[2.34, 6.78]	
5	10.51	[2.84, 8.38]	[1.84, 11.26]	[2.75, 7.81]	
6	11.84	[3.62, 9.58]	[2.21, 12.75]	[3.82, 9.28]	
7	13.15	[4.42, 10.77]	[2.58, 13.81]	[4.25, 10.30]	
8	14.43	[5.23, 11.95]	[2.94, 15.29]	[5.30, 11.32]	
9	15.71	[6.06, 13.11]	[4.36, 16.77]	[6.33, 12.79]	
10	16.96	[6.89, 14.27]	[4.75, 17.82]	[6.78, 13.81]	



**Figure 4.3** Frequentist coverage of 68% CL Garwood central intervals for the mean of a Poisson distribution. The coverage is evaluated in increments of 0.1 in the Poisson mean, and the nominal coverage of the construction is indicated by the solid horizontal line.

This result was first reported by Garwood in 1936 [12]. For a  $(1 - \alpha)$  CL two-sided interval, similar calculations yield

$$\left[ \frac{1}{2} \chi_{2n, \frac{\alpha}{2}}^2 - \nu, \frac{1}{2} \chi_{2(n+1), 1 - \frac{\alpha}{2}}^2 - \nu \right]. \quad (4.35)$$

For  $n = 0$  the lower limit of the interval is  $-\nu$ . Some numerical examples of (4.34) and (4.35), for  $\nu = 0$ , are shown in Table 4.2. Also shown there are *Feldman–Cousins intervals*, which are based on a likelihood-ratio ordering rule [8]. A significant advantage of Feldman–Cousins intervals is that they are never unphysical, regardless of how large the background contamination  $\nu$  is. This is not the case for Garwood intervals. The frequentist coverage of 68% CL Garwood central intervals is plotted in Figure 4.3 (again with  $\nu = 0$ ). Viewed as a function of the true Poisson mean, the coverage is highly discontinuous due to the discreteness of the Poisson distribution.

#### 4.3.4 Asymptotic Approximations

In Section 4.3.1.3 we mentioned the likelihood-ratio ordering rule as an option for the construction of Neyman confidence sets. Given data  $x$ , a parameter of interest  $\theta$  and its maximum-likelihood estimate (MLE)  $\hat{\theta} = \hat{\theta}(x)$ , this rule includes in the confidence set any  $\theta$  value that is not rejected by an  $\alpha$ -level test based on the likelihood ratio  $\lambda(x; \theta) \equiv L(x; \theta)/L(x; \hat{\theta})$ . For large samples it turns out that the confidence level constraint is easy to implement thanks to Wilks's theorem [13]. The latter states that, under standard regularity conditions,  $-2 \ln \lambda(x; \theta)$  is asymptotically distributed as a  $\chi^2$  variate for  $d$  degrees of freedom, where  $d$  equals the dimensionality of  $\theta$  (in the terminology of Section 4.3.3,  $-2 \ln \lambda(x; \theta)$  is an asymptotic pivot). This provides a simple way to construct a  $(1 - \alpha)$  CL interval, by taking the set of  $\theta$  values for which

$$-2 \ln \lambda(x; \theta) \leq \chi_{d, 1 - \alpha}^2, \quad (4.36)$$

where  $\chi^2_{d,1-\alpha}$  is the  $(1 - \alpha)$ -quantile of a  $\chi^2$  distribution for  $d$  degrees of freedom. Thus, if  $\theta$  is one-dimensional, use  $\chi^2_{1,0.68} \approx 1$  for a 68% CL interval,  $\chi^2_{1,0.95} \approx 4.00$  for a 95% CL interval, and so on.

If nuisance parameters are present, collectively labelled  $\nu$ , the same result applies provided the likelihood ratio is defined by

$$\lambda(x; \theta) \equiv \frac{L(x; \theta, \hat{\nu}(\theta))}{L(x; \hat{\theta}, \hat{\nu})}, \quad (4.37)$$

where  $\hat{\nu}(\theta)$  is the profile likelihood estimate of  $\nu$ , that is, its MLE evaluated at a fixed value of  $\theta$ , and  $\hat{\nu}$  is the global MLE of  $\nu$ , without constraining to a fixed  $\theta$ .

For one-dimensional  $\theta$  it is often helpful to plot a graph of  $-2 \ln \lambda(x; \theta)$  versus  $\theta$ , since this allows the interval to be determined at various confidence levels, and to assess the ‘Gaussianity’ of the problem, for example whether 95% CL intervals have twice the length of 68% CL intervals. In addition, it may happen that confidence sets obtained by this method consist of two or more disjoint intervals, in which case a plot is most useful. For a two-dimensional parameter vector  $\boldsymbol{\theta}$  one can plot contours of  $-2 \ln \lambda(x; \boldsymbol{\theta})$  in the plane of  $\boldsymbol{\theta}$  values. Then for example, the contour corresponding to  $-2 \ln \lambda(x; \boldsymbol{\theta}) = \chi^2_{2,0.68} \approx 2.30$  encloses a 68% confidence region for  $\boldsymbol{\theta}$ , and for 95% confidence one should use  $\chi^2_{2,0.95} \approx 6.18$ . In high energy physics these constructions are typically done with the help of the routine *Minos* in the *MINUIT* program package [14]. A general treatment of likelihood asymptotics for high energy physics can be found in [15].

#### 4.3.5

#### Bootstrapping

The confidence interval constructions we have examined so far all assume that it is possible to write down explicitly the probability distribution of the data in analytical form, including its dependence on the parameter of interest and on nuisance parameters. Unfortunately this is not always the case in high energy physics. A good example is the measurement of the top quark mass, where the dependence of data distributions on the parameter of interest is buried deeply in complex Monte Carlo simulations of physics processes and detector responses. The bootstrap method provides a powerful way to circumvent this difficulty. It is a bridge between exact methods, which cannot be used in complex physics analyses, and asymptotic methods, which lack coverage accuracy in finite samples. There are two ideas at the core of bootstrap methods [16]: the *plug-in principle* and *resampling*. The plug-in principle sounds rather obvious as it states that in order to estimate a quantity of interest, one should replace the unknown data cdf  $F$  by an estimate  $\hat{F}$ . If nothing is known a priori about  $F$ , and all we have is a data sample  $x_1, x_2, \dots, x_n$ , then  $F$  can be estimated by the *empirical* distribution of the data, which assigns probability  $1/n$  to each data point:

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n}. \quad (4.38)$$

Another possibility is that  $F$  has a known functional form that depends on some unknown parameter  $\psi$ , in which case it could be estimated by substituting the maximum-likelihood estimate (MLE)  $\hat{\psi}$  for  $\psi$ .

Suppose now that we are interested in estimating a quantity  $\theta$ , which could be something as simple as the mean of a population characteristic or as complex as the mass of the top quark. The true value of  $\theta$  is defined as the result of applying the appropriate estimating procedure to the true distribution, which we write as  $\theta = \theta(F)$ , whereas the plug-in estimate of  $\theta$  is obtained by applying the same procedure to the estimated distribution,  $\hat{\theta} = \theta(\hat{F})$ . For example, when  $\theta$  is a mean and  $\hat{F}$  is an empirical distribution we have:

$$\theta = \theta(F) = \int x dF(x) \quad \text{and} \quad \hat{\theta} = \theta(\hat{F}) = \int x d\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n x_i. \quad (4.39)$$

Thus, quantities of interest should be viewed as functionals of distributions, or as outcomes of procedures or algorithms applied to distributions.

The second core idea of the bootstrap is resampling, whereby difficult analytical calculations are replaced by simulations. Resampling comes in two versions, *parametric* and *non-parametric*. In the parametric version it is assumed that the distribution  $F$  of the data is known up to some parameter  $\psi$ . An estimate of  $\psi$ , typically the MLE, is then substituted in the expression for  $F$  in order to allow the generation of random data samples. In non-parametric resampling no assumption is made about the form of  $F$ . Instead, the data  $x_1, \dots, x_n$  themselves are used to approximate statistical fluctuations according to  $F$ . This is done by *resampling with replacement* from the set  $\{x_1, \dots, x_n\}$ : for each resample,  $n$  data points are successively selected at random from this set, and each selected data point is ‘put back’ in the set before selecting the next one. Thus, some of the original data points will appear more than once in a resampled dataset, and some will not appear at all. A (parametric or non-parametric) resampled dataset is often called a *bootstrap sample*.

There exists a bewildering array of bootstrap methods for computing confidence intervals [17]. These methods can be broadly classified in three categories: pivotal, non-pivotal, and test inversion. Section 4.3.5.1 discusses the bootstrap- $t$  interval as an example of pivotal methods, and makes the important point that the best way to improve the theoretical coverage accuracy of an interval is to bootstrap a pivot (or an asymptotic pivot). Unfortunately theoretical coverage accuracy is not everything, and other important considerations lead to the definition of the non-pivotal percentile intervals in Section 4.3.5.2, first in a ‘simple’ version and then an improved, ‘automatic’ version that incorporates a test-inversion technique. Finally, Section 4.3.5.3 describes a calibration procedure that can improve the coverage accuracy of any confidence interval construction.

#### 4.3.5.1 The Bootstrap- $t$ Interval

If we have a dataset  $\{x_1, \dots, x_n\}$  from which we can derive an estimate  $\hat{\theta}$  of the parameter of interest  $\theta$ , as well as an estimate  $\hat{\sigma}$  of the standard deviation of  $\hat{\theta}$ ,

then we can form a  $(1 - \alpha)$  CL standard interval for  $\theta$ ,

$$\left[ \hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{\sigma}, \hat{\theta} - z_{\frac{\alpha}{2}} \hat{\sigma} \right], \quad (4.40)$$

where  $z_\gamma$  is the  $\gamma$ -quantile of the standard normal distribution. If  $\hat{\theta}$  is asymptotically normal, and the estimators  $\hat{\theta}$  and  $\hat{\sigma}$  are *consistent*, then the asymptotic coverage of the standard interval is  $1 - \alpha$ . In finite samples the actual coverage,

$$P \left[ \hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{\sigma} \leq \theta \leq \hat{\theta} - z_{\frac{\alpha}{2}} \hat{\sigma} \right] = P \left[ z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta} - \theta}{\hat{\sigma}} \leq z_{1-\frac{\alpha}{2}} \right], \quad (4.41)$$

typically differs from  $1 - \alpha$  by a term of order  $n^{-1}$ , where  $n$  is the sample size. The above expression suggests that one way to reduce this difference would be to correct the  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}}$  coefficients by bootstrapping the quantity

$$t \equiv \frac{\hat{\theta} - \theta}{\hat{\sigma}}. \quad (4.42)$$

The idea is to simulate the distribution of  $t$  and replace  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}}$  by the corresponding quantiles of this  $t$  distribution. Since we do not know the true value of  $\theta$ , we need to apply the plug-in principle: replace  $\theta$  by its estimate  $\hat{\theta}$ , and  $\hat{\theta}$  and  $\hat{\sigma}$  by their bootstrapped estimates  $\hat{\theta}^*$  and  $\hat{\sigma}^*$  (the  $*$  superscript is a conventional way to indicate a bootstrapped quantity). The following pseudo-code illustrates the calculation:

1. Obtain  $\hat{\theta} = \theta(\hat{F})$  and  $\hat{\sigma} = \sigma(\hat{F})$  from the original dataset  $\{x_1, \dots, x_n\}$ .
2. For  $i = 1$  to  $b$ :
3. Generate  $\{x_{1i}^*, \dots, x_{ni}^*\}$  from  $\hat{F}$  to obtain  $\hat{F}_i^*$ .
4. Compute  $\hat{\theta}_i^* = \theta(\hat{F}_i^*)$  and  $\hat{\sigma}_i^* = \sigma(\hat{F}_i^*)$ .
5. Set  $t_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{\hat{\sigma}_i^*}$ .
6. Estimate the bootstrap quantiles  $t_{[\frac{\alpha}{2}]}^*$  and  $t_{[1-\frac{\alpha}{2}]}^*$  from the sample of  $t_i^*$ .

The  $(1 - \alpha)$  CL bootstrap- $t$  interval for  $\theta$  is then given by

$$\left[ \hat{\theta} - t_{[1-\frac{\alpha}{2}]}^* \hat{\sigma}, \hat{\theta} - t_{[\frac{\alpha}{2}]}^* \hat{\sigma} \right]. \quad (4.43)$$

The quantiles  $t_{[\gamma]}^*$  at step 6 can be estimated by taking  $t_{[\gamma]}^* = t_{(k)}^*$ , where  $k = \gamma b$  and  $t_{(k)}^*$  is entry number  $k$  in the list of sorted bootstrap values  $t_{(1)}^* \leq t_{(2)}^* \leq \dots \leq t_{(b)}^*$ . If  $k$  is not integer, a linear interpolation can be used:

$$t_{(k)}^* = t_{(k')}^* + (k - k') \left( t_{(k'+1)}^* - t_{(k')}^* \right). \quad (4.44)$$

Here  $k'$  is the largest integer smaller than  $k$ . An appropriate value for the number of bootstrap replications  $b$  is typically 1000 for confidence interval estimation. Note that, unlike the standard interval (4.40), the bootstrap- $t$  interval (4.43) is not

necessarily symmetric around  $\hat{\theta}$ . This asymmetry contributes to the better coverage of the bootstrap- $t$  interval, which typically differs from nominal by a term of order  $n^{-2}$ . Although this constitutes a theoretical improvement with respect to the standard interval, it is important to distinguish theoretical from numerical accuracy. In particular, if the estimated standard deviation  $\hat{\sigma}$  in (4.42) has itself a large variance, the actual numerical accuracy of the interval (4.43) may not be much better than that of the standard interval (4.40). Furthermore, the bootstrap- $t$  interval is primarily designed to work for location parameters such as the mean or median of a sample; it does not work well for parameters such as a standard deviation or correlation coefficient. This is because of the form of the bootstrapped quantity (4.42), which is a pivot for location parameters but not for scale parameters or correlations. The bootstrap- $t$  interval shares a couple of other disadvantages with the standard interval: it does not respect physical boundaries and is not equivariant under parameter transformations. For all these reasons we now turn to percentile intervals.

#### 4.3.5.2 Percentile Intervals

The endpoints of the standard interval (4.40) can be reinterpreted in terms of percentiles of the distribution of the bootstrap estimates  $\hat{\theta}_i^*$ . Indeed, under the conditions of validity of that interval, the  $\hat{\theta}_i^*$  are normal with mean  $\hat{\theta}$  and standard deviation  $\hat{\sigma}$ , so that

$$\begin{aligned} P\left[\hat{\theta}^* \leq \hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{\sigma}\right] &= P\left[\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}} \leq -z_{1-\frac{\alpha}{2}}\right] \\ &= P\left[\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}} \leq z_{\frac{\alpha}{2}}\right] = \frac{\alpha}{2}. \end{aligned} \quad (4.45)$$

This suggests the following definition of a  $(1 - \alpha)$  CL bootstrap interval for  $\theta$ :

$$\left[\hat{\theta}_{[\frac{\alpha}{2}]}^*, \hat{\theta}_{[1-\frac{\alpha}{2}]}^*\right], \quad (4.46)$$

where  $\hat{\theta}_{[\gamma]}^*$  is the  $\gamma$ -quantile of the distribution of bootstrap estimates  $\hat{\theta}_i^*$ . This interval is known as the *simple percentile interval*. Its endpoints are quantiles, making it equivariant under parameter transformations. Thus, if  $\hat{\theta}$  itself is not distributed according to a Gaussian, but a transformation to a Gaussian exists, the percentile method will be able to take advantage of this in producing an interval with accurate coverage. Another advantage of simple percentile intervals is that they respect physical boundaries on the parameter provided the estimator does so. On the other hand, the construction of this interval is based on the quantity  $\hat{\theta}$ , which is generally not a pivot, not even asymptotically. Therefore, its coverage accuracy, outside of the special case just mentioned, is not better than that of the standard interval, of order  $n^{-1}$ .

Several methods have been developed to improve the coverage properties of simple percentile intervals [7], some of them requiring non-trivial analytical calculations. Here we focus on one method, known as the *automatic percentile bootstrap*

because it does not require such calculations [18]. Suppose that  $F(\hat{\theta}; \theta)$  is the cumulative distribution of the plug-in estimate  $\hat{\theta}$ . Having observed  $\hat{\theta} = \hat{\theta}_{\text{obs}}$ , an exact,  $(1 - \alpha)$  CL equal-tailed interval  $[\theta_1, \theta_2]$  for  $\theta$  can be obtained by solving the equations

$$F(\hat{\theta}_{\text{obs}}; \theta_1) = 1 - \frac{\alpha}{2} \quad \text{and} \quad F(\hat{\theta}_{\text{obs}}; \theta_2) = \frac{\alpha}{2} \quad (4.47)$$

(see Table 4.1 in Section 4.3.1.3). In the automatic percentile method the solution to these equations is approximated with a bootstrap simulation. Taking the first equation as an example, one chooses a starting value for  $\theta_1$ , bootstraps the corresponding distribution of  $\hat{\theta}$ , computes its  $(1 - \frac{\alpha}{2})$ -quantile and adjusts  $\theta_1$  until that quantile equals  $\hat{\theta}_{\text{obs}}$ . A good starting value for  $\theta_1$  would be the output of the simple percentile algorithm. If the cdf  $F$  depends on nuisance parameters  $v$ , the latter should be replaced by their profile likelihood estimate  $\hat{\nu}(\theta)$  when performing the bootstrap (as in Section 4.3.4).

As defined above, the automatic percentile interval is equivariant under reparameterisation, respects physical boundaries provided  $\hat{\theta}$  does and has the same coverage accuracy as bootstrap- $t$  intervals, that is  $\mathcal{O}(n^{-2})$ .

#### 4.3.5.3 Bootstrap Calibration

The bootstrap can be used to recalibrate approximate interval constructions. In the case of an upper limit, for example, one first generates a large number of bootstrap samples in order to estimate the *calibration function*, which is the actual coverage  $1 - \alpha_{\text{true}}$  of the upper limit as a function of its nominal coverage  $1 - \alpha_{\text{nom}}$  (the same set of bootstrap samples is used at each  $1 - \alpha_{\text{nom}}$  value). The recalibrated upper limit is then the upper limit computed with the  $1 - \alpha_{\text{nom}}$  corresponding to the desired  $1 - \alpha_{\text{true}}$ . However, this is still an approximation since the calibration function was determined by a bootstrap method. In principle one could recalibrate the calibrated limit and obtain an even better result, but such calculations quickly become very complex.

Typical candidates for recalibration are the standard interval (4.40) and the percentile interval (4.46). In the latter case the recalibration procedure amounts to a double bootstrap.

#### 4.3.6 Nuisance Parameters

In principle the Neyman construction can be performed when there is more than one parameter; it simply becomes a multi-dimensional construction, and the confidence belt becomes a ‘hyperbelt’. If some parameters are nuisances, they can be eliminated by projecting the final confidence region onto the parameter(s) of interest at the end of the construction. However, there are two difficulties: the conceptual one of designing an ordering rule that minimises the amount of overcoverage introduced by the projection [19], and the more practical one of performing multi-dimensional constructions.

Several simpler, approximate solutions are available. We already discussed two of them: asymptotic approximations in Section 4.3.4 and the bootstrap in Section 4.3.5. A third approach inverts the order of the steps in the multi-dimensional Neyman construction: first eliminate the nuisance parameters  $\nu$  from the pdf  $f(x; \theta, \nu)$  of the data  $x$  and then perform a one-dimensional interval construction on the parameter of interest  $\theta$ . The elimination step can be done by integration over a proper prior distribution  $\pi(\nu)$ :

$$f(x; \theta, \nu) \rightarrow f^\dagger(x; \theta) \equiv \int f(x; \theta, \nu) \pi(\nu) d\nu. \quad (4.48)$$

Although this is clearly a Bayesian step, nothing prevents one from studying the frequentist properties of intervals derived from  $f^\dagger$  [20, 21].

Another possibility is to eliminate the nuisance parameters by profiling the pdf:

$$f(x; \theta, \nu) \rightarrow f^*(x; \theta) \equiv f(x; \theta, \hat{\nu}(\theta)). \quad (4.49)$$

Here  $\hat{\nu}(\theta)$  is the profile MLE of  $\nu$ , which maximises  $f(x; \theta, \nu)$  at the observed value of  $x$  and at the given value of  $\theta$ . Even though  $\hat{\nu}(\theta)$  depends on the data, the interval for  $\theta$  is constructed under the assumption that, for a given  $\theta$ , the true value of  $\nu$  is known and equal to  $\hat{\nu}(\theta)$ . In other words,  $f^*(x; \theta)$  is treated as a properly normalised pdf for  $x$  [22, 23].

It is important to keep in mind that the coverage of the simpler solutions is not guaranteed. It must therefore be checked, at least at a few representative points of parameter space (in both  $\theta$  and  $\nu$ ).

To illustrate various techniques for handling nuisance parameters we consider a slight generalisation of the background-subtraction problem analysed in Section 4.3.3.4. We have measured a number of events  $n$  that follows a Poisson distribution with mean  $\theta + \nu$ , where  $\theta$  is a signal of interest and  $\nu$  a background contamination. Neither  $\theta$  nor  $\nu$  is known, but we have an auxiliary measurement of  $\nu$  in the form of a Poisson-distributed number of events  $k$ , with mean  $\tau\nu$ , where  $\tau$  is a known constant. The joint probability mass function of  $n$  and  $k$  is

$$f(n, k; \theta, \nu) = f_1(n; \theta, \nu) f_2(k; \nu) = \frac{(\theta + \nu)^n e^{-\theta-\nu}}{n!} \frac{(\tau\nu)^k e^{-\tau\nu}}{k!}. \quad (4.50)$$

The likelihood ratio for testing a given value of  $\theta$  is

$$\lambda(n, k; \theta) = \frac{f(n, k; \theta, \hat{\nu}(\theta))}{f(n, k; \hat{\theta}, \hat{\nu})}, \quad (4.51)$$

where  $(\hat{\theta}, \hat{\nu})$  is the MLE of  $(\theta, \nu)$  and  $\hat{\nu}(\theta)$  is the profile MLE of  $\nu$ . All these MLEs are constrained to be positive for physical reasons. Assuming we have observed  $n = n_{\text{obs}}$  and  $k = k_{\text{obs}}$ , from which we can derive estimates  $\hat{\theta}_{\text{obs}}$ ,  $\hat{\nu}_{\text{obs}}$ , and  $\hat{\nu}_{\text{obs}}(\theta)$ , one can consider the following eight methods for constructing a  $(1 - \alpha)$  CL interval for  $\theta$ :

- a) *Likelihood-ratio test inversion:* This is an ‘exact’ frequentist method, in the sense that it never undercovers. The interval is defined as the set of  $\theta$  values for which

$$\min_{\nu} \{ P[-2 \ln \lambda(N, K; \theta) \leq -2 \ln \lambda(n_{\text{obs}}, k_{\text{obs}}; \theta) | \theta, \nu] \} \leq 1-\alpha . \quad (4.52)$$

The notation  $P[E|\theta, \nu]$  indicates the probability of event  $E$  when  $f(n, k; \theta, \nu)$  is the true distribution of  $(N, K)$ . With  $q_{1-\alpha}(\theta, \nu)$  the  $(1-\alpha)$ -quantile of the distribution of  $-2 \ln \lambda(N, K; \theta)$ , (4.52) is equivalent to:

$$-2 \ln \lambda(n_{\text{obs}}, k_{\text{obs}}; \theta) \leq q_{1-\alpha}(\theta) \equiv \max_{\nu} q_{1-\alpha}(\theta, \nu) . \quad (4.53)$$

The minimisation and maximisation in these interval definitions can be viewed as a kind of worst-case analysis that guarantees frequentist coverage, and possibly overcoverage, for all physical values of  $\nu$  at a given  $\theta$ .

- b) *Naive method:* Given the relative computational complexity of the test-inversion method, it may be worthwhile to compare it to the following simple approach. Under model (4.50), the MLE of  $\theta$  is  $\hat{\theta}_{\text{obs}} = n_{\text{obs}} - k_{\text{obs}}/\tau$ . Ignoring the physical constraint  $\hat{\theta}_{\text{obs}} \geq 0$ , the variance of this MLE is  $\theta + \nu + \nu/\tau$ , which can be estimated by  $n_{\text{obs}} + k_{\text{obs}}/\tau^2$ . Thus, an approximate  $(1-\alpha)$  CL interval for  $\theta$  is given by the intersection of

$$\left[ \hat{\theta}_{\text{obs}} - z_{1-\frac{\alpha}{2}} \sqrt{n_{\text{obs}} + \frac{k_{\text{obs}}}{\tau^2}}, \hat{\theta}_{\text{obs}} + z_{1-\frac{\alpha}{2}} \sqrt{n_{\text{obs}} + \frac{k_{\text{obs}}}{\tau^2}} \right] , \quad (4.54)$$

with the physical region  $\theta \geq 0$ , where  $z_{\gamma}$  is the  $\gamma$ -quantile of the standard normal distribution.

- c) *Asymptotic likelihood-ratio test:* This is the method described in Section 4.3.4; a test inversion as in (4.53), but with  $q_{1-\alpha}(\theta)$  approximated by the  $(1-\alpha)$ -quantile  $\chi^2_{1,1-\alpha}$  of a  $\chi^2$  distribution for one degree of freedom.  
d) *Bayesian elimination:* Here the auxiliary measurement  $f_2(k; \nu)$  is replaced by a prior  $\pi(\nu)$  for  $\nu$ . With proper normalisation this is:

$$\pi(\nu) = \frac{\tau(\tau\nu)^k e^{-\tau\nu}}{\Gamma(k+1)} , \quad (4.55)$$

and  $f_1(n; \theta, \nu)$  is integrated over  $\pi(\nu)$  to obtain a distribution of  $n$  that depends on  $\theta$  only:

$$f^\dagger(n; \theta) = \int f_1(n; \theta, \nu) \pi(\nu) d\nu . \quad (4.56)$$

The likelihood ratio is now

$$\lambda^\dagger(n_{\text{obs}}; \theta) = \frac{f^\dagger(n_{\text{obs}}; \theta)}{f^\dagger(n_{\text{obs}}; \hat{\theta})} , \quad (4.57)$$

where  $\hat{\theta}$  maximises  $f^\dagger$  at the observed value  $n_{\text{obs}}$  of  $N$ . One then obtains the  $(1-\alpha)$ -quantile  $q_{\text{Bayes}, 1-\alpha}(\theta)$  of the distribution of  $-2 \ln \lambda^\dagger(N; \theta)$  under  $f^\dagger(n; \theta)$ , and  $\theta$  values for which  $-2 \ln \lambda^\dagger(n_{\text{obs}}; \theta) \leq q_{\text{Bayes}, 1-\alpha}(\theta)$  form the desired interval.

- e) *Simple percentile*: This is the bootstrap method of Section 4.3.5.2. Intervals are computed from the  $\frac{\alpha}{2}$ - and  $(1 - \frac{\alpha}{2})$ -quantiles of the distribution of the estimator  $\hat{\theta} = \max(N - K/\tau, 0)$ . This distribution is derived from  $f(n, k; \hat{\theta}_{\text{obs}}, \hat{\nu}_{\text{obs}})$ , where  $\hat{\theta}_{\text{obs}}$  and  $\hat{\nu}_{\text{obs}}$  are determined from  $n_{\text{obs}}$  and  $k_{\text{obs}}$ .
- f) *Automatic percentile*: This is the second method described in Section 4.3.5.2. Let  $G(\hat{\theta}; \theta, \nu)$  be the cumulative distribution of the estimate  $\hat{\theta}$ . The interval endpoints  $\theta_1$  and  $\theta_2$  are the solutions of  $G(\hat{\theta}_{\text{obs}}; \theta_1, \hat{\nu}(\theta_1)) = 1 - \frac{\alpha}{2}$  and  $G(\hat{\theta}_{\text{obs}}; \theta_2, \hat{\nu}(\theta_2)) = \frac{\alpha}{2}$ .
- g) *Likelihood-ratio bootstrap*: Again a test inversion as in (4.52), but instead of minimising the likelihood-ratio tail probability with respect to  $\nu$ , it is evaluated at  $\nu = \hat{\nu}_{\text{obs}}$ . Thus, the confidence region is defined as the set of  $\theta$  values for which

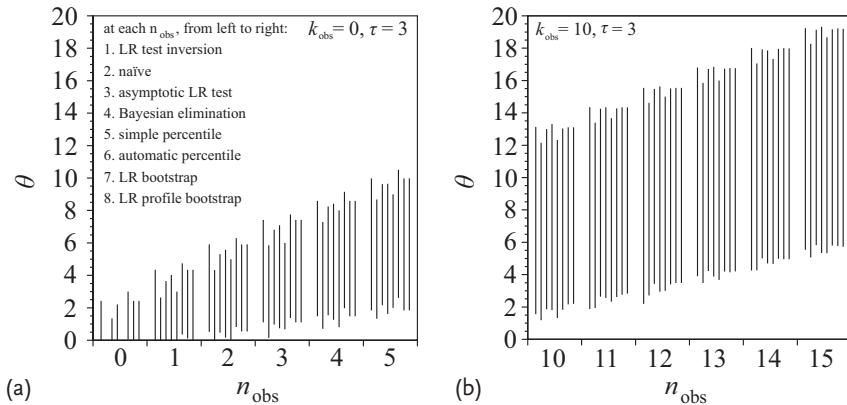
$$P[-2 \ln \lambda(N, K; \theta) \leq -2 \ln \lambda(n_{\text{obs}}, k_{\text{obs}}; \theta) | \theta, \hat{\nu}_{\text{obs}}] \leq 1 - \alpha. \quad (4.58)$$

If we write  $q_{\text{bootstrap}, 1-\alpha}(\theta, \hat{\nu}_{\text{obs}})$  for the  $(1 - \alpha)$ -quantile of the distribution of  $-2 \ln \lambda(N, K; \theta)$  under  $f(n, k; \theta, \hat{\nu}_{\text{obs}})$ , the above inequality is equivalent to

$$-2 \ln \lambda(n_{\text{obs}}, k_{\text{obs}}; \theta) \leq q_{\text{bootstrap}, 1-\alpha}(\theta, \hat{\nu}_{\text{obs}}). \quad (4.59)$$

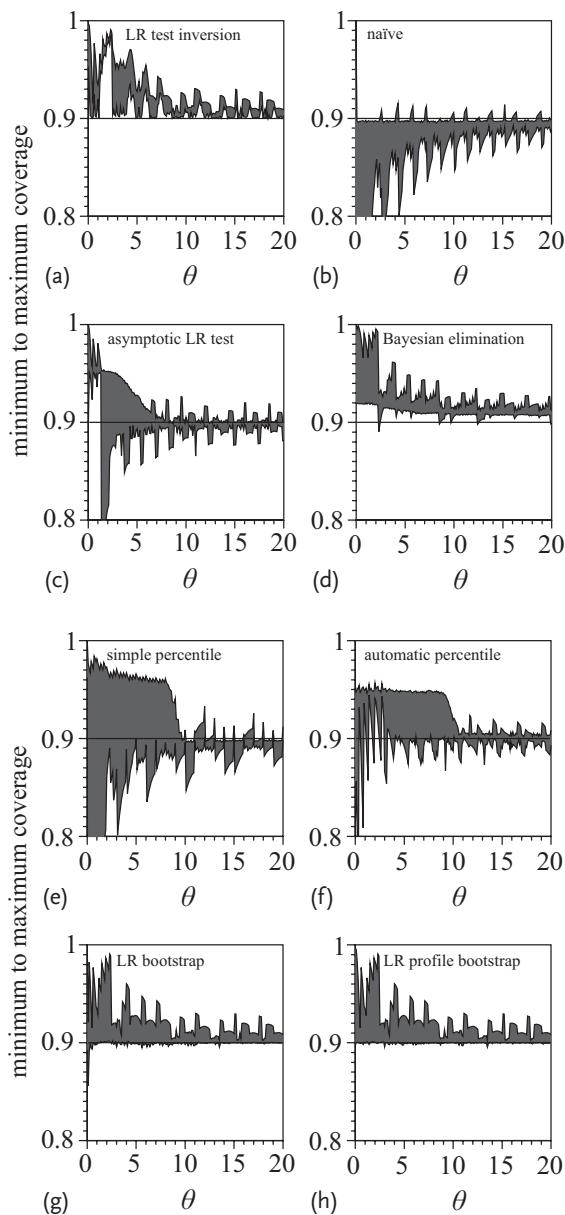
- h) *Likelihood-ratio profile bootstrap*: This is a variation on the previous method, where  $\hat{\nu}_{\text{obs}}$  is replaced by  $\hat{\nu}_{\text{obs}}(\theta)$  in (4.58) and (4.59). It essentially corresponds to the profile method of (4.49).

These interval constructions are compared in Figure 4.4 for the cases  $\tau = 3$  with  $k_{\text{obs}} = 0$  (a) and  $k_{\text{obs}} = 10$  (b), and for several values of  $n_{\text{obs}}$ . Differences between



**Figure 4.4** Interval constructions for the signal rate  $\theta$  in a Poisson signal plus background problem;  $n_{\text{obs}}$  and  $k_{\text{obs}}$  are the numbers of events observed in the signal + background and background-only measurements, respectively, and  $\tau$  is the known ratio of mean backgrounds between the two

measurements. (a) The case  $k_{\text{obs}} = 0$ ; (b) the case  $k_{\text{obs}} = 10$ . At each value of  $n_{\text{obs}}$ , eight vertical line segments represent 90% CL intervals for  $\theta$  according to the methods listed in the legend. A missing line segment indicates that the corresponding interval is empty or collapsed to the singleton  $\{0\}$ .



**Figure 4.5** (a–h) Bands showing the frequentist coverage versus  $\theta$  of eight interval constructions for the Poisson signal plus background problem discussed in the text. The parameter  $\tau$  is set to 3. The boundaries of the

bands indicate the minimum and maximum coverage obtained by varying  $\nu$  in the range  $[0, 20]$ . The 90% nominal coverage is indicated by a horizontal line in each plot.

the methods are particularly pronounced at low  $n_{\text{obs}}$  and  $k_{\text{obs}}$ . In particular for  $n_{\text{obs}} = k_{\text{obs}} = 0$  the naive and simple percentile intervals have zero length. In the latter case this is due to the use of MLEs to form the bootstrap distribution  $f(n, k; \hat{\theta}_{\text{obs}}, \hat{\nu}_{\text{obs}})$ , which is degenerate when the parameter estimates are both zero. In principle this could be remedied by choosing different estimators. At low  $n_{\text{obs}}$  values, the naive interval has the additional problem of extending into the unphysical region, resulting in unreasonably tight constraints on  $\theta$ . This interval can certainly not be recommended. Among the other constructions, one notes that the asymptotic interval tends to be systematically shorter than the exact one, whereas the likelihood-ratio bootstrap and profile bootstrap intervals are often in good agreement with the latter. The performance of these methods is perhaps more easily judged by examining their frequentist coverage. One can plot the coverage as a function of  $\theta$  at a fixed value of  $\nu$ , or make a two-dimensional plot of coverage versus  $\theta$  and  $\nu$ , or plot as a function of  $\theta$  the minimum and maximum coverages obtained when  $\nu$  varies over a given range. Figure 4.5 shows the latter option, for  $0 \leq \nu \leq 20$ . As expected, the exact test-inversion method never undercovers, but it can substantially overcover. The naive, asymptotic, and simple percentile methods all have significant undercoverage at low values of  $\theta$ . Conversely, the likelihood-ratio bootstrap and profile bootstrap methods both perform remarkably well. For all methods the coverage tends to improve as  $\theta$  increases. This conforms with the expectation that these methods should perform well in the large sample limit, which for Poisson processes is attained as the mean  $\theta$  goes to infinity.

#### 4.4 Bayesian Methods

As already emphasised in the introductory Section 4.1, the output of a Bayesian analysis is *always* the complete posterior distribution for the parameter(s) of interest. However, it is often useful to summarise the posterior by quoting a region with a given probability content. Such a region can be an interval or a union of intervals. Several schemes, or ‘ordering rules’, are available:

- *Highest-posterior-density regions (HPD)*: Any parameter value inside such a region has a higher posterior probability density than any parameter value outside the region, guaranteeing that the region will have the smallest possible length (or volume). Unfortunately this construction is not invariant under reparameterisations, and as Example 4.4 will show, this lack of invariance can result in poor frequentist coverage for some parameter values (of course this will only be of concern to a frequentist or an objective Bayesian).
- *Equal-tailed intervals*: These are intervals that span equal posterior probabilities on each side of the posterior median. For example, a 68% equal-tailed interval extends from the 16th to the 84th posterior percentiles. These intervals are equivariant under one-to-one reparameterisations that are continuous from the

left.<sup>6</sup> However, they typically only make sense when the posterior is unimodal,<sup>7</sup> and their generalisation to multi-dimensional parameters is non-trivial. Furthermore, if a parameter is constrained to be non-negative, an equal-tailed interval will usually not include the value zero (an exception may occur if the posterior has a substantial probability mass at zero); this may be problematic if zero is a value of special physical significance.

- *Upper and lower limits:* For one-dimensional posterior distributions, these one-sided intervals can be defined using percentiles.
- *Likelihood regions:* These are standard likelihood contours, that is regions of parameter values for which the likelihood is larger than for any parameter value outside the region. The size of the region is determined by the desired posterior or credibility. Such regions are metric independent and robust with respect to the choice of prior [24]. In one-dimensional problems with physical boundaries and unimodal likelihoods this construction yields intervals that have a smooth transition from one-sided to two-sided.
- *Lowest posterior loss regions:* A more foundational approach to Bayesian interval construction starts with a loss structure [25]. Suppose that we can in some way quantify the loss  $\ell\{\theta_0, \theta\}$  incurred by using the parameter value  $\theta_0$  when the true value is  $\theta$ . After having observed data  $x$ , our posterior expected loss is

$$l\{\theta_0; x\} = \int \ell\{\theta_0, \theta\} p(\theta|x) d\theta , \quad (4.60)$$

where  $p(\theta|x)$  is the posterior density. A natural point estimate of  $\theta$  is then the value that minimises this posterior expected loss, and a natural credible region is the set of  $\theta$  values for which the posterior expected loss is smaller than for any value outside the set, subject to a credibility constraint. A possible choice of loss function is the quadratic loss  $-\ell\{\theta_0, \theta\} = (\theta_0 - \theta)^2$  – which yields the posterior mean as a point estimate. Another choice is zero-one loss  $-\ell\{\theta_0, \theta\} = 0$  if  $|\theta_0 - \theta| \leq \epsilon$ , and  $\ell\{\theta_0, \theta\} = 1$  otherwise, where  $\epsilon$  is a constant. As  $\epsilon$  goes to zero this loss function leads to the posterior mode as a point estimate and to credibility regions that have highest posterior density. Many more loss functions can be devised, but in the absence of any subjective preference, information-theoretic arguments lead to the concept of *intrinsic discrepancy loss* [26], which is defined as the symmetrised *Kullback–Leibler divergence* between the model indexed by  $\theta_0$  and that indexed by  $\theta$ :

$$\delta\{\theta_0, \theta\} = \min \{\kappa\{p(x; \theta_0); p(x; \theta)\}, \kappa\{p(x; \theta); p(x; \theta_0)\}\} , \quad (4.61)$$

with

$$\kappa\{p(x; \theta_0); p(x; \theta)\} = \int p(x; \theta) \ln \frac{p(x; \theta)}{p(x; \theta_0)} dx \quad (4.62)$$

(for discrete sample spaces the integral is replaced by a sum). From this definition it follows that the intrinsic discrepancy loss between two models can be

6) A reparameterisation  $\theta \rightarrow \eta(\theta)$  is continuous from the left if  $\lim_{\theta \uparrow \theta_0} \eta(\theta) = \eta(\theta_0)$ .

7) A probability density function with a single maximum is called unimodal.

interpreted as the minimum expected log-likelihood ratio in favour of the model that generated the data. Credible regions derived from this loss function are labelled ‘intrinsic’. They enjoy many useful properties, including equivariance under parameter transformation, and they are available in multi-dimensional settings. A one-dimensional example is given in Example 4.4.

Users of Bayesian procedures are generally advised to assess the sensitivity of their result to the choice of prior. Furthermore, if the prior is of the so-called non-informative variety, the behaviour of the result under repeated sampling (i.e. the frequentist coverage) should also be investigated.

In the context of interval construction, it is worth mentioning that non-informative priors can be designed in such a way that the resulting posterior intervals have a frequentist coverage that matches their Bayesian credibility to some order in  $1/\sqrt{n}$ ,  $n$  being the sample size. When there are no nuisance parameters and the parameter of interest is one-dimensional, the matching prior to  $\mathcal{O}(1/n)$  for one-sided intervals is *Jeffreys’ prior* (see also Section 1.5.3.2):

$$\pi_J(\theta) \propto \sqrt{E\left[-\frac{d^2}{d\theta^2} \ln L(x; \theta)\right]}. \quad (4.63)$$

Frequentist coverage is harder to achieve in higher dimensions, but the *Bayesian reference analysis approach* has obtained good results [25]. This is an objective Bayesian method based on information-theoretic considerations. In spite of being non-subjective, it provides results with a credibility interpretation: such results would be obtained by a person whose prior beliefs have minimal effect, relative to the data, on posterior inferences. An application to cross-section measurements in high energy physics is described in [27].

The next subsection illustrates Bayesian interval constructions with an example that appears simple and yet can lead to serious difficulties if not handled properly. Sections 4.4.1 and 4.4.2 summarise the calculation of Bayesian intervals for binomial efficiencies and Poisson means, thereby complementing the two frequentist interval calculations given in Sections 4.3.3.3 and 4.3.3.4. The Bayesian calculations are based on Jeffreys’ prior, and the resulting intervals are therefore known as Jeffreys intervals.

#### Example 4.4 Measuring track momenta

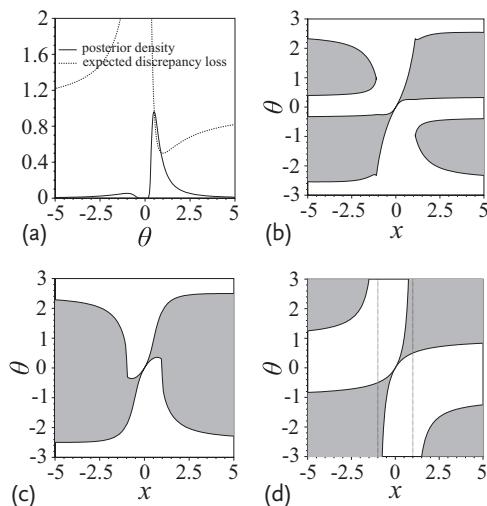
Consider the measurement of particle transverse momenta in a tracking chamber immersed in a solenoidal magnetic field. A simple model is that for a given particle the charge-signed transverse momentum is the inverse of the radius of curvature  $\rho$  of its track, and that the measured curvature radius has a Gaussian distribution with standard deviation  $\sigma$  proportional to the chamber resolution and inversely proportional to the magnetic field strength. Thus, if  $x$  is the measured transverse momentum and  $\theta$  its true value, the likelihood function has the form:

$$L(x; \theta) \propto e^{-\frac{1}{2}\left(\frac{1/x - 1/\theta}{\sigma}\right)^2}. \quad (4.64)$$

A straightforward calculation shows that Jeffreys' prior is proportional to  $1/\theta^2$ . The properly normalised posterior is therefore

$$p(\theta|x) = \frac{e^{-\frac{1}{2}(\frac{1/x-1/\theta}{\sigma})^2}}{\sqrt{2\pi}\sigma\theta^2}. \quad (4.65)$$

This posterior is shown in Figure 4.6a for the case  $\sigma = 1$  and  $x = 1$ . There are two local maxima, at  $\theta_{\pm} = (-1 \pm \sqrt{1 + 8x^2\sigma^2})/(4x\sigma^2)$ , corresponding to two possible charge assignments to the observed track. As  $|x| \rightarrow \infty$ , the posterior density reaches equal heights at these maxima, reflecting the ambiguity in charge determination at large momenta. However, the posterior mode is a very biased estimate of  $\theta$  since  $|\theta_{\pm}|$  never exceeds  $1/(\sqrt{2}\sigma)$ . Highest-posterior-density (HPD) credible regions are shown in Figure 4.6b: they consist of a single interval at low  $|x|$ , and of the union of two intervals at large  $|x|$ . At large  $|x|$  the credibility 'belt' (i.e. the set of credible regions viewed in  $(x, \theta)$  space) consists of two horizontal bands that are bounded away from large  $\theta$  values. As a result, the frequentist coverage of HPD regions is zero at large  $|\theta|$ ! This may surprise in view of the facts that the posterior (4.65) for the transverse momentum  $\theta$  can be derived from a Gaussian posterior for the curvature radius  $\rho$  via the transformation  $\rho \rightarrow \theta = 1/\rho$ , and that HPD intervals for a Gaussian posterior have exact frequentist coverage. The problem, of course, is that HPD intervals are not equivariant under reparameterisation. This suggests



**Figure 4.6** Credible region construction for a transverse momentum with true value  $\theta$  and observed value  $x$  in a tracking chamber immersed in a magnetic field. (a) The posterior density for  $x = 1$  (solid line), superimposed on the posterior expected intrinsic loss function (dashes). (b–d): The 68% credible intervals (shaded regions) in  $\theta$  as a function of  $x$ ,

for highest posterior density, equal-tails, and lowest intrinsic loss constructions, respectively. The vertical dotted lines in (d) are asymptotes of the boundaries of the credibility belt. For  $x$  values between these lines the credible region is a single interval; outside these lines it is the union of two open intervals (4.66).

a simple solution, which is to construct an HPD interval for  $\rho$  and to invert the endpoints to obtain a credible region for  $\theta$ , taking care of the case where the  $\rho$  interval contains zero. Applying this idea to the 68% credible interval  $[1/x - \sigma, 1/x + \sigma]$  for  $\rho$  leads to the following credible region for  $\theta$ :

$$\begin{aligned} & \left[ \frac{1}{1/x + \sigma}, \frac{1}{1/x - \sigma} \right] && \text{if } |x| < 1/\sigma, \quad \text{and} \\ & \left[ -\infty, \frac{1}{1/x - \sigma} \right] \cup \left[ \frac{1}{1/x + \sigma}, +\infty \right] && \text{if } |x| > 1/\sigma. \end{aligned} \quad (4.66)$$

This is not an HPD region in the  $\theta$  parameterisation, but its coverage is 68%, exactly matching its credibility.

Figure 4.6c shows the 68% credibility belt for Bayesian equal-tailed intervals. Again, the outer contour of the belt becomes horizontal at large  $|x|$ , resulting in zero coverage for large  $\theta$  values. The transformation  $\rho \rightarrow \theta = 1/\rho$  is one-to-one but not continuous from the left, explaining why the nice properties of equal-tailed intervals for  $\rho$  do not transfer to  $\theta$ .

Finally we examine intrinsic loss credible regions. The intrinsic discrepancy loss, (4.61), becomes

$$\delta\{\theta_0, \theta\} = \frac{1}{2} \left( \frac{1/\theta_0 - 1/\theta}{\sigma} \right)^2. \quad (4.67)$$

The posterior expected intrinsic discrepancy loss is then

$$d\{\theta_0; x\} = \int \delta\{\theta_0, \theta\} p(\theta|x) d\theta = \frac{1}{2} \left[ 1 + \left( \frac{1/\theta_0 - 1/x}{\sigma} \right)^2 \right], \quad (4.68)$$

and is plotted as a dashed line in Figure 4.6a. The minimum-loss estimate of  $\theta$  is  $x$ , and minimum-loss credible regions are given by (4.66) and shown in Figure 4.6d. In this case the intrinsic discrepancy loss formalism has automatically produced point and interval estimates that correspond to HPD in the curvature radius parameterisation.

#### 4.4.1 Binomial Efficiencies

The binomial likelihood for an efficiency  $\epsilon$ , after having observed  $x$  successes out of  $n$  trials, is

$$L(n, x; \epsilon) = \binom{n}{x} \epsilon^x (1 - \epsilon)^{n-x}, \quad (4.69)$$

from which Jeffreys' prior is found to be

$$\pi_J(\epsilon) \propto \sqrt{\frac{n}{\epsilon(1-\epsilon)}}. \quad (4.70)$$

The properly normalised posterior is a  $\text{Beta}(x + \frac{1}{2}, n - x + \frac{1}{2})$  distribution:

$$p(\epsilon|x) = \frac{\epsilon^{x-\frac{1}{2}}(1-\epsilon)^{n-x-\frac{1}{2}}}{B(x + \frac{1}{2}, n - x + \frac{1}{2})}. \quad (4.71)$$

The endpoints of an equal-tailed,  $(1 - \alpha)$  CL Bayesian interval  $[\epsilon_{\text{low}}, \epsilon_{\text{up}}]$  for  $\epsilon$  are the  $\frac{\alpha}{2}$ - and  $1 - \frac{\alpha}{2}$ -quantiles of this posterior:

$$\epsilon_{\text{low}} = B_{x+\frac{1}{2}, n-x+\frac{1}{2}, \frac{\alpha}{2}} \quad \text{and} \quad \epsilon_{\text{up}} = B_{x+\frac{1}{2}, n-x+\frac{1}{2}, 1-\frac{\alpha}{2}}. \quad (4.72)$$

These can be compared with the frequentist formulæ (4.29) and (4.30). In contrast with Clopper–Pearson intervals, Jeffreys intervals tend to be shorter but do not guarantee exact coverage. The coverage of both constructions oscillates as a function of the true value of  $\epsilon$ . For Clopper–Pearson intervals these oscillations all remain above the nominal confidence level  $1 - \alpha$ , whereas for Jeffreys intervals they straddle  $1 - \alpha$  [28].

#### 4.4.2

##### Poisson Means

For a Poisson-distributed number of events  $n$  the likelihood is

$$L(n; \theta) = \frac{(\theta + \nu)^n e^{-\theta-\nu}}{n!}, \quad (4.73)$$

where, as before,  $\theta$  is the signal strength of interest and  $\nu$  is the level of a known background contamination. Jeffreys' rule (4.63) gives:

$$\pi_j(\theta) \propto \frac{1}{\sqrt{\theta + \nu}}, \quad (4.74)$$

and the corresponding posterior is a shifted Gamma distribution:

$$p(\theta|n) = \frac{(\theta + \nu)^{n-\frac{1}{2}} e^{-\theta-\nu}}{\Gamma(n + \frac{1}{2}) [1 - P(n + \frac{1}{2}, \nu)]},$$

with  $P(a, \nu) \equiv \int_0^\nu \frac{t^{a-1} e^{-t}}{\Gamma(a)} dt.$

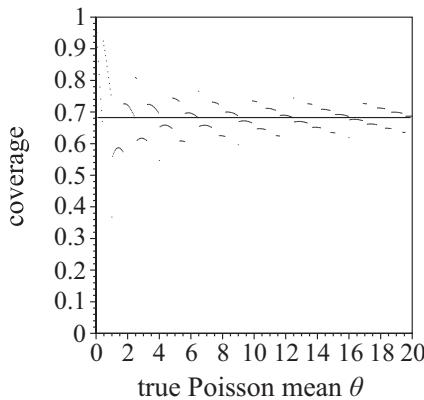
$$(4.75)$$

A  $(1 - \alpha)$  CL upper limit  $\theta_{1-\alpha}$  is given by the  $(1 - \alpha)$ -quantile of this posterior:

$$1 - \alpha = \int_0^{\theta_{1-\alpha}} p(\theta|n) d\theta = \frac{P(n + \frac{1}{2}, \nu + \theta_{1-\alpha}) - P(n + \frac{1}{2}, \nu)}{1 - P(n + \frac{1}{2}, \nu)}$$

$$= \frac{P[\chi^2_{2n+1} \leq 2(\nu + \theta_{1-\alpha})] - P(n + \frac{1}{2}, \nu)}{1 - P(n + \frac{1}{2}, \nu)},$$

$$(4.76)$$



**Figure 4.7** Frequentist coverage of 68% CL Bayesian central intervals for the mean of a Poisson distribution, using Jeffreys' prior. The coverage is evaluated in increments of 0.1

in the value of the true Poisson mean, and the Bayesian credibility of the construction is indicated by the solid horizontal line.

where, similarly to what was done in Section 4.3.3.4, we converted an incomplete Gamma function into the tail probability of a  $\chi^2$  distribution. Solving for the latter yields

$$P \left[ \chi_{2n+1}^2 \leq 2(\nu + \theta_{1-\alpha}) \right] = 1 - \alpha' ,$$

with  $1 - \alpha' \equiv 1 - \alpha + \alpha P(n + \frac{1}{2}, \nu)$ . (4.77)

Hence we find

$$\theta_{1-\alpha} = \frac{1}{2} \chi_{2n+1,1-\alpha'}^2 - \nu , \quad (4.78)$$

which can be compared to (4.34). In contrast with the frequentist result, the Jeffreys upper limit never becomes negative, thanks to the dependence of  $\alpha'$  on  $\nu$ .

Figure 4.7 shows how the coverage of the Jeffreys limit oscillates as a function of the true value of  $\theta$ , with downward oscillations dipping below the Bayesian credibility of the interval. For this reason a flat prior is sometimes preferred, as the resulting coverage oscillations remain above the credibility. For a flat prior the upper limit is given by:

$$\theta_{1-\alpha}^{\text{flat}} = \frac{1}{2} \chi_{2n+2,1-\alpha''}^2 - \nu , \quad \text{where } 1 - \alpha'' \equiv 1 - \alpha + \alpha P(n + 1, \nu) . \quad (4.79)$$

The good coverage properties of the flat prior are only true for upper limits, however; for lower limits and two-sided intervals Jeffreys' rule performs better.

For the Poisson model the intrinsic discrepancy loss is given by

$$\delta\{\theta_0, \theta\} = |\theta_0 - \theta| - [\nu + \min(\theta_0, \theta)] \left| \ln \frac{\nu + \theta_0}{\nu + \theta} \right| , \quad (4.80)$$

**Table 4.3** Bayesian interval constructions for the mean of a Poisson distribution when  $n$  events are observed. All results were obtained with Jeffreys' prior. The ordering rules shown are 95% CL upper limit (column 2), 68% CL equal-tailed interval (column 3), and 95% and 68% CL lowest posterior-expected intrinsic discrepancy loss (columns 4 and 5).

$n$	Bayesian intervals with Jeffreys' prior			
	Upper limit 95% CL	Equal-tailed 68% CL	Lowest post. exp. intr. loss 95% CL	68% CL
0	1.92	[0.02, 0.99]	[0.00, 1.92]	[0.02, 0.91]
1	3.91	[0.42, 2.59]	[0.01, 3.93]	[0.41, 2.58]
2	5.54	[1.03, 3.98]	[0.28, 5.82]	[1.03, 3.97]
3	7.03	[1.72, 5.28]	[0.69, 7.48]	[1.72, 5.28]
4	8.46	[2.46, 6.54]	[1.18, 9.03]	[2.46, 6.54]
5	9.84	[3.23, 7.78]	[1.73, 10.51]	[3.23, 7.77]
6	11.18	[4.02, 8.99]	[2.32, 11.94]	[4.02, 8.98]
7	12.50	[4.82, 10.18]	[2.94, 13.33]	[4.82, 10.18]
8	13.79	[5.64, 11.36]	[3.59, 14.69]	[5.64, 11.36]
9	15.07	[6.47, 12.53]	[4.26, 16.04]	[6.47, 12.53]
10	16.34	[7.31, 13.69]	[4.94, 17.36]	[7.31, 13.69]

and its posterior expectation can be computed numerically. Bayesian intervals derived from this loss function are shown in Table 4.3, together with regular upper limits and equal-tailed intervals. Note that the 95% CL intrinsic interval coincides with the upper limit when zero events are observed. This is due to the fact that in order to obtain a higher credible interval one has to tolerate a higher loss, which eventually becomes larger than the loss at  $\theta = 0$ . At 99% CL for example, the intrinsic interval coincides with the upper limit for  $N = 0, 1$ , and 2. This unification of two-sided intervals and upper limits is reminiscent of Feldman–Cousins intervals in the frequentist case, where it could be used to test a parameter value on the boundary. However, it does not have the same significance here, because the duality between confidence intervals and hypothesis tests only exists in the frequentist paradigm.

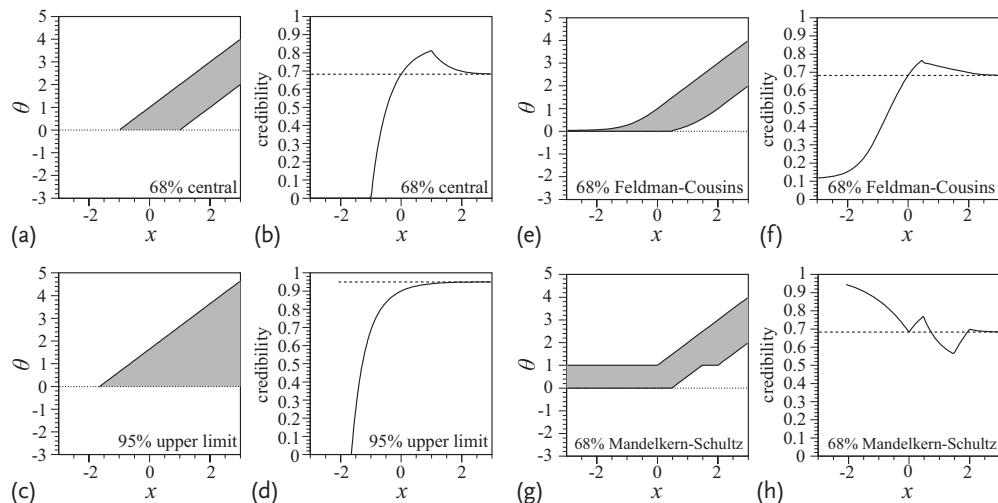
## 4.5

### Graphical Comparison of Interval Constructions

The effect of a physical boundary on frequentist and Bayesian interval constructions is illustrated in Figures 4.8 and 4.9 for the measurement of the mean  $\theta$  of a Gaussian with unit standard deviation. The true mean  $\theta$  is constrained to be positive. All intervals are based on a single observation  $x$ , which can be positive or negative due to resolution effects. This is a simplified model corresponding for example to the measurement of the square of a neutrino mass discussed in [8]. As pointed out in Section 4.2, intervals have many properties that are worth studying: here we only examine the Bayesian credibility of frequentist constructions and the frequentist coverage of Bayesian constructions.

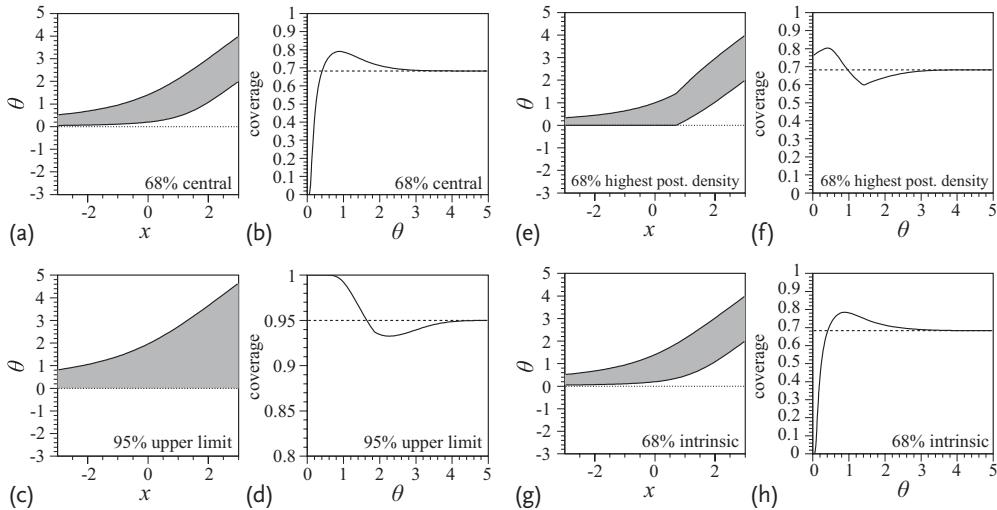
Figure 4.8 shows only frequentist constructions. Because of the positivity constraint on  $\theta$ , the 68% CL equal-tailed (or central) interval (Figure 4.8a) is empty whenever the observation  $x$  is below  $-1$ . For  $x$  between  $-1$  and  $+1$  the interval is an upper limit, and for  $x$  higher than  $+1$  it is two-sided. Since this is an exact frequentist construction, its coverage is 68% for all physical values of  $\theta$ . From a frequentist point of view empty intervals are not meaningless: they simply indicate that no physical value of  $\theta$  can account for the observation *at the stated confidence level*. However, empty intervals have a drastic effect on Bayesian credibility. We can investigate this with the help of Jeffreys' prior, which for this problem is zero for  $\theta < 0$  and a positive constant for  $\theta \geq 0$ . For each value of  $x$  the integral of the posterior density over the corresponding  $\theta$  interval yields the latter's credibility. The result is shown in Figure 4.8b: the credibility vanishes for  $x < -1$ , then rises sharply up to a maximum at  $x = 1$ , and finally for  $x > 2$  it settles down to a value very close to the frequentist coverage.

The remaining pairs of panels in Figure 4.8 are similarly organised, showing a frequentist construction in Figure 4.8a,c,e,g and the corresponding Bayesian credibility in Figure 4.8b,d,f,h. It can be seen that upper limits have the same credibility problem as central intervals. The remaining two frequentist constructions mitigate the credibility problem by avoiding empty intervals. Feldman–Cousins intervals, shown in Figure 4.8e,f, use  $x$  as an estimator of  $\theta$  and are based on a likelihood ratio ordering rule [8]. They still have low credibility for negative  $x$  values. Mandelkern–Schultz intervals, presented in Figure 4.8g,h, use  $\max\{0, x\}$  as an estimator of  $\theta$  and are based on an equal-tails ordering rule [29]. These intervals are the same for any negative  $x$  as for zero  $x$ , resulting in excess credibility at negative  $x$ .



**Figure 4.8** Frequentist interval constructions. Panels (a,c,e,g) show graphs of  $\theta$  versus  $x$ , with dotted lines indicating the lower boundary of the physical region. Panels (b,d,f,h)

show the corresponding Bayesian credibility levels based on Jeffreys' prior, with dashed lines indicating the frequentist coverage.



**Figure 4.9** Bayesian interval constructions. Panels (a,c,e,g) show graphs of  $\theta$  versus  $x$ , with dotted lines indicating the lower boundary of the physical region. Panels (b,d,f,h) show the corresponding frequentist coverage levels, with dashed lines indicating the Bayesian credibility.

Figure 4.9 shows four Bayesian constructions in paired panels. Figure 4.9a,c,e,g show the credibility belt and Figure 4.9b,d,f,h the corresponding frequentist coverage. All constructions use Jeffreys' prior for  $\theta$  and differ only by the ordering rule used. Panel pairs (a) and (b), (c) and (d), and (e) and (f) use equal-tailed, upper limit and highest posterior density ordering, respectively. On panel (g) and (h) the ordering is according to the intrinsic discrepancy loss, which for this problem equals  $\delta\{\theta_0, \theta\} = (\theta - \theta_0)^2/2$  and coincides with quadratic loss. All four constructions have reasonable frequentist coverage, except near  $\theta = 0$ , where the curves for equal-tailed and intrinsic intervals dip to zero.

A noteworthy feature of both Figures 4.8 and 4.9 is that frequentist coverage and Bayesian credibility always agree with each other when one is far enough from the physical boundary.

## 4.6

### The Role of Intervals in Search Procedures

Suppose that we are using a collider experiment to search for a new particle with unknown production rate  $\theta$ . From a statistical point of view this problem can be formulated as a hypothesis test of

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0 , \quad (4.81)$$

since  $H_0$  corresponds to non-existence of the particle and  $H_1$  to its existence. In a frequentist approach we calculate a  $p$ -value  $p_0$  to test  $H_0$ , and claim discovery

if  $p_0 \leq \epsilon$ , where  $\epsilon$  is a pre-specified type I error rate (typically  $2.87 \cdot 10^{-7}$ , corresponding to  $5\sigma$  in the tail of a Gaussian distribution). It is then customary to accompany the discovery claim with point and interval estimates of  $\theta$ , with the interval being two-sided and providing 68% confidence. Conversely, if  $p_0 > \epsilon$ , we fail to reject  $H_0$ . However, this decision does not imply that all values of  $\theta$  under  $H_1$  are now excluded. In particular, there are values of  $\theta$  that the experiment is simply not sensitive to, and values that the data won't allow us to exclude. Hence we need to investigate this more closely: for which values  $\theta_0$  can the hypothesis  $H'_1 : \theta = \theta_0$  be excluded? More precisely, we need to test:

$$H'_1[\theta_0] : \theta = \theta_0 \quad \text{versus} \quad H'_0[\theta_0] : \theta < \theta_0, \quad (4.82)$$

where for later convenience we specified the tested value of  $\theta$  as an argument to the hypotheses. Comparing with (4.3) in Section 4.3.2 shows that the set of  $\theta_0$  values for which  $H'_1[\theta_0]$  cannot be excluded, that is, the set of particle production rates that our data cannot exclude, is of the form  $[0, \theta_{\text{up}}]$  for some upper limit  $\theta_{\text{up}}$ . The value of  $\theta_{\text{up}}$  depends on the size of test (4.82). A common choice in high energy physics is 5%, so that the upper limit  $\theta_{\text{up}}$  will have 95% CL.

From a frequentist point of view there are two problems with the scenario of discovery versus non-discovery just outlined: one concerns coverage, and the other measurement sensitivity. We discuss these issues in the next two subsections.

#### 4.6.1

##### Coverage

When we claim discovery, we typically quote a 68% CL, two-sided interval on the new particle production rate  $\theta$ . When we fail to claim discovery, we quote a 95% CL upper limit. What is the reference ensemble (see Section 4.3.1.2) of these confidence level statements? It might seem sensible to refer the 68% CL intervals to the ensemble  $\mathcal{E}_1$  of all searches that claim discovery, and the 95% CL limits to the ensemble  $\mathcal{E}_2$  of all searches that don't. Unfortunately this doesn't work, because in  $\mathcal{E}_2$  the upper limits undercover at large values of  $\theta$  and in  $\mathcal{E}_1$  the two-sided intervals undercover at low values of  $\theta$ . One might perhaps think that this problem would disappear if we had just one common reference ensemble instead of two; and that this could be achieved if we decided to quote the same confidence level for both upper limits and two-sided intervals, say 90%. However, as shown in [8] this doesn't work either, because there is then a set of intermediate values of  $\theta$  where the coverage is only 85%.

The real source of the problem lies in the fact that the decision regarding the type of interval to quote is based on the observation itself; [8] calls this policy ‘flip-flopping based on the data’. There would be no undercover if somehow the decision could be made *before* looking at the data. Since this is not possible in the context of a search for new physics, [8] advocates the use of the likelihood ratio ordering rule described in Section 4.3.1.3. With this rule, intervals of a given confi-

dence level are two-sided (see Figure 4.8e) when the observation is above a certain threshold, and one-sided when it is below. However, this does not yet completely solve the problem, because as mentioned previously, the search procedure used in practice involves three confidence levels:  $5\sigma$  to decide on a discovery claim, 95% for upper limits, and 68% for two-sided intervals. Thus, for example, there would still be undercoverage if one were to report 68% CL likelihood ratio intervals *only* when claiming discovery, and so on. The solution here is to *always* report both the 68% and 95% CL intervals.

#### 4.6.2

##### Sensitivity

The sensitivity issue arises when there is no convincing evidence favouring the existence of a new particle, and we cannot reject the background-only hypothesis  $H_0 : \theta = 0$ . We then proceed to set an upper limit  $\theta_{\text{up}}$  on the new particle production rate  $\theta$ , typically at the 95% confidence level. The *desired* interpretation of  $\theta_{\text{up}}$  is that  $\theta$  values above it are both within the sensitivity range of the experimental apparatus and excluded by the observations;  $\theta$  values below  $\theta_{\text{up}}$  are either outside the sensitivity range or not excluded by the observations. Unfortunately, when  $\theta_{\text{up}}$  is determined by a frequentist method, there is a finite probability that it will exclude  $\theta$  values to which the experiment is not sensitive. As a simple illustration, consider the case where the observation is a Poisson distributed number of events  $x$  with mean  $\theta + \nu$ , where  $\nu$  is a known background contamination. We first encountered this example in Section 4.3.3.4, where the frequentist upper limit is given by (4.34). Remarkably, that upper limit decreases as the background contamination increases, and could even be negative. However, a negative upper limit means that all physical values of  $\theta$  are excluded by the experiment, which is clearly implausible. In the case where no events are observed, the formula gives  $\theta_{\text{up}} = -\ln \alpha - \nu$ , which is negative whenever  $\nu \geq -\ln \alpha$ . In the absence of signal, the probability of no events is  $e^{-\nu}$  and therefore the probability of a negative upper limit could be as high as  $e^{\ln \alpha} = \alpha$ . For a 95% CL limit this is 5%, which is considered quite substantial by many physicists.

Several attempts have been made to handle this problem [30], none entirely satisfactory. All have to deal with the ambiguity of deciding which  $\theta$  values are outside the sensitivity reach of the experiment, and whether this set of values depends on the confidence level of the upper limit or even on the strength of evidence provided by the data [31]. To compare approaches it is convenient to introduce some notation: let  $p_0$  be the  $p$ -value used to test  $H_0$  in test (4.81) and  $p_1(\theta_0)$  the  $p$ -value used to test  $H'_1[\theta_0]$  in test (4.82). The standard frequentist  $(1 - \alpha)$  CL upper limit construction rejects  $\theta$  values for which  $p_1(\theta) < \alpha$ . A simple modification of this construction that addresses the sensitivity problem is to reject  $\theta$  values for which *both*  $p_1(\theta) < \alpha$  and  $\theta \in \mathcal{S}$ , where  $\mathcal{S}$  is the subset of parameter space that contains all the  $\theta$  values to which the experiment is deemed to be sensitive. There

is no unique way of defining the sensitivity set  $\mathcal{S}$ . One approach [32] defines it as containing all  $\theta$  values that have probability at least  $\beta$  of being detected at the  $\gamma$  significance level if  $H_1$  is true:

$$\mathcal{S} = \{\theta : P[p_0 \leq \gamma \mid H_1[\theta]] \geq \beta\}. \quad (4.83)$$

Thus, in addition to the confidence level  $1 - \alpha$ , this method requires the choice of two probabilities,  $\beta$  and  $\gamma$ .

An alternative definition of  $\mathcal{S}$  is as the set of  $\theta$  values for which the inequality  $p_1(\theta) < \alpha$  is *expected* to occur with probability at least  $\beta$  if  $H_0$  is true:

$$\mathcal{S} = \{\theta : P[p_1(\theta) \leq \alpha \mid H_0] \geq \beta\}. \quad (4.84)$$

The advantage here is that one needs to choose only one additional probability, namely  $\beta$ . The  $\theta \in \mathcal{S}$  requirement is sometimes called a power constraint, due to the fact that the probabilities calculated in definitions (4.83) and (4.84) are power functions, that is, they are probabilities for rejecting one hypothesis when the other is true.

Although power constraint methods provide some protection against excluding parameter values to which the experiment is not sensitive, they fail to address another problem of the frequentist limit (4.34), which is that if two experiments have different background contaminations but observe the same number of events, the experiment with the larger contamination will be able to exclude more signal [33]. An approach which is arguably more successful at dealing with all manner of sensitivity problems is the so-called  $CL_s$  prescription [33, 34]. A  $(1 - \alpha)$  CL  $CL_s$  upper limit construction rejects  $\theta$  values for which

$$CL_s \equiv \frac{p_1(\theta)}{1 - p_0} < \alpha. \quad (4.85)$$

Note that this is a *stronger* requirement than the standard frequentist rejection criterion  $p_1(\theta) < \alpha$ . As a result,  $CL_s$  upper limits overcover from a frequentist point of view. On the other hand, in simple problems such as setting an upper limit on a Gaussian or Poisson mean, the  $CL_s$  result agrees with the Bayesian one for a constant prior.

It should be kept in mind that the  $CL_s$  prescription is nothing more than the rejection criterion (4.85). Just as with standard  $p$ -values, there is complete freedom in the choice of test statistic and method for handling nuisance parameters. Experiments at LEP, the Tevatron, and the LHC have all adopted different conventions and strategies in this regard, and one should be careful when attempting comparisons. In contrast with  $p$ -values however, the  $CL_s$  prescription is only used to compute upper limits.

Finally, we emphasise that Bayesian methods do not suffer from sensitivity problems due to the fact that they fully condition on the observations.

**4.7****Final Remarks and Recommendations**

One way to view the great assortment of interval constructions discussed in this chapter is as a set of answers to slightly different questions. Frequentist and Bayesian intervals with various ordering rules can all produce different inferences from the same dataset. Whether these differences matter depends on the biases and expectations of the analyst, but also on objective factors such as the evidence available prior to the measurement, the sample size, systematic effects, and instrumental sensitivity. Thus, if the consumer of the measurement result is provided with more than one interval estimate, for example a frequentist, a Bayesian, and an asymptotic construction, then he or she will better be able to judge the robustness and significance of the final result.

A recurring problem in high energy physics is the handling of nuisance parameters. When the sample size is large enough, asymptotic approximations based on the likelihood function can be trusted. However, care is required in small samples. An approximate frequentist approach is to first eliminate the nuisance parameter(s) by profiling or Bayesian integration, and then apply a test-inversion method on the parameter of interest. Although past experience with this approach has shown it to be reliable, one is always well advised to perform a few spot checks of the coverage. When using a Bayesian interval construction on small samples, one should of course evaluate the sensitivity of the final result to reasonable changes in the prior.

Another issue arises when the parameter space has physical boundaries, especially when the experiment has only weak sensitivity in the vicinity of such a boundary. The main concern is to avoid reporting intervals that exclude parameter values to which the apparatus is not sensitive. Bayesian methods appear to behave properly in this situation, but no single frequentist method is entirely satisfactory. This is another argument for reporting more than one type of interval.

**4.8****Exercises****Exercise 4.1 Measurements with Gaussian uncertainties**

The position of a charged particle is measured in a silicon strip detector to be  $y = 150 \mu\text{m}$ . Determine the 68%, 95% and 99.7% CL equal-tailed intervals for the true position  $y$  for the cases that the measurement has

- a) a Gaussian uncertainty with width  $\sigma = 10 \mu\text{m}$ ;
- b) an uncertainty described by the sum of two Gaussians: a first one with  $\sigma = 10 \mu\text{m}$  and a second one with  $\sigma = 200 \mu\text{m}$ . The first Gaussian contributes with 90% and the second one with 10%. Such a case could happen if the detector has

an inefficiency for signal hits and there is noise creating fake hits at random positions.

#### **Exercise 4.2 Rate of rejects (binomial statistics)**

A manufacturer has invented a new electronics chip and delivers a test series of ten chips to a customer. Three of the ten chips fail the functional test.

- Determine equal-tailed 95% CL intervals for the single-chip failure probability  $\epsilon$ , using the frequentist (Clopper–Pearson) and Bayesian (with Jeffreys' prior) constructions.
- Estimate a 95% CL upper limit on  $\epsilon$ .

#### **Exercise 4.3 Poisson statistics**

One of the many new physics studies at the LHC is the search for the production of a hypothetical heavy  $Z'$  particle in the decay channel  $Z' \rightarrow ll$ , where  $ll$  denotes a lepton pair. The particle would produce a signal peak in the distribution of the invariant mass of the two decay leptons that are reconstructed in the detector. Let us assume a hypothetical  $Z'$  mass of 2 TeV. After collecting a certain amount of luminosity, one expects in a signal mass window from 1.8–2 TeV a certain number of background events from Standard Model processes. The number is represented by the Poisson parameter  $\mu_b$  which is assumed to be known. The goal is to determine confidence regions for the corresponding unknown signal parameter  $\nu_s$ . Let us assume the following four measurement cases:

- One observes  $N = 5$  events and  $\mu_b = 1.3$ ;
- one observes  $N = 5$  events and  $\mu_b = 6.5$ ;
- one observes  $N = 90$  events and  $\mu_b = 100$ ;
- one observes  $N = 132$  events and  $\mu_b = 100$ .

Calculate and compare for each of the four cases

- the 95% CL upper limits on  $\nu_s$ , using the frequentist (Garwood) and Bayesian (with Jeffreys' prior) constructions, and
- 68% CL equal-tailed intervals for  $\nu_s$ , using the frequentist (Garwood) and Bayesian (with Jeffreys' prior) constructions and the frequentist 68% CL Feldman–Cousins intervals.

#### **Exercise 4.4 Bootstrapping a distribution**

A high energy physics experiment uses two production lines for building detector sensors (for example for a silicon strip detector). A test series of ten sensors is delivered from each production line. The detection inefficiencies<sup>8)</sup> of the ten sensors are found to be as follows:

8) For each sensor the inefficiency is already an average over the sensor strips.

- first production line: 9, 12, 11, 8, 7, 5, 8, 9, 10, 7%;
- second production line: 8, 13, 4, 7, 7, 8, 6, 9, 10, 5%.

Determine the sample mean and estimate its variance. Resample the data, that is draw ten values (with replacement) to obtain a first bootstrap sample and determine its sample mean. Repeat the resampling 100 times and obtain an empirical bootstrap distribution of the sample mean. Determine 68% and 95% CL bootstrap intervals, using the *simple percentile intervals*. Is there any hint towards a different inefficiency performance for the sensors from the two production lines?

#### Exercise 4.5 Eliminating nuisance parameters by conditioning

In the frequentist paradigm, handling nuisance parameters can be a thorny problem. A method that sometimes works is based on the idea of *conditioning*. To illustrate this approach, suppose we measure an event count  $N$  that is Poisson-distributed with mean  $\mu\nu$ , where  $\mu$  is the parameter of interest and  $\nu$  a nuisance parameter. Assume that  $\nu$  is constrained by the auxiliary measurement of a Poisson variate  $K$  with mean  $\tau\nu$ , where  $\tau$  is a known constant:

$$N \sim \text{Poisson}(\mu\nu), \quad (4.86)$$

$$K \sim \text{Poisson}(\tau\nu). \quad (4.87)$$

In high energy physics one could think of  $\mu$  as the production cross section for some process of interest and  $\nu$  as a product of efficiencies, acceptances, and integrated luminosity. One can argue that the *sum*  $M \equiv N + K$  provides no information about the *ratio*  $\mu/\tau$  of the above two Poisson means, or about  $\mu$  itself. It is therefore interesting to seek inferences that condition on  $M$ . First, show that the conditional distribution of  $N$  given  $M$  is given by:

$$P[N = n \mid M = m] = \binom{m}{n} \left( \frac{\mu}{\tau + \mu} \right)^n \left( 1 - \frac{\mu}{\tau + \mu} \right)^{m-n}. \quad (4.88)$$

This is a binomial distribution that does not involve the nuisance parameter  $\nu$ ; it can therefore be used for inference about  $\mu$ . Using the results from Section 4.3.3.3 on binomial efficiencies for example, one can compute a confidence interval for the binomial parameter  $\mu/(\tau + \mu)$ . Assuming you have such an interval, transform it into an interval for  $\mu$  and examine what happens when  $n = 0$ . What about when  $k = 0$ ? Or when  $n = k = 0$ ?

Next, suppose that the mean of  $N$  is the sum of  $\mu$  and  $\nu$  instead of their product, so we have:

$$N \sim \text{Poisson}(\mu + \nu), \quad (4.89)$$

$$K \sim \text{Poisson}(\tau\nu). \quad (4.90)$$

Can we still apply the conditioning method to eliminate the nuisance parameter  $\nu$  here?

**Exercise 4.6 Bayesian intervals for an exponential lifetime**

Consider the exponential decay example of Section 4.3.3.2, where the probability density of the data  $t$  is  $f(t; \tau) = e^{-t/\tau}/\tau$ . Derive Jeffreys' prior for this problem and compute the corresponding posterior. Construct equal-tailed intervals from this posterior and compare them to the corresponding frequentist intervals in Table 4.1. What can you conclude about the relationship between Bayesian credibility and frequentist coverage for this problem?

Show that the intrinsic discrepancy loss for this problem is given by

$$\delta\{\tau_0, \tau\} = \min\left\{\frac{\tau_0}{\tau}, \frac{\tau}{\tau_0}\right\} - 1 + \left|\ln\left(\frac{\tau_0}{\tau}\right)\right|, \quad (4.91)$$

and the posterior expectation of this loss by

$$\begin{aligned} d\{\tau_0; t\} = & -\left(1 + \frac{\tau_0}{t}\right)e^{-t/\tau_0} + \frac{\tau_0}{t} - 1 + \gamma + \ln\left(\frac{t}{\tau_0}\right) \\ & + \left(2 + \frac{t}{\tau_0}\right)E_1\left(\frac{t}{\tau_0}\right), \end{aligned} \quad (4.92)$$

where  $\gamma = 0.577\,215\,664\,901\,532\,860\,60\dots$  is the Euler–Mascheroni constant and  $E_1(x) = \int_x^\infty (e^{-t}/t)dt$  is the exponential integral. Plot  $d\{\tau \mid t\}$  as a function of  $\tau$ , for  $t = 1$ , and compare the minimum-loss estimate of  $\tau$  with its maximum-likelihood estimate. Intrinsic loss intervals can only be computed numerically for this problem. How do they compare with likelihood intervals? With frequentist intervals?

**Exercise 4.7 Graphical representation of search procedures**

Suppose we make a measurement  $X$  that has a Gaussian distribution with unknown mean  $\theta$  and unit width. Suppose also that the value  $\theta = 0$  has the special physical significance of ‘no signal’, whereas  $\theta > 0$  represents ‘signal’. In this simple model, the measurement sensitivity can be quantified by the difference  $\Delta\theta$  between the  $\theta$  values under a given signal hypothesis and under the no-signal hypothesis. Following the discussion in Section 4.6.2 about sensitivity, make a plot with  $p_1$  along the  $y$  axis and  $p_0$  along the  $x$  axis, and draw contours of constant  $\Delta\theta$  (i.e. for a fixed value of  $\Delta\theta$ , how does  $p_1$  vary with  $p_0$  as the data  $X$  run through its range?) Note that the line of no sensitivity,  $\Delta\theta = 0$ , coincides with the second diagonal. Draw the line  $p_0 = \epsilon$ , corresponding to the threshold for rejecting the no-signal hypothesis. A line at  $p_1 = \alpha$  corresponds to the standard frequentist exclusion limit. Draw the sensitivity sets  $S$  defined in (4.83) and (4.84) and draw the  $CL_s$  threshold of (4.85). Note that under the no-signal hypothesis  $p_0$  values are uniformly distributed between 0 and 1. Therefore, the standard frequentist probability of excluding a signal hypothesis is given by 1 minus the abscissa of the intersection of the corresponding  $\Delta\theta$  contour with the line  $p_1 = \alpha$ . Show how this probability of exclusion is non-zero even when there is no sensitivity. Show how the  $CL_s$  criterion and the other two methods avoid this problem.

## References

- 1 Kass, R.E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.*, **91**, 1343.
- 2 Pratt, J.W. (1961) Length of confidence intervals. *J. Am. Stat. Assoc.*, **56**, 549.
- 3 Hodges, Jr., J.L. and Lehmann, E.L. (1963) Estimates of location based on rank tests. *Ann. Math. Stat.*, **34**, 598.
- 4 Neyman, J. (1937) Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. London A*, **236**, 333.
- 5 Cox, D.R. (1958) Some problems connected with statistical inference. *Ann. Math. Stat.*, **29**, 357.
- 6 Bondar, J.V. (1988) Discussion of “Conditionally acceptable frequentist solutions” by G. Casella, in *Statistical Decision Theory and Related Topics IV*, Vol. 1, (eds S.S. Gupta and J.O. Berger), Springer, p. 91.
- 7 Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall.
- 8 Feldman, G.J. and Cousins, R.D. (1998) Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, **57**, 3873.
- 9 Clopper, C.J. and Pearson, E.S. (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404.
- 10 Press, W.H. et al. (2007) *Numerical Recipes, the Art of Scientific Computing*, 3rd edn, Cambridge University Press.
- 11 Cousins, R.D., Hymes, K.E., and Tucker, J. (2010) Frequentist evaluation of intervals estimated for a binomial parameter and for the ratio of Poisson means. *Nucl. Instrum. Methods A*, **612**, 388.
- 12 Garwood, F. (1936) Fiducial limits for the Poisson distribution. *Biometrika*, **28**, 437.
- 13 Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, **9**, 60.
- 14 James, F. and Roos, M. (1975) Minuit – a system for function minimization and analysis of the parameter errors and correlations. *Comput. Phys. Commun.*, **10**, 343.
- 15 Cowan, G. et al. (2011) Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, **71**, 1554.
- 16 Davison, A.C., Hinkley, D.V., and Young, G.A. (2003) Recent developments in bootstrap methodology. *Stat. Sci.*, **18**, 141.
- 17 Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, **19**, 1141.
- 18 DiCiccio, T.J. and Romano, J.P. (1995) On bootstrap procedures for second-order accurate confidence limits in parametric models. *Stat. Sinica*, **5**, 141.
- 19 Punzi, G. (2006) Ordering algorithms and confidence intervals in the presence of nuisance parameters, in *Statistical Problems in Particle Physics, Astrophysics and Cosmology. Proceedings of PHYSTAT05*, (eds L. Lyons and M. Karagöz Ünel), Imperial College Press, p. 88.
- 20 Cousins, R.D. and Highland, V.L. (1992) Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Methods A*, **320**, 331.
- 21 Tegenfeldt, F. and Conrad, J. (2005) On Bayesian treatment of systematic uncertainties in confidence interval calculation. *Nucl. Instrum. Methods A*, **539**, 407.
- 22 Chuang, C.S. and Lai, T.L. (2000) Hybrid resampling methods for confidence intervals. *Stat. Sinica*, **10**, 1.
- 23 Sen, B., Walker, M., and Woodroofe, M. (2009) On the unified method with nuisance parameters. *Stat. Sinica*, **19**, 301.
- 24 Wasserman, L.A. (1989) A robust Bayesian interpretation of likelihood regions. *Ann. Stat.*, **17**, 1387.
- 25 Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*, John Wiley & Sons.
- 26 Bernardo, J.M. (2005) Intrinsic credible regions: an objective Bayesian approach to interval estimation. *Test*, **14**, 317.

- 27** Demortier, L., Jain, S., and Prosper, H.B. (2010) Reference priors for high energy physics. *Phys. Rev. D*, **82**, 034002.
- 28** Cai, T.T. (2005) One-sided confidence intervals in discrete distributions. *J. Stat. Plan. Inference*, **131**, 63.
- 29** Mandelkern, M. and Schultz, J. (2000) The statistical analysis of Gaussian and Poisson signals near physical boundaries. *J. Math. Phys.*, **41**, 5701.
- 30** Highland, V. (1987) Estimation of upper limits from experimental data. Temple University preprint C00-3539-38.
- 31** Cousins, R.D. (2011) Negatively biased relevant subsets induced by the most-powerful one-sided upper confidence limits for a bounded physical parameter. arXiv:1109.2023 [physics.data-an].
- 32** Kashyap, V.L. *et al.* (2010) On computing upper limits to source intensities. *Astrophys. J.*, **719**, 900.
- 33** Read, A.I. (2002) Presentation of search results: the CLs technique. *J. Phys. G: Nucl. Part. Phys.*, **28**, 2693.
- 34** Junk, T. (1999) Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Methods A*, **434**, 435.

## 5

# Classification

*Helge Voss*

This chapter introduces various techniques of event classification used in high energy physics and elsewhere. One of the steps in a typical data analysis is to select the events of interest, hence to classify the available data into different categories. However, even before this step, various classifications have typically been performed on the basis of the raw event observables like trigger decisions, identification of tracks, clusters and particle types, and so on.

In this chapter, we will always refer to *events* which are classified as either *signal* or *background*. These, of course, should be seen as general terms for any type of instances, signatures or observations that are classified as being of interest (e.g. belonging to the signal class) or not (e.g. background). We also restrict ourselves to binary classification as this is the most common use case in high energy physics. Multi-class classification, where events are classified into one of several different possible classes, is a common problem in other applications like recognition of handwritten letters or numbers, protein classification and so on. In general, any multi-class classification can be written as a series of binary classifications.<sup>1)</sup>

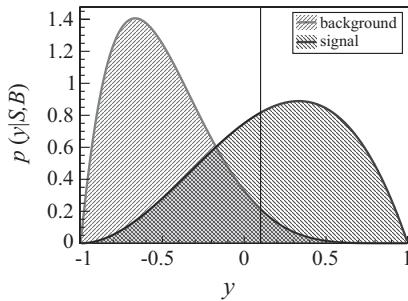
### 5.1

#### Introduction to Multivariate Classification

In comparison to the classical rectangular cut selection, that is the sequential application of requirements to the individual observables in an event, modern classification methods have become increasingly sophisticated and powerful. So-called *multivariate classification methods* or *algorithms* (MVA), which are closely related to what is called *pattern recognition* or *data mining* in a general statistics context, use the multi-dimensional observable space rather than each observable separately. Mul-

1) Possible variations are, for example, the combination of binary classifiers that are built for each class to discriminate against all the rest (*one-versus-all* approach) and the combination of classifiers where for each pair of classes an individual classifier is

designed to discriminate one class against the other (*all-versus-all* or *one-versus-one*). For more information about these and other techniques the reader is referred to the vast literature. See for example [1] for a short review.



**Figure 5.1** Distributions of an MVA variable  $y$  for signal and background events. If the measured  $y$  is above the indicated threshold value (vertical line) then the event is selected as signal, otherwise it is rejected as background.

tivariate methods use the statistical distributions of the events in the observable space to decide on a class membership of a particular event. Rather than assigning a definite class membership to an event, it is also common to attribute a probability for it to belong to a certain class.

In general, multivariate techniques combine the information of all observables of an event, often referred to as the *feature vector*  $\mathbf{x}$  of the event, into one single variable,  $y$ . This MVA variable can then be used in order to decide if the event is selected as signal or rejected as background, depending on whether the variable passes a previously set threshold or not, as indicated in Figure 5.1. In general, one can regard a multivariate selection algorithm as a mapping function of the  $D$ -dimensional feature space of the observables to a real number,  $\mathbb{R}^D \rightarrow \mathbb{R}: y = y(\mathbf{x} = \{x_1, \dots, x_D\})$ . Each constant value  $y(\mathbf{x}) = c$  then represents a hypersurface in the original observable space. Simply classifying all events with  $y(\mathbf{x}) > c$  as signal then corresponds to classifying all events on one side of this, possibly very complicated, hypersurface as signal and rejecting the others as background.<sup>2)</sup>

Some classifiers, for example the naive Bayes classifiers that are often referred to as *likelihood-selection* techniques in the context of high energy physics, are simple to code. When applying more advanced techniques the use of standard program packages that offer either individual classifiers (e.g. JETNET [2]) or a collection of many different classifiers (e.g. WEKA [3] and TMVA [4]) have become increasingly popular. The latter group of packages, even though perhaps in some cases less sophisticated for the individual classifiers, offers the possibility to easily compare different MVA techniques against each other, allowing the user to choose the simplest technique without sizable sacrifice in performance. Many articles haven been written about the comparison of different statistical approaches in data mining (see for ex-

2) Note that – rather than applying a hard selection cut using the MVA output  $y(\mathbf{x})$  – one can also use the expected signal and background MVA output distributions directly in a likelihood fit to estimate the sizes of the signal and background

contributions to a data sample. Alternatively, one can transform the output for a given event into a probability for the event to be signal or background (see (5.4)) and use this probability as an event weight in the following physics analysis.

ample [5]), which shows that it is often impossible to identify the best-performing classifier without considering the specifics of the problem at hand.

The step from classical cuts to MVA techniques is, of course, motivated by the better performance of the latter in terms of higher efficiency for the same misclassification rate. Obviously individual cuts in each observable are not able to exploit possible correlations among the different observables. In addition, a signal event that might look background-like in only a single observable will inevitably be misclassified as background in a cut-based analysis. However, it might very well be correctly classified with a multivariate classification approach that is able to compensate for this one background-like feature by exploiting all the other observables that might look very signal-like.

Constructing the perfect multivariate classifier would be easy if we had access to the full differential cross section in the actually observed event features, that is if we had the theoretical differential cross section in the observables, folded with the detector response for both signal and background events (see Section 5.2.1.1 for a detailed discussion of the *Neyman–Pearson lemma*). Then the differential cross section is the probability density function (pdf)  $p(\mathbf{x}|C)$ , where  $C$  can be either signal ( $S$ ) or background ( $B$ ). According to the Neyman–Pearson lemma, a selection algorithm based on the ratio of the pdfs (i.e. the likelihood ratio)  $p(\mathbf{x}|S)/p(\mathbf{x}|B)$  is optimal for retaining the highest signal efficiency for a given background efficiency. Unfortunately, one typically does not know the exact pdfs (for example, the detector description or effects of fragmentation or final-state radiation can only be estimated using Monte Carlo simulations). One possibility is then to approximate the true pdf using Monte Carlo-simulated events. Another possibility is to construct a suitable variable which can be used as a multivariate classifier. These two approaches are the subject of most of the rest of this chapter.

## 5.2

### Classification from a Statistical Perspective

In statistical literature one typically talks about hypothesis testing. As described in detail in Chapter 3, these tests are usually done by formulating a *null hypothesis* ( $H_0$ ). In the context of event classification, where we want to select the (signal) events we are interested in, the null hypothesis would be that an event is ‘background’. The null hypothesis is then either rejected or not, depending on the value of a *test statistic*  $y(\mathbf{x})$  which in our case is the MVA variable.<sup>3)</sup>

In most cases the probability densities of the observables for signal and background events overlap. This means that there are regions in phase space where one can find both signal and background events, leading to unavoidable errors in the decisions made to classify the events. One either misclassifies background events as signal (*type I error*), or one fails to identify a signal event as such and classifies

3) Note that in other contexts – for example in the hypothesis testing as discussed in Chapter 3 – the test statistic is often denoted as  $t$ , while in the classification context  $y$  is chosen.

**Table 5.1** The different types of errors that are made when either failing to classify a signal event as such (type II error) or misclassifying a background event as signal (type I error).

	Reject $H_0$ (select as signal)	Accept $H_0$ (select as background)
$H_0$ is false (event is signal)	Right decision with probability $1 - \beta = \text{power} = \text{efficiency}$	Wrong decision; type II error with probability $\beta$
$H_0$ is true (event is background)	Wrong decision; type I error with probability $\alpha = \text{size} = \text{significance}$	Right decision with probability $1 - \alpha = \text{background rejection}$

it as background (*type II error*). These errors have been introduced in Section 3.1.4. Their probabilities of occurrence are denoted as  $\alpha$  (*the size of the test or significance*) and  $\beta$ , where  $(1 - \beta)$  is called *power of the test* or *signal efficiency*. The quantity  $(1 - \alpha)$  is referred to as *background rejection*. Table 5.1 shows the various situations that occur when testing for a hypothesis, and the corresponding error types.

For each individual classification problem, one has to find the optimal working point, that is the best balance between type I and type II errors. It is not always sufficient to simply choose the classification that gives the smallest number of misclassifications (sum of type I and type II errors). Instead, a figure of merit has to be defined that reflects the nature of the respective analysis. Analyses can loosely be categorised as

- precision measurements: These require high purity  $p$ , that is a large fraction of signal events in the selected sample. This is achieved by keeping type I errors small;
- trigger selections: These need high efficiency  $\epsilon = 1 - \beta$  which is achieved by keeping type II errors small;
- cross-section measurements: These are optimised by maximising the signal significance which is often approximated by  $\nu_s / \sqrt{(\nu_s + \nu_b)}$  (or equivalently by maximising the quantity  $\sqrt{\epsilon \cdot p}$ );
- searches for new particles: These typically have  $\nu_s \ll \nu_b$  and are optimised by maximising  $\nu_s / \sqrt{\nu_b}$ .

In the critical region, – the part of the parameter space where  $\gamma(x) > c$  – the null hypothesis is rejected and the events are accepted as signal; the type I(II) error rates  $\alpha$  ( $\beta$ ) are given by:

$$\alpha = \int_C p(x|H_0)dx = \int_C p(x|B)dx = \int_{\gamma(x)>c} p(x|B)dx . \quad (5.1)$$

Similarly we have

$$\beta = \int_{\gamma(\mathbf{x}) < c} p(\mathbf{x}|S) d\mathbf{x}. \quad (5.2)$$

The boundary of the critical region  $C$  is also called the *decision boundary*. Rather than integrating over the multi-dimensional observable space, we can also revert to using the distributions of the test statistic  $\gamma(\mathbf{x})$ . Once the distributions  $p(\gamma|S)$  and  $p(\gamma|B)$  have been determined, the integrals (5.1) and (5.2) can easily be evaluated:

$$\int_{y=c}^{\infty} p(\gamma|B) dy = \alpha \quad \text{and} \quad \int_{-\infty}^{y=c} p(\gamma|S) dy = \beta. \quad (5.3)$$

Rather than defining hard decision boundaries, the probability densities, expressed either in terms of the multi-dimensional observables  $p(\mathbf{x}|S(B))$  or the one-dimensional projection  $p(\gamma = \gamma(\mathbf{x})|S(B))$ , can also be used to calculate the probability of an observed event to be of either signal or background origin. For this one needs the prior probabilities of randomly picking either a signal or a background event given by the relative fraction of signal ( $f_s$ ) and background events ( $f_b = 1 - f_s$ ) in the sample. The probability of an event observed with features  $\mathbf{x}$ , resulting in  $\gamma(\mathbf{x}) = \gamma$ , is then given by<sup>4</sup>

$$P_s(\gamma) \equiv P(S|\gamma) = \frac{p(\gamma|S) \cdot f_s}{p(\gamma|S) \cdot f_s + p(\gamma|B) \cdot (1 - f_s)}. \quad (5.4)$$

Assuming  $\nu_{s(b)}$  to be the expected number of signal (background) events in the sample on which the classification algorithm is run, the relative abundance  $f_s$  is given by  $f_s = \nu_s / (\nu_s + \nu_b)$ .

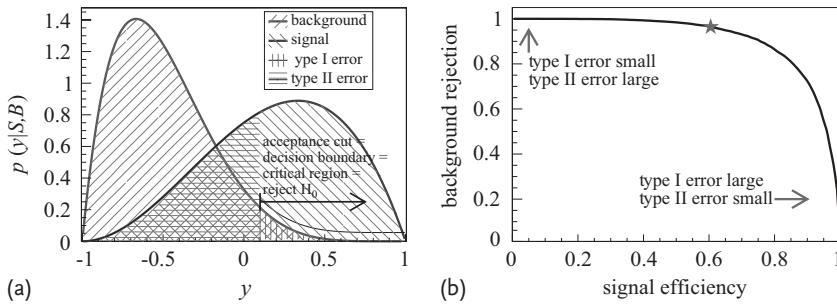
### 5.2.1

#### Receiver-Operating-Characteristic Curve and the Neyman–Pearson Lemma

Both type I and type II errors can be made arbitrarily small individually, at the expense of the other one becoming large. As indicated earlier, the best working point, that is the cut value  $c$ , has to be found for each analysis individually. This is equivalent to choosing a particular cut value  $c$  on the MVA output variable  $\gamma(\mathbf{x}) = c$ , which in turn is equivalent to picking a particular point on the so-called *Receiver-Operating-Characteristic* (ROC) curve which is often used to display the performance of a classification algorithm. It shows the relation between the signal efficiency and the background rejection. An example is shown in Figure 5.2.

Since typically the MVA variable distributions for signal and background overlap (as is also the case for the example shown in Figure 5.2a), the background rejection becomes worse for higher signal efficiency, and vice versa. This is also illustrated

4) If the true distributions  $p(\gamma|S)$  and  $p(\gamma|B)$  are not known, estimates can be used which are often referred to as  $\hat{p}(\gamma|S)$  and  $\hat{p}(\gamma|B)$ .



**Figure 5.2** (a) The distributions  $p(y|S)$ ,  $p(y|B)$  of an MVA variable  $y$  for signal and background events. The classification is based on a cut on the MVA variable  $y$  which in this example is chosen to be at  $y = 0.1$ . (b) The

ROC curve, showing the background rejection as a function of the signal efficiency achieved by varying the cut on the MVA output variable. The working point according to the cut value of 0.1 is indicated by a star.

by the ROC curve in Figure 5.2b. For a hypothetical cut value of  $y = 0.1$ , the areas which lead to type I and type II errors, respectively, are also indicated in Figure 5.2a. The hatched areas of the pdfs for signal and background in Figure 5.2a are proportional to the number of misclassified signal and background events, respectively. In case the total number of signal and background events are the same, the sum of the hatched areas is directly proportional to the total number of misclassified events. This area, and hence the misclassification error, is minimised for a cut where the two curves intersect, which for the given example would be at  $-0.175$ .

The overall performance for all possible cut values for a given classification algorithm is easily visualised using the ROC curve: the algorithm with the largest area underneath the curve has on average the best performance. Note that for a particular analysis with specific requirements on the type I or type II error, an algorithm with excellent performance in special regions of the ROC curve might be more suited, even if its overall performance is worse.

### 5.2.1.1 Neyman–Pearson Lemma

The Neyman–Pearson lemma [6] states that a classification algorithm based on the likelihood ratio

$$\gamma(x) = \frac{p(x|S)}{p(x|B)}, \quad (5.5)$$

as test statistic results, for all given background contaminations  $\alpha$ , in the critical region with the largest signal efficiency  $1 - \beta$ . To exploit the Neyman–Pearson lemma, the true underlying pdfs need to be known, which – as already stated – is hardly ever the case. There are two different approaches to dealing with this problem:

- Estimating the pdfs for signal and background events and then exploiting the Neyman–Pearson lemma to construct the MVA variable. This is done for Bayes-type classifiers.

- Determining the decision boundaries directly without the detour over the explicit pdfs, as for example in the linear discriminants or artificial neural networks. Typically these classifiers have some sort of parameterised decision boundaries, and the parameters are fitted minimising a *loss function*  $L(C, \gamma(x))$  which quantifies the penalty for misclassified events. A typical loss function would be the total number of misclassified events or, equivalently, the sum of type I and type II errors.

### 5.2.2

#### **Supervised Machine Learning**

*Machine learning* refers to the automated determination of the decision boundary according to a chosen algorithm. *Supervised machine learning* uses so-called *training events* for which the class memberships are known. The decision boundary is chosen by minimising a loss function – a process which is referred to as the *training of the classifier*. The resulting decision boundary is used in the following for the classification of events with yet unknown class membership.

Conversely, *unsupervised learning techniques* do not exploit training events and are typically used to find clusters in a distribution or, in general, to approximate the overall probability density of a given data sample. One example are self-organising maps [7].

In this chapter we are interested in classifying events into signal or background, and we typically have training events from Monte Carlo simulations or from data sidebands.<sup>5)</sup> We therefore concentrate on supervised machine learning.

### 5.2.3

#### **Bias–Variance Trade-Off**

The choice of the classifier which is best suited for a given classification task is very much dependent on the particular problem at hand. There are very simple classifiers with a small number of degrees of freedom that, for example, allow only for a linear decision boundary. Others allow for very complicated, non-linear features in the decision boundary and feature a large number of degrees of freedom.

Obviously, it is best to use a linear discriminator rather than one with more freedom if, according to the underlying true probability density, the classes are separable by a linear boundary. More complex classifiers will result in boundaries that are not exactly linear but tend to follow the statistical distribution of the training sample along the boundary. Conversely, if the training data already show a clear non-linearity beyond statistical uncertainties, a simple linear classifier would certainly underperform and systematically misclassify events in certain regions of the phase space.

5) The term *sideband* is often used in high energy physics where it refers to a relatively clean background sample which nonetheless shares most of the properties of the signal events. These are events which, in one dimension of the phase space, lie just next to where the signal is – for example just outside the expected signal particle mass peak.

Classifiers with a small number of degrees of freedom are naturally less prone to statistical fluctuations in the training sample: decision boundaries calculated from different sets of training data which were all drawn from the same underlying pdf would be very similar, that is they would have small *variance*. However, these decision boundaries will systematically deviate from the ideal ones given by the underlying probability density if the latter has more complicated features which cannot be described with the limited number of degrees of freedom of the classifier. This systematic deviation is referred to as the *bias* of the classifier.

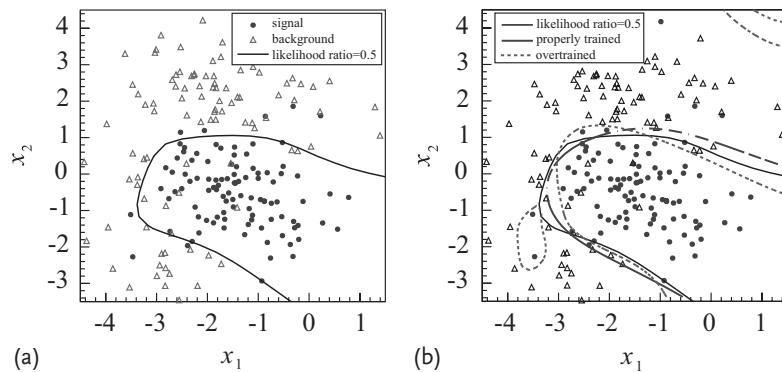
Conversely, a classifier with many degrees of freedom, allowing for very detailed decision boundaries, will have a larger dependence on the statistical fluctuations of the training sample, that is it will have a large variance (variation of the decision boundary when using different training samples), but typically a smaller bias. Defining the necessary flexibility of the model and fixing the model configuration parameters is an optimisation procedure that can also be automated, but is typically performed by the experimenter. This includes, for example, choosing the number of nodes and layers in an artificial neural network. Finding a balance between variance and bias is often referred to as the *bias-variance trade-off*.

In order to find an adequate model flexibility, one can use a so-called *validation sample*, which – like the training dataset – contains events with known class memberships. Starting with very low model flexibility and moving to ever increasing flexibility, the separation between signal and background will become better and better on the training sample because the decision boundary will become better adapted to the actual data sample. The performance of the classification tested on the independent validation sample, however, will increase only up to a certain point and will then decrease again once one starts to *overtrain*. The term ‘overtraining’ refers to the scenario where small-scale features in the decision boundary are dominated by statistical fluctuations in the training sample rather than by actual features of the underlying pdfs. This obviously leads to a worse performance of the trained classifier compared to properly trained classifiers with an appropriate degree of flexibility. One also speaks of the *generalisation properties* of a classifier. An overtrained classifier does not generalise properly as it does not capture the general features of the underlying distribution, and vice versa. It is worth mentioning that, with more training data available, more flexible models can be applied and properly fitted.

Once the best model flexibility has been determined and the model has been trained, both the validation and the training sample have been used. In order to get a truly unbiased estimate of the final performance of the classifier, a third sample – typically referred to as the *testing sample* – is used.<sup>6)</sup>

- 6) In high energy physics analyses, the validation sample is often also used as the test sample. This is acceptable if the optimisation of the model flexibility is limited, if no overtraining is expected, if the number of training events is sufficiently large such that statistical fluctuations are not significant, or if the true selection efficiency

and background contamination are in any case checked independently. Substantial biases might, however, be observed for very large background rejection or signal efficiency, where – despite a large total number of training events – only few training events dominate the calculated efficiencies.



**Figure 5.3** (a) Training events generated according to a probability density given by a multiple Gaussian mixture. Shown is also the decision boundary which results in a minimal misclassification rate (type I and type II errors). (b) The same training events as in (a). In addition, the decision boundaries obtained

by two different neural networks offering different flexibility are shown. While the fat solid line ('properly trained') approximates the true decision boundary fairly well, the dashed decision boundary ('overtrained') clearly shows some overtraining.

Figure 5.3a shows a hypothetical training sample and the decision boundary that leads to minimal misclassification of events (i.e. the smallest sum of type I and type II errors). It is calculated from the underlying pdf and is labelled 'likelihood ratio = 0.5' in the plot. Using this training sample, two different neural networks have been trained, one with adequate flexibility and the other one with a larger number of nodes and layers, offering too large flexibility. Figure 5.3b shows the corresponding decision boundaries found by these two neural networks. It is clearly visible that the decision boundary of the well-suited network is a better approximation of the optimal decision boundary than the overtrained network.

5.2.4

## Cross-Validation

If training events are scarce, one might not have enough events to separate them into a training and a validation sample. In this case, one might apply *cross-validation* for finding the appropriate model. The training sample  $T$  is split into  $K$  independent subsets  $T_k$  ( $K$  is typically chosen as 5 or 10), and then  $K$  classifiers – all using the same configuration parameters – are trained on each of the sets  $T \setminus T_k$  ( $T$  excluding subset  $T_k$ ).

For each individual training event, there is now one out of the  $K$  classifiers that has been trained without using this particular training event. This classifier is then applied to this event during the validation process. In this way, the model configuration parameters can be optimised. The optimised configuration parameters are then used to train the final classifier with the full training sample.

### 5.3

#### Multivariate Classification Techniques

Bayesian classifiers in general model the probability densities that underlie the signal and background events. Applying Bayes' theorem with the knowledge of the total number of signal and background events in the sample, one can then transform the  $p(\mathbf{x}|C = S, B)$  for a test event observed at  $\mathbf{x}$  into a probabilistic statement about the event being signal or background.

##### 5.3.1

###### Likelihood (Naive Bayes Classifier)

Assuming the absence of correlations between the observables, it is easy to estimate the multi-dimensional pdf, which for each event  $i$  is then simply the product of the one-dimensional pdfs  $p_{s(b),k}(x_k^{(i)})$ . The latter can be estimated with sufficient statistical precision from the normalised projections of the training events onto the individual variables.<sup>7)</sup> In the context of high energy physics the resulting classifier, the *naive Bayes classifier*, is often called the *likelihood classifier*. The product of these probability densities  $p_{s(b),k}$  is called the likelihood  $L_{s(b)}$ <sup>8)</sup> and is given by

$$L_{s(b)}^{(i)} = \prod_{k=1}^D p_{s(b),k}(x_k^{(i)}) . \quad (5.6)$$

Here  $D$  denotes the dimension of the phase space, and the densities fulfil the normalisation condition

$$\int_{-\infty}^{+\infty} p_{s(b),k}(x_k) dx_k = 1 , \quad \forall k . \quad (5.7)$$

As the classifier one uses the likelihood ratio  $y_L^{(i)}$  for event  $i$ , which is defined by<sup>9)</sup>

$$y_L^{(i)} = \frac{L_s^{(i)}}{L_s^{(i)} + L_b^{(i)}} . \quad (5.8)$$

If correlations between the observables were indeed zero, then according to the Neyman–Pearson lemma this classifier would be close to optimal because the pdf estimate from the training events in one dimension is typically very accurate. However, correlations are typically neither negligible nor purely linear and thus cannot

- 7) The projections are typically histogrammed and then smoothed in order to obtain a better estimate of the true, underlying pdfs.
- 8) The term ‘likelihood’ is used here because this product is not a real probability density. The multiplication of probabilities only results in another probability if the individual probabilities are independent, that is if the observables are uncorrelated.
- 9) The Neyman–Pearson lemma states that the likelihood ratio  $y(x) = p(x|S)/p(x|B)$ , or any monotone function of it, is the best possible classifier (see Section 5.2.1.1). The monotone transformation leading to  $p(x|S)/(p(x|S) + p(x|B))$  is given by  $y' = y/(1 + y)$ .

be eliminated using simple variable transformations (see also Section 5.4). Under these circumstances other classifiers that do not assume zero correlations are preferable.

### 5.3.2

#### **k**-Nearest Neighbour and Multi-dimensional Likelihood

Multi-dimensional likelihood methods approximate the likelihood ratio as closely as possible in the multi-dimensional observable space. If the approximation were perfect, this would give the best possible classifier as stated by the Neyman–Pearson lemma. The general idea is that the distribution of the training events, that is the *density* of events of a given class in a given region of the observable space, reflects the underlying probability density. The *only* problem with this approach is that the likelihood ratio needs to be known at points in the phase space, but we can only determine averages over phase-space regions. The number of signal and background events in various phase-space regions in the training sample only tells us something about the integral of the likelihood ratios over these regions. This might seem a very small shortcoming. However, for small numbers of training events in a certain region, the statistical uncertainties of the density estimates can become large. Thus, in the case of a large number of dimensions, one has to integrate over comparatively large portions of phase space. This is known as the so-called *curse of dimensionality*. As an example, in a ten-dimensional Euclidean phase space of size  $1^{10}$ , a cube that should capture only 1% of the phase space needs to have an edge length of  $\Delta x^{10} = 0.001 \rightarrow \Delta x \simeq 0.5$ , hence spanning half of the observable space in each dimension. A region like this would certainly not qualify as a ‘small volume’ around the phase-space point at its centre.

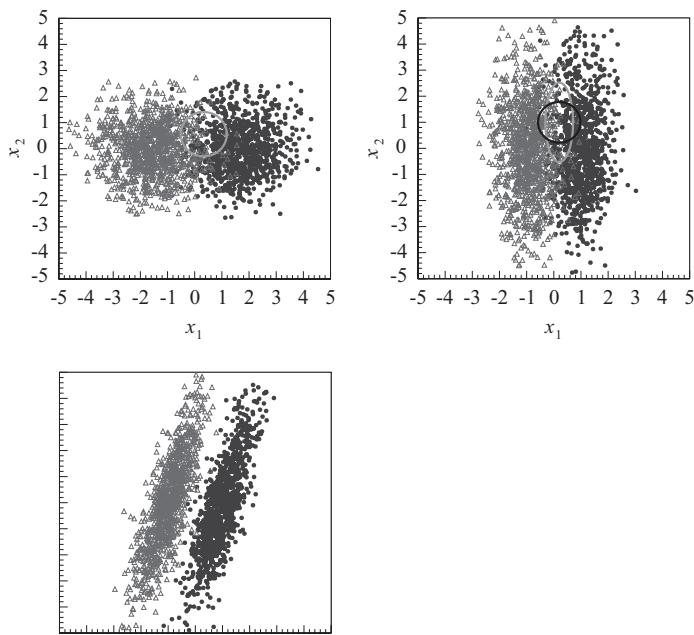
The *k*-Nearest Neighbour (kNN) algorithm provides an estimate for a multi-dimensional likelihood where the densities  $p(\mathbf{x}|S)$  and  $p(\mathbf{x}|B)$  are approximated by the number of signal and background events of the training sample,  $k_s(\mathbf{x})$  and  $k_b(\mathbf{x})$ , that lie in a small phase-space volume around the point  $\mathbf{x}$ . The likelihood ratio defined in (5.5) can then be written as

$$\frac{p(\mathbf{x}|S)}{p(\mathbf{x}|B)} \propto \frac{P(S|\mathbf{x})}{P(B|\mathbf{x})} \simeq \frac{k_s(\mathbf{x})}{k_b(\mathbf{x})}. \quad (5.9)$$

If the ratio of signal and background events in the training sample corresponds to the true value, this can be translated directly into an estimate of the probability of a test event at the phase-space point  $\mathbf{x}$  being of signal type:

$$P_s(\mathbf{x}) = \frac{k_s(\mathbf{x})}{k_s(\mathbf{x}) + k_b(\mathbf{x})} = \frac{k_s(\mathbf{x})}{k(\mathbf{x})}. \quad (5.10)$$

In order to define a ‘small volume’, the kNN algorithm specifies as a resolution parameter the number  $k = k_s + k_b$  of training events that is supposed to be inside the volume around the phase-space point  $\mathbf{x}$ . This means the volume around  $\mathbf{x}$  over which the pdf gets averaged is chosen such that it is just large enough to contain  $k$



training events. One then typically starts off at the point  $\mathbf{x}$  and increases the volume around it until  $k$  events fall inside. Large values of  $k$  do not allow one to capture small features in the pdfs, while small values of  $k$  lead to a large dependence on statistical fluctuations in the training data (see the discussion of the bias-variance trade-off in Section 5.2.3). The advantage of this approach is that it automatically adapts the size of the region over which the pdf is averaged to the available training data at each phase-space point.

The only thing left to specify is the *metric* that defines how one increases the phase space in the different dimensions in order to include the  $k$  events. The simplest choice is the Euclidean metric which, for a  $D$ -dimensional phase space, defines the distance

$$R = \left( \sum_{k=1}^D |\mathbf{x}_k - \mathbf{y}_k|^2 \right)^{\frac{1}{2}}, \quad (5.11)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  denote two points in this phase space. In fact, the choice of the metric is crucial and has a large influence on the performance of the kNN algorithm. Obviously with the simple Euclidean metric the kNN classification is not scale-invariant and, moreover, the algorithm might be dominated by a feature (dimension, observable) with very little separation power simply because this observable has small numerical values and hence the volume is effectively expanded mostly into this dimension. It is therefore customary to at least scale the metric in each dimension such that all variables are within the same numerical parameter range. Going one step further (see Figure 5.4) one can apply the full inverse covariance matrix  $\mathbf{V}^{-1}$  as a scaling of the metric to account also for correlations as it is done in the *Mahalanobis distance* [8]:

$$R_{\text{rescaled}} = \left[ \sum_{k,\ell=1}^D (\mathbf{x}_k - \mathbf{y}_k)(\mathbf{V}^{-1})_{k,\ell}(\mathbf{x}_\ell - \mathbf{y}_\ell) \right]^{\frac{1}{2}}. \quad (5.12)$$

The approach of the kNN to model the pdf using (relative) numbers of training events results in piecewise constant values of the pdf. This also means that there are discontinuities whenever an individual training event falls into or out of the corresponding volume. Using so-called *kernel functions*, that is representing each event not by a point in space but with a (for example: Gaussian) density distribution results in a smooth pdf estimate, albeit considerably increasing computation time.

### 5.3.3

#### Fisher Linear Discriminant

A linear discriminant, that is a classifier that constructs a linear decision boundary, is among the simplest classifiers. Nonetheless, it might be a good choice in many cases, particularly when the amount of training data is very limited and essentially does not allow more complicated decision boundaries to be reliably constrained.

The general form of such a linear classifier  $y(\mathbf{x})$  is given by

$$y(\mathbf{x}) = w_0 + \sum_{k=1}^D x_k \cdot w_k = \mathbf{x}^\top \mathbf{w} + w_0 , \quad (5.13)$$

where  $\mathbf{w}$  is called the *weight vector*.

As usual, any constant value  $y(\mathbf{x}) = c$  defines a specific decision boundary of the classifier. For this type of classifier, the decision boundaries are linear, that is hyperplanes with a possible offset from the origin in phase space. This classifier corresponds to a plane in the multi-dimensional phase space with the normal vector  $\mathbf{w}$  and the offset from the origin given by  $w_0/|\mathbf{w}|$ . Choosing a different constant cut value  $c$  on the test statistic is equivalent to choosing another  $w_0$ , also called *bias*. Events with  $y(\mathbf{x}) > 0$  lie on one side of the plane and would, for example, be classified as signal, and all events with  $y(\mathbf{x}) < 0$  lie on the other side and would be called background.

We are now left with the task of finding the set of parameters  $w_i$  that give the best-performing set of boundaries. This is a standard optimisation problem and can, in principle, be solved by defining and minimising a loss function that codes the quality of the separation achieved by the boundary between signal and background. An example is

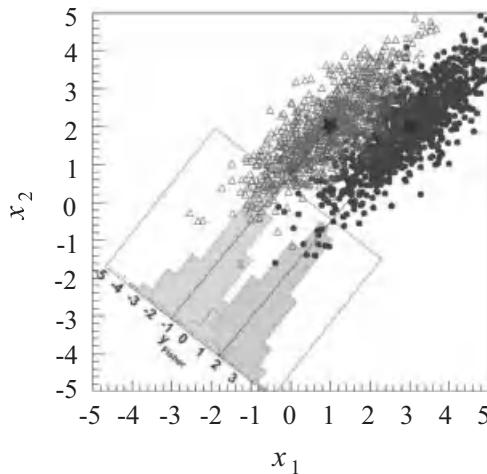
$$L = L(\mathbf{w}) = \sum_i^{\text{events}} [y^{(i)} - y(\mathbf{x}^{(i)}; \mathbf{w})]^2 , \quad (5.14)$$

where  $\mathbf{x}^{(i)}$  is the vector of observables of event  $i$  and  $y^{(i)}$  is the class membership of the event (coded for example as  $-1$  for background and  $+1$  for signal). Obviously the resulting classifier – weight vector – depends on the choice of the loss function, and there are several different functions available that are less sensitive to outliers compared to the squared loss in (5.14). Here, rather than going into detail, we concentrate on a different, very popular approach to finding an optimal linear discriminator [9].

The approach can be understood by interpreting the weight vector  $\mathbf{w}$  as the direction of an axis in the phase space onto which the events are projected. The bias or cut value ‘ $c$ ’ is then the point on this axis that separates the events classified as signal from those classified as background. The choice of the projection axis which will result in the best separation of the two classes will be the direction where the mean values of the projections for the two classes are as far apart as possible while at the same time the spread in both classes is kept minimal. An example is shown in Figure 5.5. Mathematically this condition can be expressed by maximising the ratio

$$J(\mathbf{w}) = \frac{(m_s - m_b)^2}{\sigma_s^2 + \sigma_b^2} , \quad (5.15)$$

where  $m_{s(b)} = E_{s(b)}[y(\mathbf{x})]$  is the expectation value of the projection of the signal (background) pdf onto  $\mathbf{w}$ , and  $\sigma_{s(b)}^2$  are the respective variances of the projections.



**Figure 5.5** Hypothetical distribution of signal and background events in two correlated variables  $x_1$  and  $x_2$ . The insert shows the distribution of the one-dimensional Fisher MVA variable  $y_{\text{Fi}}$  according to (5.21) and (5.22). This distribution can also be interpreted as a projection of the two variables on the one-

dimensional axis given by the Fisher discriminant. The orientation of this axis is given by the orientation of the abscissa of the inset. It maximises the difference between the mean values of the signal and background projections while at the same time minimising the spread in each of the projected distributions.

This formula can be rewritten as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{B} \mathbf{w}}{\mathbf{w}^T \mathbf{W} \mathbf{w}} . \quad (5.16)$$

**B** is the so-called *between-class matrix* which is related to the difference of the expectation (mean) values of the classes,<sup>10)</sup> and **W** is called the *within-class matrix* and is given by the expectation values of the covariances (spread) of the distributions within each class.<sup>11)</sup> It is worth noting that the matrices **B** and **W** only depend on the pdfs of the signal and background events and are independent of the parameters (weights) of the classifier. In addition, it is not necessary to obtain estimates of the full pdfs to build the classifier, but only mean values and covariances are needed. Estimates of these matrices can be conveniently calculated from the covariances of the event variables  $\mathbf{x}$  in the training

10) The elements of the *between-class matrix* are given by

$$\mathbf{B}_{k\ell} = (E_s[x_k] - E_b[x_k])(E_s[x_\ell] - E_b[x_\ell]) \quad (5.17)$$

$$= (\boldsymbol{\mu}_s - \boldsymbol{\mu}_b)(\boldsymbol{\mu}_s - \boldsymbol{\mu}_b)_{k\ell}^T , \quad (5.18)$$

where  $E_s[x_k] = (\mu_s)_k$  is the expectation value of signal events for variable  $x_k$ . An estimate of this matrix can be obtained by replacing the expectation values by averages over the sample.

11) The elements of the *within-class matrix* are given by

$$\begin{aligned} \mathbf{W}_{k\ell} &= (E_s[x_k \cdot x_\ell] - E_s[x_k]E_s[x_\ell]) \\ &\quad + (E_b[x_k \cdot x_\ell] - E_b[x_k]E_b[x_\ell]) \\ &= \mathbf{C}_{s,k\ell} + \mathbf{C}_{b,k\ell} , \end{aligned} \quad (5.19)$$

with  $\mathbf{C}_{s(b),k\ell}$  denoting the covariance matrix of the distribution of signal (background) events. Again, an estimate of the covariance matrix can of course be obtained from the covariance of the sample.

sample. The projection axis leading to the best classification is found by setting the gradient of  $J(\mathbf{w})$  to zero,  $\nabla J(\mathbf{w}) = 0$ , which, after some algebra, results in

$$\mathbf{w} \propto \mathbf{W}^{-1}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_b) . \quad (5.20)$$

Here  $\boldsymbol{\mu}_{s(b)}$  are the vectors of the expectation values in the various coordinates for signal (background) and  $\mathbf{W} = \mathbf{C}_s + \mathbf{C}_b$  is the within-class matrix, written as the sum of the covariances of the signal and background pdfs. Using instead the estimates  $\hat{\mu}_{s(b),k} = \bar{x}_{s(b),k}$  and  $\hat{\mathbf{W}} = \hat{\mathbf{C}}_s + \hat{\mathbf{C}}_b$  obtained in the training sample ultimately gives the Fisher coefficients  $F_k$  as

$$F_k = \alpha \sum_{\ell=1}^D \hat{\mathbf{W}}_{k\ell}^{-1} (\bar{x}_{s,\ell} - \bar{x}_{b,\ell}) , \quad (5.21)$$

where  $\alpha$  is an arbitrary scaling factor. The Fisher discriminant  $y_{Fi}^{(i)}$  for event  $i$  is then given by

$$y_{Fi}^{(i)} = F_0 + \sum_{k=1}^D F_k x_k^{(i)} . \quad (5.22)$$

The offset  $F_0$  can be used to centre the sample mean  $\bar{y}_{Fi}$  of all  $\nu_s + \nu_b$  events at zero.

The same recipe to compute the weights in the linear discriminant as derived by Fisher is obtained when optimising the squared loss function as given in (5.14) and the encoding of  $y_{s(b)} = +1(-1)$  and having the same number of signal and background events in the training sample. For different numbers of training events, the Fisher result is obtained for the quadratic loss function if the events are weighted accordingly or, equivalently, when the encoding is done with  $y_{s(b)} = \nu_s/\nu_s(-\nu_s/\nu_b)$ , that is  $y_s = 1$  and  $y_b = -\nu_s/\nu_b$ .

The linear decision boundary determined by the Fisher discriminant is obviously not suited for problems where the mean values of the signal and background distributions are equal and the two distributions only differ by their widths.<sup>12)</sup> On the other hand, for cases in which the covariance matrices of the two classes are equal and correlations are only linear, the Fisher discriminant yields the optimal classification boundaries.

### 5.3.4

#### Artificial Neural Networks – Feed-Forward Multi-layer Perceptrons

The concept of the linear discriminator of the previous section could be stated in the following way: ‘Construct a linear function  $f(\mathbf{x})$  of the observables  $\mathbf{x}$  that is

<sup>12)</sup> For the application of a Fisher discriminant to such problems one can transform the variables. If, for example, the signal and background distributions are both centred at zero and symmetric, the absolute values of the variables have mean values that differ for both classes and are suited again as input to the Fisher discriminant.

then fitted to the training data such that hyperplanes of this function  $f$ , given by constant values of  $f$ , separate regions in phase space that are dominated by either signal or background.' This prescription is obviously suited only for cases reasonably separated by a linear boundary. The general concept, however, can be easily transferred to virtually any non-linear case by simply dropping the linearity requirement. Equation 5.13 can be viewed as a linear combination of linear basis functions  $x_k$  (with the  $k$ th function taken to be the observable  $x_k$  itself). By replacing the linear basis functions with some general, non-linear ones,  $h_m(\mathbf{x})$ , we end up with

$$y(\mathbf{x}) = w_0 + \sum_{m=1}^M w_m \cdot h_m(\mathbf{x}), \quad (5.23)$$

where the *weights*  $w_m$  denote the coefficients of the linear combination. Since the number of basis functions is not limited to the number of dimensions as was the case for the basis functions  $x_k$ , the summation has been extended to  $M$ , a number one can choose freely. Given enough flexibility in the basis functions, it is easy to imagine that any function, and consequently any decision boundary, can be approximated. A popular choice for neural networks are error-like functions like the *sigmoid*,  $h(t) = 1/(1 + e^{-t})$ , or the *hyperbolic tangent*,  $h(t) = (e^t - e^{-t})/(e^t + e^{-t})$ . The output of these functions depends non-linearly on the input variable  $t$  for a limited range while being almost constant elsewhere. Like this, they are well suited for a piecewise approximation of the desired decision boundary. The different basis functions  $h_m(\mathbf{x}) = h((\mathbf{w}_m)^T(\mathbf{x}))$  are constructed using  $h(t)$  with a suitable linear combination of the observables as input  $t = \mathbf{w}^T \mathbf{x}$ . While in (5.23) we had an expansion in different basis functions  $h_m$ , this is now replaced by one single type of so-called *activation functions*  $h$ ,<sup>13)</sup>

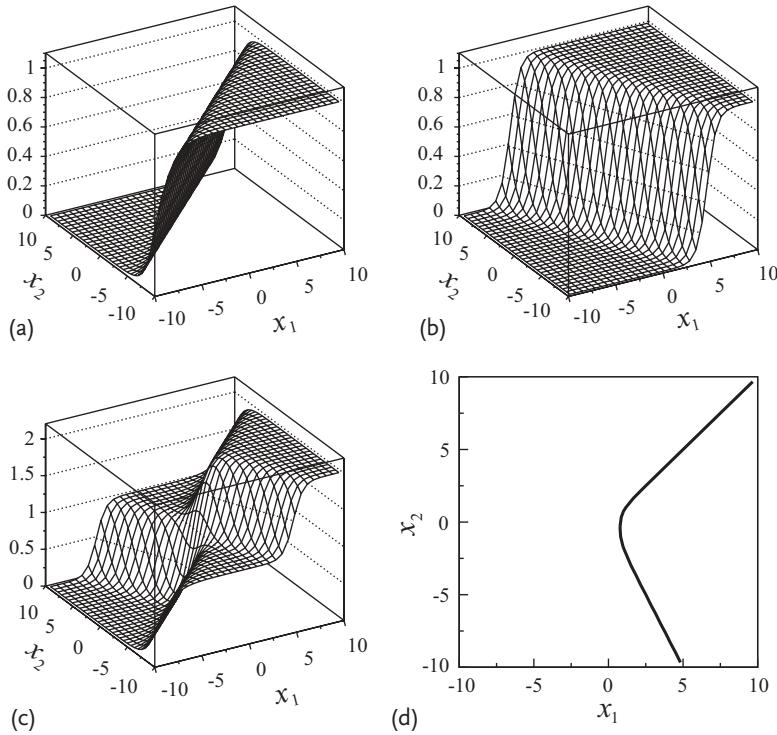
$$y(\mathbf{x}) = w_0^2 + \sum_{m=1}^M \left[ w_m^2 \cdot h \left( w_{0m}^1 + \sum_{k=1}^D w_{km}^1 x_k \right) \right]. \quad (5.24)$$

The sum over the basis functions has also been extended by a constant *bias* added with a weight  $w_{0m}^1$  which allows the linear combination of input variables to the node to be easily shifted to any *working point* of the activation function: it is more effective to introduce a single weight rather than to scale all input variables. An illustration of how two such basis functions allow the decision boundary to be defined piecewise is given in Figure 5.6.

The discriminant in (5.24) can be displayed graphically as shown in Figure 5.7 for  $D = 4$  (number of observables) and the choice of  $M = 5$  (number of basis functions).

Figure 5.7 also illustrates why one often speaks of *artificial neural networks* in this context. The basis functions can be viewed as the *neurons* (or *nodes*) that are fed by the input data via their synapses. The strength of an individual synapse is given by

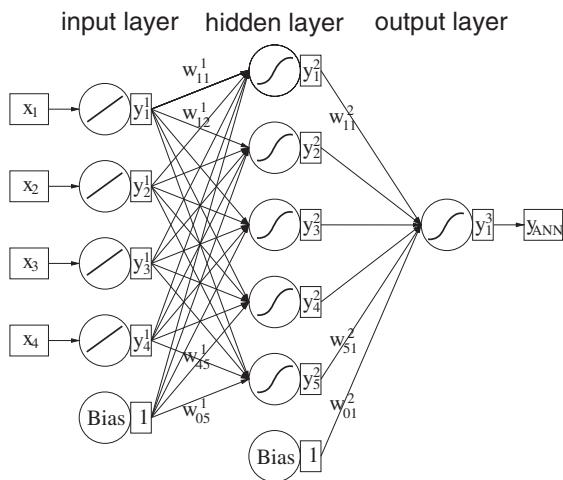
<sup>13)</sup> There are many publications showing that, given a summation over enough such basis functions, any possible function can be approximated [10–12].



**Figure 5.6** Illustration of piecewise construction of a decision boundary following an arbitrary functional form with non-linear activation functions. (a) The result of the sigmoid function of the linear combination of the input variables  $h(-1 \cdot x_1 + 1 \cdot x_2)$ ; (b) the same for a different linear combination of the inputs, namely  $(-2 \cdot x_1 - 1 \cdot x_2)$ ; (c) the sum

of these two sigmoid functions; (d) one decision boundary for an example constant value  $c = 1.5$ . Changing the input weights to the two sigmoid functions allows the boundary angles to be moved and changed. If more complicated shapes are needed, obviously more sigmoid functions need to be added.

its weight in the linear combination. The nodes are made up of a function which gives a non-constant output only in a limited range of the input. This can be seen as a neuron that only ‘reacts’ in just this input-value range and which is otherwise ‘quiet’, just like neurons in the brain only react to particular stimulations. In this example, a linear combination of the outputs of the neurons in the middle layer is fed to the output node. It is, however, straightforward to enlarge the architecture by adding more layers of neurons. For the output node, a sigmoidal activation function maps the output value onto a range between 0 and 1. In applications where signal and background processes have to be separated, neural networks are typically trained such that signal events peak close to 1 and background events close to 0. Networks with this kind of architecture, where the output of one layer of nodes is only used as input for the following layer(s), not allowing for feedback loops into previous layers, are called *feed-forward networks* or (*multi-layer*) *perceptrons* (MLPs).



**Figure 5.7** Graphical representation of the architecture of a feed-forward neural network with four input observables, one hidden layer of five nodes and one output node. The constant input (*bias*) at each layer allows a certain offset to be given to each node of the following layer.

After fixing the architecture of the network, one is again left with the task of finding the right weight values – coefficients in the various linear combinations – to actually achieve the signal and background discrimination for the particular problem at hand. In order to do so, we first define a *loss function* that quantifies the performance of the network on the training data. The different weights are highly correlated, and the loss function is typically a very complicated function of these weights with many local minima. Standard minimisation techniques are not useful for adjusting the weights in this case. The most common approach to finding the weights which give the best classification results for an MLP is *backpropagation*. Typically one starts with random weights.<sup>14)</sup> The idea of backpropagation is then to use the training events and calculate the gradient of the loss function as a function of the weights; these are then stepwise adjusted towards smaller loss functions. This can be done either after evaluation of the loss function for the full training sample (*batch learning*) or after each training event (*online learning*). Because of the structure of the network being built up as a series of identical activation functions that are always fed simply by different linear combinations of the outputs from the previous layer, the derivative of the network response with respect to the individual weights can easily be calculated using the chain rule of differentiation and evaluated at the current position in the *weight space* using the training events. This procedure is called backpropagation because one first evaluates the network output for the training event(s) with the current weights. Then one calculates the derivatives and feeds them back into the network by adjusting the weights by a certain amount,  $w \rightarrow w + \eta \cdot \nabla_w$ . Here  $\eta$  is called the *learning rate* and specifies how

14) The best starting point for weight optimisation are weights for which the sigmoid functions are almost linear, which is the case if the input to the sigmoid is close to zero.

aggressively one moves into the indicated direction of smaller loss values.

While in standard fitting procedures one often uses a square loss function, for the purpose of classification one typically reverts to more appropriate loss functions like the cross-entropy:<sup>15)</sup>

$$L(\mathbf{w}) = \sum_i^{\text{events}} [\gamma^{(i)} \ln(\gamma(\mathbf{x}^{(i)}, \mathbf{w})) + (1 - \gamma^{(i)}) \ln(1 - \gamma(\mathbf{x}^{(i)}, \mathbf{w}))] . \quad (5.25)$$

Here,  $\gamma^{(i)}$  is the coded class membership ( $\gamma^{(i)} = 1$ : signal;  $\gamma^{(i)} = 0$ : background) of training event  $i$  and  $\gamma(\mathbf{x}^{(i)}, \mathbf{w})$  is the output of the neural network for event  $i$  and weights  $\mathbf{w}$ . Variations of the simple backpropagation approach not only use the gradient of the loss function but also its second derivative, the Hessian matrix.

If a neural network offers too much flexibility, overtraining may occur as demonstrated in Figure 5.3. If the training procedure is started with weights that are in the range where the model is linear, then overtraining will not occur during the early iterations of the backpropagation. Traditionally, overtraining was avoided by early stopping, that is by allowing only a limited number of training epochs (or weight updates) thus preventing the non-linearities in the model from producing too detailed decision boundaries. More elegant methods are the inclusion of regularisation parameters in the loss function, that is  $L(\mathbf{w}) \rightarrow L(\mathbf{w}) + \sum \mathbf{w}^T \mathbf{w}$ , that penalise large weights.

### 5.3.5

#### Support Vector Machines

The *support vector machine* (SVM) is probably the most challenging classification technique presented here – conceptually, mathematically and probably also linguistically.<sup>16)</sup> This and the fact that it is also a fairly recent development might explain its limited use in high energy physics so far. However, due to its very robust optimisation during the training phase, it is a very powerful method. While neural networks suffer from the problem of finding the optimal weights when optimising a loss function that has many local minima, the training of support vector ma-

15) The distribution for observing an event as  $C = 1$  (signal) or as  $C = 0$  (background) for given measured features  $\mathbf{x}$  is best described by a Bernoulli distribution

$$P(C; \mathbf{x}) = \gamma(\mathbf{x}^{(i)}, \mathbf{w})^C [1 - \gamma(\mathbf{x}^{(i)}, \mathbf{w})]^{1-C} ,$$

with the mean value  $\gamma(\mathbf{x}^{(i)}, \mathbf{w})$ . This mean value gives the expected fraction of events of type  $C = 1$  and measured features  $\mathbf{x}$  given the model with parameters  $\mathbf{w}$ . Then the likelihood of the actually observed data for the model prediction  $\gamma(\mathbf{x}^{(i)}, \mathbf{w})$  is given by

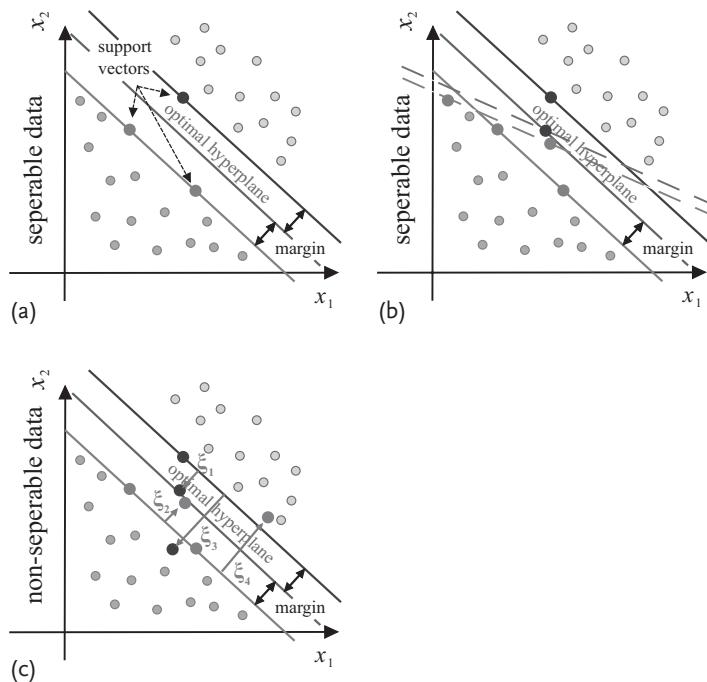
$$L(\mathbf{w}) = \prod_i^{\text{events}} \gamma(\mathbf{x}^{(i)}, \mathbf{w})^C [1 - \gamma(\mathbf{x}^{(i)}, \mathbf{w})]^{1-C} ,$$

and again, the likelihood of the observed data is maximal when the model is well fitted, that is the model is close to the true distribution. Equivalently one can minimise the negative logarithm of the likelihood,

$$\sum_i^{\text{events}} [C \ln(\gamma(\mathbf{x}^{(i)}, \mathbf{w})) + (1 - C) \cdot \ln(1 - \gamma(\mathbf{x}^{(i)}, \mathbf{w}))] ,$$

which is the loss function given in (5.25).

16) Is anyone (who has not yet heard the reason behind it) able to guess what the term *support vector* refers to? Probably not!



**Figure 5.8** Example sketches of linear separation boundaries between two event classes. (a) The largest margin and the support vectors for an example dataset. (b) A slightly different dataset with two additional events that would lead to a different margin indicated by the dashed lines. Note that the original margin might still result in a better perfor-

mance of the classifier because of its superior margin size. (c) The extension to a case where additional events make the two classes not completely separable by a linear boundary. The margin condition is relaxed, and a penalty is added for events inside the margin or on the wrong side of the boundary.

chines involves only the fitting of convex quadratic functions that exhibit only one (global) minimum. The idea behind support vector machines can be summarised in four points:

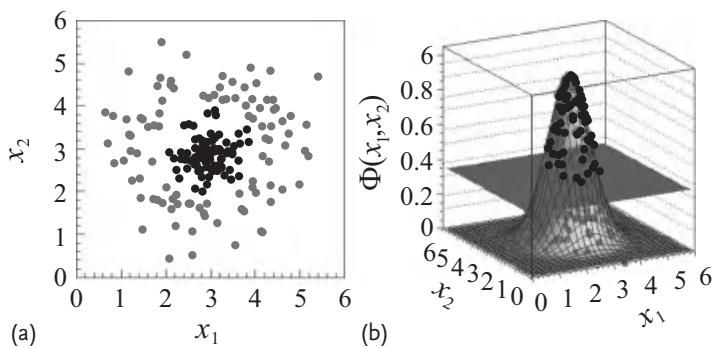
1. For linearly separable data, a good linear decision boundary would be one that lies furthest away from the nearest events on each side of the boundary. That should make it most robust against statistical fluctuations which would make the test data look slightly different than the training data. The events that happen to lie nearest to the decision boundary are called *support vectors*. As the boundary is drawn along those events, as illustrated in Figure 5.8a, it is them and only them that define the actual position of the decision boundary.<sup>17)</sup> The distance of the support vectors to the boundary is called the *margin*.

17) Note that this is different to the linear decision boundary fitted in Section 5.3.3, where the position of the actual decision boundary is given by the mean and spread of the whole ensemble of the training events and is thus influenced even by events far away from the boundary.

One could also have chosen slightly different separating boundaries which, however, would have a smaller margin and perhaps different support vectors. During the training process, all that one needs to do is to vary the decision boundary, make sure all events lie on the correct side of the boundary, determine the corresponding support vectors and margin, and maximise the latter. Depending only on distance measures, this can be formulated in a way that only depends on scalar products of the feature vectors of the events. Rather than strictly looking for the largest margin which is characterised by the fact that absolutely no event lies within its boundaries, it might be beneficial in some cases to allow for some softening of the margin constraints. An example is given in Figure 5.8b. Ignoring a few events close to the margin (which anyhow presumably are outliers) may allow another boundary to be set (e.g. the one in Figure 5.8a). This boundary would represent a much larger and hence more robust margin which will most likely give a classifier with better generalisation properties.

There is a trade-off between how many events are allowed inside the margin and the achievable margin size. In the optimisation process for the maximal margin this is achieved by adding a penalty term for events within the margin boundaries. This is called the *soft-margin approach*.

2. If the data are not totally separable by a linear boundary, one has to allow for some misclassification of events in the training process (see Figure 5.8c). This can be realised by adding the distance of the misclassified events from the margin boundary as a penalty in the optimisation process, in the same way as a penalty was added in the soft-margin approach. Again, the scale factor in this penalty is a tunable parameter of the algorithm specifying how to weigh misclassifications versus the size of the margin.
3. In general, simple linear boundaries will not result in an optimal classification. But since we have a machinery for finding linear decision boundaries, we might simply transform the data in a non-linear manner such that the transformed data are nicely separable by a linear boundary. Figure 5.9 shows an example of such a non-linear transformation of the  $D$ -dimensional observable space into a higher-dimensional feature space. Finding the appropriate non-linear transformation is not trivial and requires detailed knowledge of the data. However, basically any non-linear transformation into some higher-dimensional space will allow a better linear separation via hyperplanes compared to what had been possible in the original phase space – at least it cannot worsen the separability as one could simply ignore the additional possibilities given by the higher dimensions. In the end, the separation power merely depends on the number of dimensions of the transformed feature space rather than the actual transformation itself, as in the extreme case we could imagine all training events in their ‘own dimension’ so that they could trivially be separated by the hyperplanes along the coordinate axis.
4. As the algorithm of finding the optimal linear separation boundary was formulated in terms of scalar products of vectors, it can be immediately applied after specifying how a scalar product in some higher-dimensional feature space has



**Figure 5.9** (a) An example for a two-dimensional dataset that is not separable by a linear boundary. (b) The dataset after transformation into a three-dimensional space in which a linear separation becomes possible as indicated by the plane.

to be calculated, without ever specifying anything else about this feature space. Therefore, consider now a so-called *kernel function*  $K(\mathbf{x}, \mathbf{x}')$  which fulfils all the requirements of a scalar product and thus can be regarded as a scalar product of  $\mathbf{x}$  and  $\mathbf{x}'$  in a transformed variable space. In that case we do not even need to specify the actual transformation  $\Phi(\mathbf{x})$  as the only thing we need to know about the transformed variable space is how to calculate the scalar product:  $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$ . For some suitable kernel functions, for example the Gaussian kernel function (see (5.34)), the corresponding transformation could not even be written down, as it would actually be a transformation into a feature space with an infinite number of dimensions. But as stated before, this is not a problem as we only need the scalar product. Nonetheless, it is possible to get some insight into what actually happens: if the width of the Gaussian is sufficiently small (i.e. much smaller than the minimal distance between any two training events  $\mathbf{x}$  and  $\mathbf{x}'$ ) then all scalar products evaluate to essentially zero – meaning that each event is somehow transformed into its own dimension, allowing a perfect separation of all training events with linear boundaries along the coordinate axes. Since this will typically represent overtraining, one would rather opt for a slightly less flexible transformation – that is larger Gaussian width – and allow non-linear separability using the soft-margin approach and a reasonable penalty parameter for the misclassified training events.

A boundary as shown in Figure 5.8 can be described by  $\mathbf{w} \cdot \mathbf{x} + b = 0$ , where  $\mathbf{w}$  is a vector normal to the boundary and  $b/|\mathbf{w}|$  is the distance of the boundary from the origin. Given this definition of  $\mathbf{w}$ , the width of the margin is given by  $2/|\mathbf{w}|$ . Such a boundary will then separate signal from background if it fulfils the condition that all signal (background) training events  $i$  lie ‘above’(‘below’) the margin in Figure 5.8. This can be expressed as

$$y^{(i)} \cdot (\mathbf{x}^{(i)} \cdot \mathbf{w} + b) - 1 \geq 0, \quad \forall i, \quad (5.26)$$

with the encoding  $y^{(i)} = +1$  for signal and  $y^{(i)} = -1$  for background events. The best decision boundary, that is the one with the largest margin, is found by maximising the margin under the constraint of separating signal from background. This minimisation, together with the constraints in (5.26), can be expressed in one formula using Lagrange multipliers  $\alpha^{(i)}$ , one for each event  $i$ :

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} |\mathbf{w}|^2 - \sum_i^{\text{events}} \alpha^{(i)} \{y^{(i)} \cdot [(\mathbf{x}^{(i)} \cdot \mathbf{w}) + b] - 1\} . \quad (5.27)$$

This expression then needs to be minimised with respect to all  $w_k$  and  $b$  under the constraints that  $\alpha^{(i)} \geq 0$  and that the derivatives with respect to  $\alpha^{(i)}$  are zero. This means also that at the solution, the conditions  $\partial L / \partial b = 0$  and  $\partial L / \partial w_k = 0$  or  $\nabla_{\mathbf{w}} L = 0$  hold, leading to the following equations:

$$\mathbf{w} = \sum_i^{\text{events}} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} \quad \text{and} \quad \sum_i^{\text{events}} \alpha^{(i)} y^{(i)} = 0 . \quad (5.28)$$

These relations can be re-inserted into the Lagrangian and lead to the *dual Lagrangian*, which then needs to be maximised with respect to  $\alpha^{(i)}$ , following the constraints  $\alpha^{(i)} \geq 0$  and  $\sum_i \alpha^{(i)} y^{(i)} = 0$ :

$$L(\boldsymbol{\alpha}) = \sum_i^{\text{events}} \alpha^{(i)} - \frac{1}{2} \sum_{i,j}^{\text{events}} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \quad (5.29)$$

and

$$\sum_i^{\text{events}} \alpha^{(i)} y^{(i)} = 0 . \quad (5.30)$$

For a more in-depth and exact presentation of these procedures see *Kuhn–Tucker conditions* [13] for the extension of the Lagrangian multiplier method to constraints with inequalities, and the *Wolfe dual* [13] for the re-formulation of the original optimisation problem into this dual representation. It can be shown that the solution of a maximum of the Lagrangian  $L(\boldsymbol{\alpha})$  is also a solution to the original minimisation problem and that  $\mathbf{w}$  can be calculated from the  $\alpha^{(i)}$  that maximise  $L(\boldsymbol{\alpha})$  using (5.28).

In the soft-margin approach, a cost  $C \cdot \xi^{(i)}$  is added for events on the wrong side of the margin boundary, and the equations for margin and constraints become

$$y^{(i)} \cdot (\mathbf{x}^{(i)} \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 , \quad \xi^{(i)} \geq 0 , \quad \forall i , \quad (5.31)$$

$$W = \frac{1}{2} |\mathbf{w}|^2 + C \sum_i^{\text{events}} \xi^{(i)} . \quad (5.32)$$

The cost parameter  $C$  determines how heavily one punishes wrongly classified events, and  $\xi_i$  is the distance (in units of  $2/|\mathbf{w}|$ ) of the event to the margin boundary. A full presentation of the mathematics here is beyond the scope of this book.

Instead of replacing the scalar product  $\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$  in the Lagrangian (5.29) with the transformed feature vectors,  $\Phi(\mathbf{x}^{(i)}) \cdot \Phi(\mathbf{x}^{(j)})$ , we use the kernel function  $K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ . The dual Lagrangian in the transformed feature space then reads

$$L(\boldsymbol{\alpha}) = \sum_i^{\text{events}} \alpha^{(i)} - \frac{1}{2} \sum_{i,j}^{\text{events}} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}). \quad (5.33)$$

Typical kernel functions are

$$\begin{aligned} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) &= (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} + \theta)^d && \text{polynomial ,} \\ K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) &= \exp(-|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|^2 / 2\sigma^2) && \text{Gaussian ,} \\ K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) &= \tanh(\kappa \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} + \theta) && \text{sigmoidal ,} \end{aligned} \quad (5.34)$$

where  $\theta$ ,  $d$ ,  $\sigma$  and  $\kappa$  denote offset, degree and scale parameters of the different kernel functions. The scalar product between two events in the transformed phase space becomes essentially zero if the Euclidean distance in the original phase space becomes considerably larger than the width of the Gaussian. The scalar product between two vectors measures the size of the projection of one vector in the direction that is given by the other, just like the well-known scalar product in Euclidean space. Hence if it is zero, it means that the two vectors do not ‘share any length’ in the same direction, that is coordinate. This means that they are located in different subhyperspaces. If the width of the Gaussian kernel is smaller than the distance between *any* of the training events, all scalar products are almost zero – essentially all events live in their own dimension as they do not share any direction. Thus, the sum over all pairs in (5.33) is reduced to a sum over the events only:

$$L(\boldsymbol{\alpha}) \simeq \sum_i^{\text{events}} \alpha^{(i)} - \frac{1}{2} \sum_i^{\text{events}} (\alpha^{(i)})^2 = \sum_i^{\text{events}} \left[ \alpha^{(i)} - \frac{1}{2} (\alpha^{(i)})^2 \right]. \quad (5.35)$$

The Lagrangian  $L(\boldsymbol{\alpha})$  has its maximum at  $\alpha^{(i)} = 1 \forall i$ , and therefore all training events become support vectors. Each event being a support vector also means that each event participates in actually defining a decision boundary.

This can also be seen in the following by evaluating the resulting classifier  $y_{\text{SVM}}$  which is given by the distance of a *test event*  $\mathbf{x}$  from the boundary:

$$y_{\text{SVM}}(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b. \quad (5.36)$$

Here  $\mathbf{w}$  is the normal vector of the separation hyperplane in the higher-dimensional feature space and is given as an expansion, in terms of the support vectors, in (5.28). Substituting this leads to

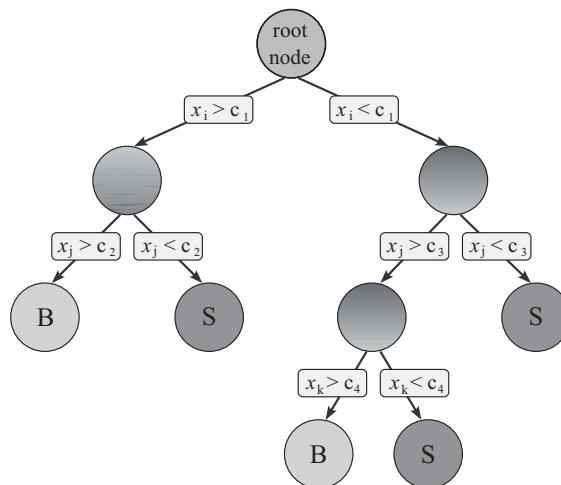
$$y_{\text{SVM}}(\mathbf{x}) = \frac{1}{M} \cdot \sum_i^{\text{events}} \alpha^{(i)} y^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}) + b. \quad (5.37)$$

Note that if the kernel width is very small,  $y_{\text{SVM}}(\mathbf{x})$  is essentially determined only by the distance to the closest support vector. The decision boundary then follows

closely the training events. Such a decision boundary would obviously represent a highly overtrained classifier and will generalise badly when applied to new events not included in the training sample. Conversely, if the kernel width is larger, the SVM output  $\gamma_{\text{SVM}}(\mathbf{x})$  in (5.37) is given by a superposition of the kernels between the event and *several* support vectors, which smoothes out the decision boundary. The ‘correct’ size of the kernel width and the cost parameter for wrongly classified events are the two most important parameters for the training of a support vector machine. They determine the size of features that can be resolved by the resulting decision boundary.

### 5.3.6 (Boosted) Decision Trees

*Decision trees* are tree-structured classifiers that consist of a series of binary splits as displayed in Figure 5.10. The tree starts from a *root node* and is built up of repeating splits and nodes down to the final or *leaf nodes*. The set of nodes and splits leading to a given leaf node is called a branch. An event is classified according to the class label of the leaf node at the end of the tree branch in which it ends up. For most decision trees the split criteria are simple cuts on individual observables (features). Each branch of a decision tree corresponds to a sequence of cuts which classifies an event as either signal or background, depending on the leaf node class label. A decision tree hence splits up the multi-dimensional observable space into many (rectangular) volumes that are attributed to either signal or background.



**Figure 5.10** A decision tree is typically a two-dimensional structure with a single root node, followed by a set of yes/no decisions (binary splits) that finally result in a set of leaf nodes. For classification, a test event is passed from

the root node down the tree and will end up in a certain leaf node depending on how it responded to the various split criteria. The event is then classified according to the class label of this leaf node.

An individual decision tree is trained by sequentially splitting the training data sample. At each step, starting from the root node which sees the whole data sample, one determines the one variable and the corresponding cut value that provide the best separation of the data. Separation is typically measured in terms of the *Gini index*  $p \cdot (1 - p)$ , where  $p$  denotes the purity, or the *cross-entropy* already defined in (5.25). These indices have maximum values when the sample is totally mixed, and they decrease monotonically for samples that become purer in either signal or background. The best split variables and values are determined by comparing the separation index before and after the split. The latter is defined by the sum of the indices of the two daughter nodes, weighted by the respective fraction of events in the nodes. The best split is then performed in order to split the training sample into two daughter nodes, for which the whole procedure is re-iterated.

Despite the aforementioned similarity with simple cuts and the easy interpretability of a decision tree, they have hardly been used in high energy physics. The way they are typically constructed (*grown* or *trained*) makes them fairly sensitive to statistical fluctuations in the training data sample, and as a multivariate classification algorithm they are typically less powerful than others. However, this changed with the advent of *boosted decision trees* (BDTs) that combine many different decision trees which together form a *forest*. All trees are trained with data samples that are derived from the training events by reweighting the events according to the *boost recipe* (see Section 5.3.7). The classification of a test event is then obtained by a (weighted) average of the responses of each of the individual decision trees in the forest.

While a ‘standard’ decision tree is best grown during the training to full size – that is until the training sample is (almost) fully divided into clean signal and background nodes – and then cut back in a *pruning procedure* to avoid overtraining, this does not hold for boosted decision trees. As the boosting procedure is designed to work with so-called *weak classifiers*, it is advantageous here to stop the individual tree growing at quite an early stage already, resulting in trees which only have a few split levels and therefore do not require pruning. The optimal tree depth depends on the degree of correlations between the variables and will have to be optimised for each use case. Boosted decision trees are very robust and powerful classifiers and have hence been referred to as the best-suited ‘off-the-shelf’ classifiers [14].

### 5.3.7

#### **Boosting and Bagging**

##### 5.3.7.1 **Boosting**

Boosting is a way of enhancing the performance of typically weak classifiers and was introduced to classification techniques in the early 1990s [15, 16]. In many cases this simple strategy results in dramatic performance increases. The idea of boosting was developed first in a more or less heuristic fashion by the invention of *adaptive boosting* (ADABOOST), see Section 5.3.7.2. The idea is to first train a classifier using the training data and to use, for the next training iteration, a modi-

fied training sample in which the previously misclassified events are given a larger weight. This procedure is then iterated, and finally the result of all the different classifiers obtained is averaged. The final classifier is then a linear combination of the so-called *base classifiers*,

$$\gamma_{\text{Boost}}(\mathbf{x}; \alpha_0, \dots, \alpha_M, \mathbf{a}_0, \dots, \mathbf{a}_M) = \sum_{m=0}^M \alpha_m b(\mathbf{x}; \mathbf{a}_m), \quad (5.38)$$

where  $b(\mathbf{x}; \mathbf{a}_m)$  denote the  $M$  trained base classifiers. The  $\mathbf{a}_m$  are the parameters of the classifier  $m$  determined in the training, and  $\alpha_m$  denotes the weight of classifier  $m$  in the final average. Boosting can be applied to all kinds of different classifiers, but is most commonly used in connection with decision trees. For boosted decision trees, the base classifier or *base learner* is a decision tree and the parameters  $\mathbf{a}_m$  specify the various node splits for the decision tree  $m$ .

It has been shown [17] that the final result is typically better if the individual classifiers in the boosted set are not too powerful. Therefore, in general, weak classifiers are boosted.

Different boosting algorithms, that is different prescriptions of how the event weights are updated in each training step and how the various base classifiers are weighted in the final linear combination, correspond to different loss functions used in the stepwise minimisation [17]. In this context, the AdaBoost algorithm described in Section 5.3.7.2 corresponds to an exponential loss function. It was, however, developed and used years before this interpretation in terms of the particular loss function was discovered. Since an exponential loss function is typically not the most robust choice, other boosting algorithms have been developed. The *gradient boost algorithm* [18, 19] algorithm allows an approximate implementation of any (differentiable) loss function.

### 5.3.7.2 Adaptive Boost (AdaBoost)

Probably the most popular boosting algorithm is the so-called AdaBoost (adaptive boost) algorithm [20]. Here, events that were misclassified during the training of the previous classifier are given a higher event weight in the following training of the classifier. Starting with the original event weights when training the first classifier, the subsequent ones are trained using modified event samples where the weights of previously misclassified events are multiplied by a common *boost weight*  $\exp(\alpha_m)$ , where the index  $m$  should refer to the weights applied before the training of the classifier  $m$ . The factor  $\alpha_m$  is derived from the fraction of misclassified training events,  $\text{err}_{m-1}$ , of the previous classifier:

$$\alpha_m = \ln \left( \frac{1 - \text{err}_{m-1}}{\text{err}_{m-1}} \right). \quad (5.39)$$

The event weights of the entire event sample are then renormalised such that the sum of weights remains constant.

We define the result of an individual classifier as  $b(\mathbf{x}) = +1$  or  $-1$  for signal and background, respectively. The boosted event classification  $y_{\text{Boost}}(\mathbf{x})$  is then given by

$$y_{\text{Boost}}(\mathbf{x}) = \frac{1}{M} \cdot \sum_m^M \alpha_m \cdot b_m(\mathbf{x}), \quad (5.40)$$

where the sum is over all of the  $M$  classifiers in the collection. Values of  $y_{\text{Boost}}(\mathbf{x})$  tending to  $+1(-1)$  indicate a more signal-(background)-like event.

### 5.3.7.3 Bagging

*Bagging* [21] denotes a technique where a classifier is repeatedly trained using resampled training events, often also referred to as *bootstrap samples*. Resampling is equivalent to regarding the training sample as being a representation of the probability density distribution of the parent sample: sampling from a dataset that follows a distribution is almost the same as sampling from the distribution itself. Just like it is more likely to ‘generate’ an event where the pdf is large, it is more likely to pick an event out of this region of phase space from the training data as it will contain more events in this region compared to elsewhere. Of course, during the sampling, one should not change the underlying pdf – that is one should leave the pool of data unchanged – which means that the same event is allowed to be (randomly) picked several times. This is typically referred to as *resampling with replacement* in the literature. A data sample generated in this way will have the same parent distribution, albeit containing statistical fluctuations. For a more detailed discussion of bootstrapping see Section 10.5.1.

Training several classifiers with different resampled training data and combining their responses into a collection results in an averaged classifier that, just as for boosting, is more stable with respect to statistical fluctuations in the training sample. A priori, bagging might not be called a *boosting* algorithm in the strict sense because boosting typically refers to iterative processes in which events are reweighted in a way that is somehow related to the previous classifier. Instead, bagging effectively smears over statistical fluctuations in the training data and is hence suitable for stabilising the response of a classifier and increasing performance by eliminating overtraining. If the variance between different classifiers in the ensemble is sufficiently large, the performance is further enhanced. A larger variance means that there is a larger pool of really different classifiers. In the combination, this has a positive effect, which seems plausible because repeatedly averaging the same classifier obviously has no effect, and because as long as the individual classifiers are all reasonable they will hardly degrade the result.

In order to get a sufficiently large variance of individual classifiers in the bagging process, it is often a good idea to choose bootstrap samples that have a smaller number of events than the training sample offers. The overall precision is retained by increasing the number of samples. The idea of increasing variance is driven even further by selecting for each bagged sample only a small random subsample of the available discriminant variables [22, 23], or in the context of decision trees, even at each node splitting as done in L. Breiman’s ‘Random Forests’ [24].

## 5.4

### General Remarks

In the previous sections we have presented the most common multivariate classifiers used in high energy physics. These classifiers are trained using simulated data, or more generally, any data where the class membership of the events is known. From these training data the classifiers *learn* how the parameters of the decision boundaries are chosen in order to get optimal separation between signal and background events.

Despite the technical finesse and ‘intelligence’ of these machine-learning tools, it is important that ‘human intelligence’ is also used in an analysis. The choice of discriminating observables is vital, and care should be taken not to include variables containing too little or no additional information. This is particularly important for neural networks where multiple local minima make the finding of the best network parameters more difficult with increasing number of observables. Also for decision trees, which are mostly inert against useless variables, it surely helps to think carefully about the input – if only to save computing time during the training.

Apart from the input variables, the choice of classifier and its specific flexibility is also important and depends on the particular selection task at hand. It is good practice to choose a model/classifier with appropriate flexibility. As the best choice is not always obvious, a comparison of the performance of different classification techniques on the same sample can be useful.

#### 5.4.1

##### Pre-processing

Most classifiers presented here are designed to take into account correlations between variables. However, if the input observables have known correlations it is certainly worthwhile to help the learning algorithm by choosing a suitable variable transformation prior to the classifier training. Such pre-processing of the input variables can result in a substantial performance increase of the classifier, particularly for classifiers that rely on uncorrelated variables like the naive Bayes classifier. For simple linear correlations, there are standard techniques – like *principal component analysis* (PCA)<sup>18)</sup> – that decorrelate the data. Decorrelation is also achieved using the transformation  $\mathbf{x}' = \mathbf{R}^{-1} \mathbf{x}$ , with  $\mathbf{R}$  being the square root of the covariance matrix  $\mathbf{C} = \mathbf{R} \cdot \mathbf{R}$ . Care needs to be taken, however, not to apply such procedures blindly, because linear decorrelation of variables with highly non-linear correlations<sup>19)</sup> might result in more harm than benefit for the classifiers.

<sup>18)</sup> ‘Principal component analysis’ refers to an orthogonal variable transformation where the transformed variables are ordered with respect to their variance, such that the first variable, or principal component, is the one with the largest variance.

<sup>19)</sup> Note that in strict statistical terms correlations always refer to what we here call ‘linear correlations’. What we call ‘non-linear correlations’ – in accordance with the typical (high energy) physics slang – is commonly called ‘dependence’ or ‘association’.

Apart from decorrelating the input variables, it is also often beneficial to transform them such that they lie within similar numerical ranges and that they do not exhibit excessive slopes, which inevitably will harm any numerical analysis of the data. Having data in the same numerical range for all variables is particularly important for multi-dimensional likelihood and support vector machines that explicitly use distances between points in the multi-dimensional observable space. These algorithms therefore often include corresponding internal transformations in one way or another.

If there are obvious symmetries in the data, for example a forward–backward symmetry in the distribution of some particle-production angle  $\theta$ , then the choice of observables presented to the learning algorithm should already reflect this symmetry by using  $|\cos(\theta)|$  rather than  $\cos(\theta)$ . In this way, the algorithm does not need to recognise this symmetry by itself, which increases the statistical power for the remaining features.

Often, one has to deal with events that for other reasons fall into different categories, for example differently performing detector regions, different numbers of reconstructed jets in the event, events with and without photons converted into electron–positron pairs, decays of  $\tau$  leptons into final states with one or three charged particles. This can create correlations between variables and will inevitably result in different optimal decision boundaries in these different event categories. While theoretically some classifiers should be able to learn these differences given enough training data, it is in general much better to also help the classifiers in this case by dividing the sample *manually* and training individual classifiers for each event category.

## 5.5

### Dealing with Systematic Uncertainties

A recurrent question in the context of multivariate classification techniques in high energy physics is how to address systematic uncertainties, and if multivariate techniques are more susceptible to these uncertainties than standard cut selections. First of all, there is no substantial difference between systematic uncertainties and their estimates in multivariate classification techniques and classical cut analyses. Major systematics arise from calibrations and other experimental effects which can mostly be spotted in one-dimensional projections, that is by studying the input observables. Obviously, if a multivariate selection explicitly tries to disentangle possibly complicated correlations in the observables, it is more prone to systematic uncertainties due to a possible mismodelling of these correlations. It is worth noting, however, that mismodelled correlations also affect cut-based analyses. Usually, this is only revealed in a careful study of the cut flows, while in a multivariate analysis the MVA output distributions will clearly show differences between samples that differ in the correlations. Hence it might be even easier to spot systematic differences between data and the simulation in an MVA analysis by comparing the MVAs. Nonetheless, once such a difference is observed, finding the underly-

ing shortcomings in the simulation remains challenging and will certainly involve looking at all possible input distributions in one- and two-dimensional projections, using only subsets of the observables in the MVA classifier, and so on – so very much the same steps as one would have to take in a classical cut analysis.

It is important to note that the MVA training itself does not introduce additional systematic uncertainties. In contrast to what is often heard, there is no such thing as a *wrong* MVA training – there can only be *bad* MVA training, in the sense of falling short of the achievable performance. Obviously, systematic uncertainties occur only once performance parameters like efficiencies or background rejection are estimated for the trained MVA classification using test samples that have systematic uncertainties. Here, again just like one would have to do with classical cut analyses, the influence of possible systematic variations on the result needs to be studied for the MVA classifier. An advantage of the MVA approach here is, however, that the influence of systematic variations needs to be studied again only on the MVA output distribution, as this is the only distribution that is finally used to determine efficiencies. Note that this does not involve a new MVA training for each possible variation, but simply the estimate of how the MVA output distribution changes for the already trained classifier, once it is applied to a different set of test events. It is advisable to apply the MVA classification to test samples with all possible systematic variations in order to determine the corresponding MVA output distributions. From this set of varied MVA output distributions it is then possible to determine the systematic uncertainty on efficiencies and background rejections.

A more complicated issue is the minimisation of systematic uncertainties entering an MVA selection; this is dealt with in Chapter 8. In a classical cut analysis, one would – for example – avoid cutting near or within steep slopes in a distribution of a variable that possibly suffers from a systematic uncertainty. By doing so, the influence of a possible variation in this variable is kept small, outweighing a loss in performance which might be encountered by the specific choice of the cut. In the automated training of a multivariate classifier, the experimenter has a difficult task in deciding how to balance systematic uncertainties against theoretical performance. One possibility is to artificially modify the training sample such that the separation power of an uncertain observable is reduced. This can be done either by shifting or by smearing the observables. The result will be that the training algorithm puts less emphasis on this observable, without totally ignoring it.

## 5.6

### Exercises

#### **Exercise 5.1** Linear classifier with quadratic loss function

Derive the linear classifier with a quadratic loss function and events weighted such that their effective numbers for signal and background are the same. Proof that this classifier is equivalent to the Fisher discriminant.

- a) Write down a linear classifier  $\gamma(\mathbf{x})$  in vector notation (as in (5.13)) but include the bias ( $w_0$ ) in the weight vector  $\mathbf{w}' = (w_0, \mathbf{w})$  and extend the observable vector to  $\mathbf{x}' = (x_0, x_1, \dots, x_D)$  with  $x_0 \equiv 1$ .
- b) What is a suitable encoding  $y$  of ‘signal’ and ‘background’ events, given that they should end up lying ‘above’ or ‘below’ a hyperplane defined by  $\gamma(\mathbf{x}') = \text{const}$ ?
- c) Write down the loss function  $L$  of this classifier and its expectation value  $E[L]$ .
- d) While the expectation value uses the integral over the probability densities, the derivation of an estimator for the expectation value requires the replacement of this integral. What is this replacement? Write down the estimate of the expectation value of this loss function for a sample of  $N_s + N_b = N$  events.
- e) In order to find the best parameters  $w_i$ , one needs to minimise the estimated loss from Exercise 5.1d), that is one must solve  $\nabla_{\mathbf{w}'} E \stackrel{!}{=} 0$ .

Assume for the following that there are either the same number of signal and background events ( $N_s = N_b = \frac{1}{2}N$ ) or that the background events have been weighted by  $N_s/N_b$ . Note that it is probably the easiest to split up the matrix equation again at some point into block matrices, splitting  $\mathbf{w}'$  into  $w_0$  and  $\mathbf{w}$ , as well as  $\mathbf{x}'$  into  $x_0 = 1$  and  $\mathbf{x}$ . This will allow you to prove and use the following three relations:

- 1) Show that  $\sum_k^N (\mathbf{w} \mathbf{x}^{(k)}) \mathbf{x}^{(k)} = N_{\text{events}} (\mathbf{V} + \boldsymbol{\mu} \boldsymbol{\mu}^T) \mathbf{w}$  with  $\mathbf{V}$  being the covariance matrix (note: of the vector  $\mathbf{x}$ , not the extended  $\mathbf{x}'$ ) and  $\boldsymbol{\mu}$  the mean values of the projections onto the coordinate axes.
- 2) Show that the total covariance matrix  $\mathbf{V}_{s+b}$ , that is the covariance matrix ignoring class membership, can be decomposed into  $\mathbf{V}_{s+b} = \frac{1}{2}(\mathbf{V}_s + \mathbf{V}_b) + \frac{1}{4}\mathbf{B}$  with  $\mathbf{B}_{ij} = (\boldsymbol{\mu}_s - \boldsymbol{\mu}_b)(\boldsymbol{\mu}_s - \boldsymbol{\mu}_b)^T_{ij}$ .
- 3) Show next that for any vector  $\mathbf{x}$ , the vector  $\mathbf{Bx}$  is pointing into the direction of  $\boldsymbol{\mu}_s - \boldsymbol{\mu}_b$ , that is  $\mathbf{Bx} = \text{const} \cdot (\boldsymbol{\mu}_s - \boldsymbol{\mu}_b)$ .

Now proceed with solving  $\nabla_{\mathbf{w}} E \stackrel{!}{=} 0$  for  $\mathbf{w}$ .

- f) Compare the result with the Fisher requirement (5.20), remembering that two linear classifiers  $\gamma(\mathbf{x})$  that only differ by the length of the weight vector  $\mathbf{w}$  and the bias are equivalent.

### Exercise 5.2 Linear discriminant analysis (LDA) and Gaussian probability densities

Show that for Gaussian probability densities with the same covariances for signal and background, the optimal decision boundary is linear and equals the one given by the Fisher discriminant, (5.22).

- a) Assume two populations (signal and background) which are both described by a Gaussian pdf. Write down the densities and the likelihood ratio (up to a proportionality constant) which, according to the Neyman–Pearson lemma, gives the best possible classifier  $\gamma(\mathbf{x})$ .

- b) Assume now that both densities have the same covariance matrix, write down the logarithm of the likelihood ratio of the two densities and show that this is a linear function in  $x$  (i.e. a linear classifier!).
- c) Show that the coefficients of the linear classifier are (up to an arbitrary factor and offset) the same as the ones obtained for the Fisher discriminant.

## References

- 1 Rifkin, R.M. and Klautau, A. (2004) In defense of one-vs-all classification. *J. Mach. Learn. Res.*, **5**, 101.
- 2 Peterson, C., Rognvaldsson, T., and Lonnblad, L. (1994) Jetnet 3.0: A versatile artificial neural network package. *Comput. Phys. Commun.*, **81**, 185.
- 3 Hall, M. et al. (2009) The WEKA data mining software: An update. *SIGKDD Explorations*, **11**, 10.
- 4 Höcker, A. et al. (2007) TMVA: Toolkit for multivariate data analysis. *PoS, ACAT*, **40**.
- 5 Jamain, A. and Hand, D. (2008) Mining supervised classification performance studies: A meta-analytic investigation. *J. Classif.*, **25** (1), 87.
- 6 Neyman, J. and Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London A*, **231**, 289.
- 7 Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43** (1), 59.
- 8 Mahalanobis, P. (1936) On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India*, **2** (1), 49.
- 9 Fisher, R. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179.
- 10 Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Math. Control Sign. Syst. (MCSS)*, **2**, 303.
- 11 Hornik, K., Stinchcombe, M.B., and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Netw.*, **2** (5), 359.
- 12 Funahashi, K.I. (1989) On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, **2** (3), 183.
- 13 Fletcher, R. (1987) *Practical Methods of Optimization*, 2nd edn, John Wiley & Sons.
- 14 Breiman, L. (1998) Arcing classifiers. *Ann. Stat.*, **26** (3), 801.
- 15 Schapire, R. (1990) The strength of weak learnability. *Mach. Learn.*, **5**, 197.
- 16 Freund, Y. (1995) Boosting a weak learning algorithm by majority. *Inform. Comput.*, **121**, 256.
- 17 Friedman, J., Hastie, T., and Tibshirani, R. (1998) Additive logistic regression: a statistical view of boosting. *Ann. Stat.*, **28**, 2000.
- 18 Friedman, J. (1999) Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, **29** (5), 1189.
- 19 Friedman, J.H. (1999) Stochastic gradient boosting. *Comput. Stat.*, **38**, 367.
- 20 Freund, Y. and Schapire, R. (1997) *J. Comput. Syst. Sci.*, **55**, 119.
- 21 Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 124.
- 22 Ho, T. (1995) Random decision forests, in *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Montreal, QC, 1995), vol. 1, IEEE Computer Society Press, p. 278.
- 23 Ho, T.K. (1998) The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20** (8), 832.
- 24 Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5.

## 6

### Unfolding

*Volker Blobel*

#### 6.1 Inverse Problems

##### 6.1.1 Direct and Inverse Processes

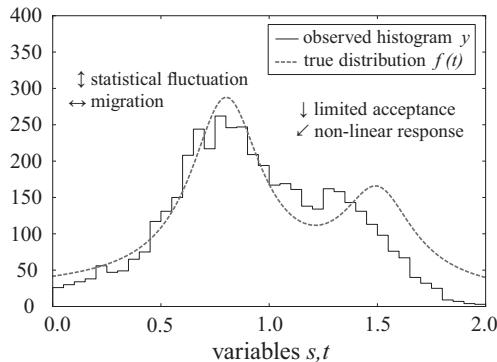
The aim of experiments in particle physics is the estimation of probability distributions of variables like energy, scattering angle, particle mass or decay length from a data sample, recorded by an often complex detector and reconstructed from raw data by involved algorithms. The data are often collected in the form of histograms, representing an estimate of the distribution, with statistical uncertainties for each bin content due to the unavoidable statistical fluctuations of random data. In addition to these fluctuations the data are subject to additional random effects of the finite resolution and limited acceptance of the detector and of the reduced (restricted) efficiency of the reconstruction. The distribution  $g(s)$  of the measured values  $s$  is related to the distribution  $f(t)$  of the true variable  $t$  by migration, distortions and transformations. Using Monte Carlo (MC) methods the *direct process* from an assumption  $f(t)^{\text{model}}$  on the true distribution  $f(t)$  to the expected measured distribution  $g(s)$  can be simulated. The *inverse process* from the actually measured distribution  $g(s)$  to the related true distribution  $f(t)$  is difficult and known to be ill-posed: small changes in the measured distribution can cause large changes in the reconstructed true distribution, if naive methods are used.

In particle physics the inverse process is usually called *unfolding*. The direct process, which can be called *folding*, and the inverse process,

$$\begin{array}{ll} \text{direct process (MC)} & \text{true/MC dist. } f(t) \implies g(s) \text{ measured dist. ,} \\ \text{inverse process (unfolding)} & \text{measured dist. } g(s) \implies f(t) \text{ true dist. ,} \end{array}$$

are described by the Fredholm integral equation of the first kind,

$$\int_{\Omega} K(s, t) f(t) dt + b(s) = g(s), \quad (6.1)$$



**Figure 6.1** Illustration of the unfolding problem. The true distribution  $f(t)$  of a variable  $t$  to be measured in a particle physics experiment is shown. A corresponding simulated measurement  $g(s)$  is shown as histogram  $y$ . See the text for further details.

with a Kernel function  $K(s, t)$  describing the physical measurement process [1–6]. The distribution  $b(s)$  in (6.1) represents a potential contribution to the measured distribution  $g(s)$  from background sources. Unfolding requires a determination of the distribution  $f(t)$  from the measured distribution, where the Kernel function  $K(s, t)$ , describing the detector response and also called the response function, is usually implicitly known from a Monte Carlo sample based on an assumption  $f(t)^{\text{model}}$ .

The unfolding problem is illustrated in Figure 6.1, which shows, as a smooth curve, the true distribution  $f(t)$  of a variable  $t$  to be measured in a particle physics experiment. A corresponding simulated measurement  $g(s)$  is shown as a histogram; it takes into account statistical and systematic effects which are typical for particle physics experiments:

- ↑ in all bins there are statistical fluctuations from the Poisson statistics which are valid for independent bins if the bin contents are counts;
- ↔ due to the finite resolution there is migration between bins (smearing);
- ↓ limited acceptance and reduced efficiency means that potential entries in a bin are missing;
- ↙ due to a non-linear detector response there is a shift, on average in a certain direction (a shift to lower values is assumed in the figure).

The effects described above are typical for the detector response in particle physics experiments, where the amount of data in unfolding is rather small, but more complex transformations between measured and true variables may occur. In contrast to this, the amount of data (e.g. millions of pixels in picture deblurring) in other applications is often large, while the corresponding response function is simple; often migrations occur only between few neighbouring points (*point spread function PSF*).

If the goal of the experiment is to test specific predictions of a theory, the predicted true distribution  $f(t)^{\text{model}}$  can be folded with the effect of the detector for the

comparison with the measured distribution  $g(s)$ . No unfolding is necessary in this case, but the sensitivity of the comparison may be difficult to estimate.

The reconstruction of the true distribution  $f(t)$  by unfolding allows a direct comparison with predictions of a theory and with other (independent) experiments, or a further statistical analysis. For a quantitative comparison a full covariance matrix is required for the reconstructed true distribution; in this respect the requirements for unfolding in particle physics differ from other fields where often no need to quantify the uncertainties in detail exists. Non-zero correlations exist between the bins of the unfolded data because of the response effects listed above.

Methods to reconstruct a true distribution by unfolding as binned data without a specific parameterisation are discussed in this chapter. The special case of unfolding assuming a certain parameterisation of the distribution, which is a simpler problem, is mentioned in Section 6.1.5.

### 6.1.2

#### Discretisation and Linear Solution

The inverse problem given by the Fredholm integral (6.1) has to be discretised in order to allow a numerical solution and can be expressed as a linear matrix equation:

$$\mathbf{A}\mathbf{x} = \mathbf{y}. \quad (6.2)$$

The relations between the functions/distributions  $f(t)$  and  $g(s)$  introduced above and the matrix  $\mathbf{A}$  and vectors  $\mathbf{x}, \mathbf{y}$  are:

$$\begin{aligned} \text{true distribution } f(t) &\Rightarrow \mathbf{x} \quad n\text{-vector of unknowns ,} \\ \text{measured distribution } g(s) &\Rightarrow \mathbf{y} \quad m\text{-vector of measured data ,} \\ \text{Kernel } K(s, t) &\Rightarrow \mathbf{A} \quad \text{rectangular } m\text{-by-}n \text{ response matrix .} \end{aligned}$$

In the following, the variables  $s, t$  and the vectors  $\mathbf{x}, \mathbf{y}$  are assumed to be one-dimensional; in practice they can be multi-dimensional, even with different numbers of dimensions for the true and the measured distributions. Several different discretisation methods are possible. The usual method is to represent the distributions as histograms, integrating the distributions over a short interval (bin), given by grids  $\{t_0, t_1, \dots, t_n\}$  for the true distribution and  $\{s_0, s_1, \dots, s_m\}$  for the measured distribution, often with equidistant bin limits:

$$\gamma_i = \int_{s_{i-1}}^{s_i} ds g(s) \quad i = 1, 2, \dots, m. \quad (6.3)$$

If the response is determined from a MC sample based on an assumption  $f(t)^{\text{model}}$ , the same method can be used for the discretisations  $K(s, t) \Rightarrow \mathbf{A}$  and  $f(t) \Rightarrow \mathbf{x}$ ; in this case element  $x_j$  is the average of  $f(t)$  in bin  $j$ . The elements  $\gamma_i$  are calculated

according to (6.2) by the product  $y_i = A_i^T \mathbf{x}$ , where the vector  $A_i$  is defined as a column vector containing the elements  $A_{i1}, A_{i2}, \dots, A_{in}$  of matrix  $\mathbf{A}$ :

$$y_i = A_i^T \mathbf{x} = A_{i1}x_1 + A_{i2}x_2 + \dots + A_{in}x_n \quad i = 1, 2, \dots, m. \quad (6.4)$$

The elements of the response matrix  $\mathbf{A}$ , describing the response function  $K(s, t)$ , are positive or zero. If the assumed distribution  $f(t)^{\text{model}}$  is a (normalised) probability density function (pdf), the element  $A_{ij}$  can be interpreted as a conditional probability  $A_{ij} = P(\text{observed in bin } i \mid \text{true value in bin } j)$ , which does not depend on the assumed distribution. The sum  $\epsilon$  of the elements  $A_{ij}$  over all bins  $i$  of the observed distribution,

$$\epsilon_j = \sum_{i=1}^m A_{ij}, \quad (6.5)$$

gives the detection efficiency of the measurement detector, that is the overall observation probability for the true bin  $j$ . It includes a geometrical acceptance and other factors. Other methods are, for example, the discretisation of  $f(t)$  by a superposition of B-splines [7], which avoids discontinuities in the unfolded distribution, or discretisation based on numerical quadrature [1, 2].

The determination of the response matrix  $\mathbf{A}$ , typically by MC simulation using a pdf  $f(t)^{\text{model}}$ , is fundamental for unfolding. The elements of the matrix  $\mathbf{A}$  from a MC simulation have statistical errors, which can delimit the effective rank (6.29) of the unfolding problem.

Assuming an accurate response matrix  $\mathbf{A}$  and the validity of the relation  $\mathbf{A}\mathbf{x}_{\text{exact}} = \mathbf{y}_{\text{exact}}$ , the measured distribution  $\mathbf{y}$  deviates from  $\mathbf{y}_{\text{exact}}$  only by data errors due to statistical fluctuations. Representing the data errors by an  $m$ -dimensional vector  $\mathbf{e}$ , the actually measured distribution  $\mathbf{y}$  is then given by

$$\mathbf{y} = \mathbf{y}_{\text{exact}} + \mathbf{e} = \mathbf{A}\mathbf{x}_{\text{exact}} + \mathbf{e}. \quad (6.6)$$

In particle physics the statistical properties of the measurement are usually well known. The elements of the vector  $\mathbf{y}$  are often counts and thus follow Poisson statistics, in which case the maximum-likelihood solution is adequate.

Assuming  $E[\mathbf{e}] = 0$ , the expectation value and variance of the measurement are generally given by

$$E[\mathbf{y}] = \mathbf{y}_{\text{exact}}, \quad (6.7)$$

$$\mathbf{V}_y \equiv V[\mathbf{y}] = V[\mathbf{e}] = E[\mathbf{e}\mathbf{e}^T]. \quad (6.8)$$

The covariance matrix  $\mathbf{V}_y$ <sup>1)</sup> is assumed to be known.

In other fields like geophysical and medical imaging or image deblurring in astronomy, the number  $n$  of parameters to be estimated in inverse problems can become very large, and no explicit knowledge of individual uncertainties is required.

1) In this chapter, covariance matrices are written with a subscript like  $\mathbf{V}_y$  (for the measured vector  $\mathbf{y}$ ); matrices  $\mathbf{V}$  without subscripts are orthogonal matrices from a decomposition (see Section 6.2).

In particle physics the number  $n$  of parameters (e.g. the number of bins of the reconstructed true distribution) is usually small, and it is essential to determine the full covariance matrix  $\mathbf{V}_x$  of the result which might be used in a further statistical analysis. If the linear Fredholm equation (6.2) is solved for the estimate  $\hat{\mathbf{x}}$  by a linear transformation of the data vector  $\mathbf{y}$  according to  $\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y}$ , the propagation of the data uncertainties to the unfolding uncertainties is straightforward:  $\mathbf{V}_x = \mathbf{A}^\dagger \mathbf{V}_y \mathbf{A}^{\dagger T}$ . The case  $m = n$  with a quadratic matrix  $\mathbf{A}$  could be solved by the inverse matrix  $\mathbf{A}^\dagger = \mathbf{A}^{-1}$ , but often the matrix  $\mathbf{A}$  has a bad condition or is even singular and  $m = n$  should be avoided [3, 8]. In the recommended case  $m > n$ , the  $n$ -by- $m$  matrix  $\mathbf{A}^\dagger$  can be constructed and used to determine the estimate  $\hat{\mathbf{x}}$ :

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y} = \mathbf{A}^\dagger \mathbf{y}_{\text{exact}} + \mathbf{A}^\dagger \mathbf{e} = \mathbf{A}^\dagger \mathbf{A} \mathbf{x}_{\text{exact}} + \mathbf{A}^\dagger \mathbf{e}. \quad (6.9)$$

The pseudo-inverse  $\mathbf{A}^\dagger$ , also called the Moore–Penrose generalised inverse, satisfies the relation  $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$  and is a generalisation of the inverse matrix. It allows the naive solution in the least-squares sense, derived from the requirement

$$\min_{\mathbf{x}} F(\mathbf{x}) \quad \text{with} \quad F(\mathbf{x}) = (\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{V}_y^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y}), \quad (6.10)$$

where the inverse of the data covariance matrix  $\mathbf{V}_y$  is included to take into account the accuracies which differ between the elements of the data vector. The least-squares solution from the normal-equations formalism can be expressed by the pseudo-inverse

$$\mathbf{A}^\dagger = \left( \mathbf{A}^T \mathbf{V}_y^{-1} \mathbf{A} \right)^{-1} \mathbf{A}^T \mathbf{V}_y^{-1}, \quad (6.11)$$

satisfying the requirement  $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$ . With defining  $\mathbf{C} = \mathbf{A}^T \mathbf{V}_y^{-1} \mathbf{A}$ , which is the Hessian matrix of  $F(\mathbf{x})$ , the estimate  $\hat{\mathbf{x}}$  and the covariance matrix  $\mathbf{V}_x$  are given by

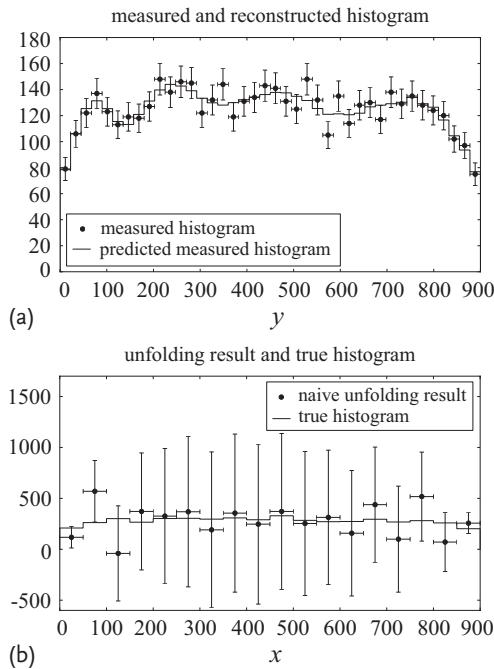
$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{A}^\dagger \mathbf{y}, \\ \mathbf{V}_x &= \mathbf{A}^\dagger \mathbf{V}_y \mathbf{A}^{\dagger T} = \left( \mathbf{A}^T \mathbf{V}_y^{-1} \mathbf{A} \right)^{-1} = \mathbf{C}^{-1}. \end{aligned} \quad (6.12)$$

An essential requirement is the acceptable description of the data  $\mathbf{y}$  by the distribution  $\hat{\mathbf{y}}$  calculated from the estimate  $\hat{\mathbf{x}}$  according to  $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$ . This agreement can, for example, be tested with a  $\chi^2$  test:

$$\chi_y^2 = (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{V}_y^{-1} (\hat{\mathbf{y}} - \mathbf{y}) \quad (6.13)$$

with  $m - n$  degrees of freedom.

Because  $\mathbf{A}^\dagger \mathbf{A} \equiv \mathbf{I}$ , the estimate  $\hat{\mathbf{x}}$  has the expectation  $\mathbf{x}_{\text{exact}}$  (see (6.9)). However, this naive solution is often not satisfactory. It can be strongly oscillating with large negative correlation coefficients between neighbouring points and large positive correlation coefficients between next-to-nearest neighbours, as demonstrated in Figure 6.2 for a measurement with Gaussian resolution equal to the bin width (see also Section 6.4.3): Figure 6.2a shows the measured distribution  $\mathbf{y}$ , which is



**Figure 6.2** Unfolding by naive least squares of an almost constant distribution with bin numbers  $m = 40$  and  $n = 18$ . The measured histogram  $y$  (a) is well described by the predicted measured histogram  $\hat{y}$ , but the naive unfolding result  $\hat{x}$  (b) is strongly fluctuating.

well described by the reconstructed distribution  $\hat{y}$ ; the unfolding result for this naive solution shows large oscillations (Figure 6.2b). The origin of this effect – which can be avoided using regularisation methods as described in Sections 6.2 and 6.3 – is discussed in Section 6.1.4.

The  $\chi^2$  test can be used in a MC simulation with a discretised version  $\mathbf{x}^{\text{test}}$  of an assumed true distribution  $f(t)^{\text{test}}$ , with

$$\chi_x^2 = (\hat{\mathbf{x}} - \mathbf{x}^{\text{test}})^T \mathbf{V}_x^{-1} (\hat{\mathbf{x}} - \mathbf{x}^{\text{test}}) \quad (6.14)$$

for  $n$  degrees of freedom. This test of the unfolding method is of course only possible if the covariance matrix  $\mathbf{V}_x$  is non-singular.

### 6.1.3

#### Unfolding Poisson-Distributed Data

In counting experiments the measured values  $y_i$  are integer numbers following a Poisson distribution about the value  $f_i(\mathbf{x}) = \mathbf{A}_i^T \mathbf{x}$  expected for the measured value  $y_i$ ; the expected value is a linear function of the components of  $\mathbf{x}$  (the vector  $\mathbf{A}_i^T$  is row  $i$  of the response matrix  $\mathbf{A}$ ). In this case the noise is correlated with the expectation.

If the  $y_i$  values are sufficiently large, one can use the Gaussian approximation of Poisson statistics and apply the least square method discussed in Section 6.1.2. Otherwise, the solution of unfolding can be found by an iterative minimisation of the *negative log-likelihood function* [9], based on Poisson probabilities:

$$\min_{\mathbf{x}} F(\mathbf{x}) \quad \text{with} \quad F(\mathbf{x}) = -\ln \mathcal{L} = \sum_{i=1}^m \left[ f_i - y_i - y_i \ln \left( \frac{f_i}{y_i} \right) \right]. \quad (6.15)$$

The function  $F(\mathbf{x})$  depends non-linearly on the parameters  $\mathbf{x}$ ; in each iteration a system of linear equations for corrections  $\Delta \mathbf{x}$  is constructed, that is  $\mathbf{H}\Delta \mathbf{x} = -\mathbf{g}$ , and solved for the corrections  $\Delta \mathbf{x}$ . The components of the gradient  $\mathbf{g}$  and the Hessian  $\mathbf{H}$  are given by:

$$g_j = \frac{\partial F}{\partial x_j} = \sum_i A_{ij} - \frac{y_i}{f_i} A_{ij}, \quad H_{jk} = \frac{\partial^2 F}{\partial x_j \partial x_k} = \sum_i \frac{y_i}{f_i^2} A_{ij} A_{ki}. \quad (6.16)$$

Convergence is typically reached after few iterations with  $\Delta \mathbf{x} \rightarrow 0$  and, as usual, the inverse of the Hessian is an estimate of the covariance matrix  $\mathbf{V}_{\mathbf{x}}$ .

#### 6.1.4

##### Convolution and Deconvolution

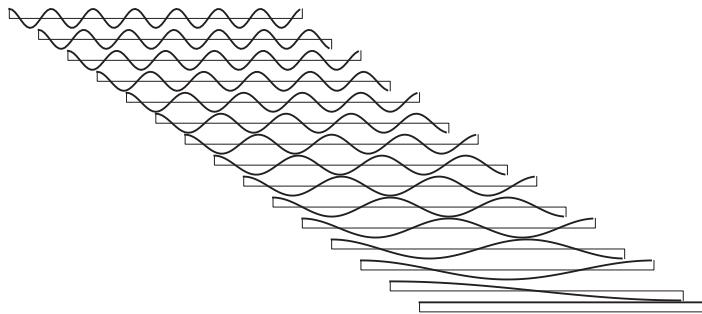
As discussed before in Section 6.1.2 and illustrated by Figure 6.2, the inverse process – the unfolding – without a specific parameterisation of the result is ill-posed as becomes apparent in the strong unphysical oscillations in (naive) solutions. The reason for this behaviour is the actual limitation of the number  $n$  of parameters that can be reconstructed, as is demonstrated below for a very simple case.

A function  $f(t)$  with period 1 can be approximated by a sum of cosine functions of the complete system  $\{1, \cos(\pi t), \cos(2\pi t), \dots\}$  which is periodic in  $[0, 1]$  and orthogonal in the interval  $0 \leq t \leq 1$ . These are the basis functions of the *discrete cosine transformation (DCT)*. The first functions of the system are shown in Figure 6.3. The approximation by  $n$  terms is given by

$$f(t) \approx a_0 + \sum_{k=1}^{n-1} a_k \cos(\pi k t). \quad (6.17)$$

The special case of a Kernel  $K(s, t) \equiv K(s - t)$  (see (6.1)) is called a *convolution*, and the corresponding inverse process is called *deconvolution*. Here a convolution of the function  $f(t)$  with a Gaussian resolution function (standard deviation  $\sigma$ ) is considered. For a single term  $\cos(\pi k t)$  the result of the convolution with the Gaussian is simple – the form of the term is not changed:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(s-t)^2}{2\sigma^2}\right] \cdot \cos(\pi k t) dt = \exp\left[-\frac{(\pi k \sigma)^2}{2}\right] \cdot \cos(\pi k s). \quad (6.18)$$



**Figure 6.3** The cosine functions  $\cos(\pi kt)$  for  $k = 0, 1, \dots, 15$  that form the basis functions of the discrete cosine transformation (DCT). See the text for more details.

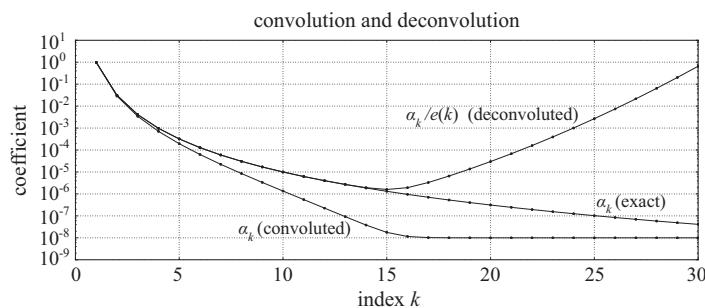
However, the amplitude of the cosine term is attenuated by an exponential factor  $e(k) = \exp[-(\pi k \sigma)^2/2]$ . This factor will become very much smaller than 1 for larger indices  $k$ . The convoluted function  $g(s)$  (see (6.1)) can again be approximated by a sum of cosine terms:

$$g(s) \approx a_0 + \sum_{k=1}^{n-1} \alpha_k \cos(\pi ks) . \quad (6.19)$$

Since the mentioned exponential factors – which form the new coefficients  $\alpha_k$  for the cosine terms in the sum of  $g(s)$  – decrease rapidly with increasing index  $k$ , the function  $g(s)$  will be *smoother* than the original  $f(t)$ . This is shown schematically in Figure 6.4 by the two lower curves.

Deconvolution is simple in this case: the relation between the coefficients  $\alpha_k$ , determined from  $g(s)$  and the coefficients  $a_k$  of the original function  $f(t)$  is given by

$$a_k = \frac{1}{e(k)} \cdot \alpha_k \quad \text{with} \quad e(k) = \exp\left[-\frac{(\pi k \sigma)^2}{2}\right] . \quad (6.20)$$



**Figure 6.4** Coefficients  $a_k$  of the exact function  $f(t)$ ,  $\alpha_k$  of the ‘measured’ (i.e. convoluted) function  $g(s)$ , and the coefficients  $a_k/e(k)$  of the deconvolution result. Because of round-

ing errors, the steeply falling coefficients  $\alpha_k$  reach an almost constant level of about  $10^{-8}$  for  $k > 15$ . Therefore, the convolution is meaningful only up to  $k = 15$ .

The extremely small true values of the coefficients  $\alpha_k$  for larger  $k$  cannot be reconstructed in practice from data on  $g(s)$ , due to rounding or measurement errors; they will reach a fixed finite lower level, as indicated in Figure 6.4 for the lowest curve. The value of the coefficients  $\alpha_k$ , reconstructed according to (6.20) by extremely large factors  $1/e(k)$  for large  $k$ , would dominate the result, as shown by the upper curve ( $\alpha_k/e(k)$ , ‘deconvoluted’). Thus, the number of terms of the original dependence  $f(t)$  which can be reconstructed is rather limited, even for well-known functions  $g(s)$ .

### 6.1.5

#### Parametrised Unfolding

So far unfolding was considered as a means to determine a discretised version  $\mathbf{x}$  of a distribution  $f(t)$  without a specific parameterisation. If a certain parameterisation  $f(t) \equiv f(t; \mathbf{a})$  depending on a vector of parameters  $\mathbf{a}$  is assumed – motivated for example by the theoretical analysis of the problem – this parameterisation can directly be used in unfolding. In contrast to unfolding without a specific parameterisation, where regularisation is in general necessary, no regularisation is required in this parameterised unfolding. In this case the aim of the measurement is to determine an estimate  $\hat{\mathbf{a}}$  of the parameter vector  $\mathbf{a}$ . For a given parameter vector  $\mathbf{a}$  the expected value of the observed bin content  $y_i$  can be calculated according to

$$y_i = \int_{s_{i-1}}^{s_i} ds g(s) = \int_{s_{i-1}}^{s_i} ds \left[ \int_{\Omega} dt K(s, t) f(t, \mathbf{a}) \right] \quad i = 1, 2, \dots, m. \quad (6.21)$$

Technically this calculation can be difficult, especially for the case of a MC simulation of the response. As an example the approximate calculation of the bin content  $y_i$  can be performed by the given response matrix  $\mathbf{A}$  with row vector  $\mathbf{A}_i^T$  (6.4) using the elements of an auxiliary vector  $\mathbf{x}$ :

$$y_i = \mathbf{A}_i^T \mathbf{x} \quad \text{with} \quad x_j(\mathbf{a}) = \int_{t_{j-1}}^{t_j} dt f(t; \mathbf{a}) \quad j = 1, 2, \dots, n, \quad (6.22)$$

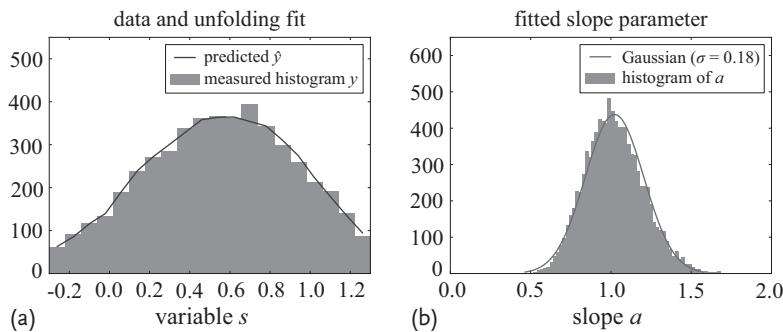
assuming a grid  $\{t_0, t_1, \dots, t_n\}$  for the variable  $t$ . Unfolding with a determination of an estimate  $\hat{\mathbf{a}}$  for the parameter vector is then the solution of the minimisation problem

$$\min_{\mathbf{a}} F(\mathbf{a}) \quad \text{with} \quad F(\mathbf{a}) = [\mathbf{A}\mathbf{x}(\mathbf{a}) - \mathbf{y}]^T \mathbf{V}_{\mathbf{y}}^{-1} [\mathbf{A}\mathbf{x}(\mathbf{a}) - \mathbf{y}]. \quad (6.23)$$

The function value  $F(\hat{\mathbf{a}}) = \chi^2_y$  should follow the  $\chi^2$  distribution with  $m - n_{\text{par}}$  degrees of freedom if the parameterisation has  $n_{\text{par}}$  parameters. A standard fit program like MINUIT [10] with numerical derivatives can be used to determine the parameter vector  $\hat{\mathbf{a}}$  and its covariance matrix.

### Example 6.1 Parameterised unfolding

An example of a parameterised unfolding, taken from [11], is shown in Figure 6.5. A pdf  $f(t) = (1 + at)/(1 + a/2)$  with  $t$  in the interval  $[0, 1]$  is measured with a Gaussian resolution with a standard deviation of 0.3. Figure 6.5a shows a simulated example for  $a = 1$  with 5000 entries in the measurement in the interval  $[-0.3, 1.3]$ . A 20-by-20 response matrix is determined by a simulation of 50 000 cases, using a uniform distribution (parameter  $a = 0$ ) in  $[0, 1]$ . The result of the parameter fit according to (6.23) in this example is  $\hat{a} = 1.09 \pm 0.18$ . Figure 6.5b shows the histogram of the fitted slope  $a$  from  $10^5$  simulated reconstructions together with a Gaussian curve of standard deviation 0.18; the fitted parameter is unbiased and has a slightly asymmetric distribution. These results agree with those of [11].



**Figure 6.5** Parametrised unfolding of the linear probability density function  $f(t) = (1 + at)/(1 + a/2)$ . (a) The measured histogram  $y$  and the predicted histogram  $\hat{y}$ ; (b) the histogram of the fitted slope  $a$  from  $10^5$  simulated reconstructions has a width of approximately 0.18. For more details see the text.

## 6.2 Solution with Orthogonalisation

Reliable solution procedures for ill-posed unfolding problems require orthogonal matrix transformations. In this section, the orthogonal decomposition of the rectangular response matrix  $\mathbf{A}$  and of the symmetric matrix  $\mathbf{C}$  of the least-squares normal equations, calculated from the response matrix  $\mathbf{A}$  (see Section 6.1.2), are discussed. These decompositions allow insight into the resolution properties of the response matrix  $\mathbf{A}$ . Least-squares methods based on the orthogonalisation have satisfactory stability properties and allow unphysical higher-order oscillation in the solution to be suppressed.

### 6.2.1 Singular Value and Eigenvalue Decomposition

**Singular value decomposition** The standard numerical method for the analysis of ill-posed problems  $\mathbf{Ax} = \mathbf{y}$  is the singular value decomposition [12, 13] of

the  $m$ -by- $n$  matrix  $\mathbf{A}$ , defined for any  $m$  and  $n$ . Assuming  $m \geq n$  (called the *thin SVD*) the SVD of  $\mathbf{A}$  is of the form

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad (6.24)$$

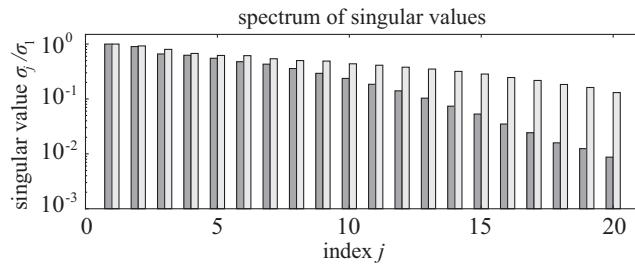
where  $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n) \in \mathbb{R}^{m \times n}$  and  $\mathbf{V}(\mathbf{v}_1, \dots, \mathbf{v}_n) \in \mathbb{R}^{n \times n}$  are matrices with orthonormal columns ( $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$ ) and the diagonal matrix  $\Sigma = \text{diag}\{\sigma_1, \dots, \sigma_n\} = \mathbf{U}^T\mathbf{A}\mathbf{V}$  has non-negative diagonal elements  $\sigma_i$ , called *singular values*, in decreasing order. The  $m$ -vectors  $\mathbf{u}_i$  and the  $n$ -vectors  $\mathbf{v}_i$  (the column-vectors of matrices  $\mathbf{U}$  and  $\mathbf{V}$ ) are called *left* and *right singular vectors* of  $\mathbf{A}$ . These singular vectors have, with increasing index (and decreasing singular value), an increasing number of sign-changes in their elements, corresponding to higher frequencies, similar to the cosine functions in Figure 6.3 of Section 6.1.4. The singular values  $\sigma_i$  correspond to the exponential factors defined in (6.20). SVD of a rectangular matrix  $\mathbf{A}$  into the orthogonal matrices  $\mathbf{U}$  and  $\mathbf{V}$  and the diagonal matrix  $\Sigma$  (6.24) is a standard algorithm of numerical linear algebra and software is available in scientific program libraries.

The matrix product  $\mathbf{Ax}$  expressed using the SVD matrices

$$\mathbf{Ax} = \mathbf{U}\Sigma\mathbf{V}^T\mathbf{x} = \sum_{j=1}^n \sigma_j (\mathbf{v}_j^T \mathbf{x}) \mathbf{u}_j = \mathbf{y} \quad (6.25)$$

shows that contributions to  $\mathbf{y}$  in the product  $\mathbf{Ax}$  with small singular values  $\sigma_j$ , corresponding to higher-frequency contributions, are suppressed. Components of  $\mathbf{x}$  in the direction of the unit vector  $\mathbf{v}_j$  are proportional to the singular value:  $\|\mathbf{Av}_j\| = \sigma_j$ . The *condition* of matrix  $\mathbf{A}$  is defined as the ratio of the largest to the smallest singular value:  $\text{cond}(\mathbf{A}) = \sigma_1/\sigma_n$ . As shown in Section 6.2.2 on unfolding by least squares using the SVD, the condition of  $\mathbf{A}$  is an upper bound on the magnification factor of the ratio of the errors of the estimate  $\hat{\mathbf{x}}$  relative to the errors of the data  $\mathbf{y}$  in least-squares solutions.

In order to take the uncertainties of the data  $\mathbf{y}$  as given by the covariance matrix  $\mathbf{V}_y$  into account, a *pre-scaling* (also called *pre-whitening*) of the linear system of equations is required before the SVD. For uncorrelated data this is achieved by dividing the rows of the linear system by the standard deviation  $\sqrt{(\mathbf{V}_y)_{ii}}$  of the data. The fastest method for correlated data is based on the Cholesky decomposition [13, 14] of the matrix  $\mathbf{V}_y = \mathbf{R}^T\mathbf{R}$  with an upper triangular matrix  $\mathbf{R}$ . The matrix  $\mathbf{A}$  and the data vector  $\mathbf{y}$  are transformed by  $\mathbf{R}^{-T}$ , the inverse of the transposed matrix  $\mathbf{R}^T$ . The covariance matrix of the transformed data vector  $\mathbf{R}^{-T}\mathbf{y}$  is  $\mathbf{R}^{-T}\mathbf{V}_y\mathbf{R}^{-1} = \mathbf{R}^{-T}\mathbf{R}^T\mathbf{R}\mathbf{R}^{-1} = \mathbf{I}$ , that is the unit matrix. In the following it is assumed that this pre-scaling of  $\mathbf{A}$  and  $\mathbf{y}$  has already been done (i.e. the elements of  $\mathbf{A}$  and  $\mathbf{y}$  are replaced by the elements of  $\mathbf{R}^{-T}\mathbf{A}$  and  $\mathbf{R}^{-T}\mathbf{y}$  respectively) before the singular value decomposition. The resulting scaled response matrix includes



**Figure 6.6** Spectrum of singular values of the response matrix  $\mathbf{A}$ . The bars are for a Gaussian resolution of  $\sigma_{\text{left}}$  equal to the bin width (left bar) and  $\sigma_{\text{right}} = \sigma_{\text{left}}/2$  (right bar).

the decomposed covariance matrix  $\mathbf{V}_y$  from the measurement, and the size of the singular values is proportional to  $\sqrt{N}$  if  $N$  is the number of measured events in an experiment. There is essentially no effect from the data statistics on the *shape* of the spectrum of singular vectors and on the matrix condition  $\sigma_1/\sigma_n$ ; the decomposed matrices represent solely the properties of the response matrix of the measurement process.

Figure 6.6 shows the spectrum of singular values for a Gaussian response matrix in two cases which differ only in the standard deviation. The decrease of the singular values is approximately described by the decrease of the exponential factor  $e(k)$  ((6.20) in the convolution example from Section 6.1.4). The left bars in the figure for a standard deviation  $\sigma_{\text{left}}$  of approximately the size of the bin width (assuming a number of bins equal to the number of coefficients) show the decrease of the singular values by a factor 100; the right bars for the smaller standard deviation  $\sigma_{\text{right}} = 1/2\sigma_{\text{left}}$  show a decrease by a factor of less than 10. Thus, the matrix condition  $\text{cond}(\mathbf{A})$  is directly related to the width of the Gaussian response function.

**Symmetric eigenvalue decomposition** An alternative but mathematically equivalent way in the framework of the normal-equations formalism of least squares is the symmetric eigenvalue decomposition. In this approach the symmetric  $n$ -by- $n$  matrix  $\mathbf{C}$  is determined from the  $m$ -by- $n$  matrix  $\mathbf{A}$  by the product  $\mathbf{C} = \mathbf{A}^T \mathbf{A}$  (assuming pre-scaling of  $\mathbf{A}$  and  $\mathbf{y}$  as discussed above). In the singular value decomposition of this square matrix  $\mathbf{C}$  the left and right singular vectors are identical:

$$\mathbf{C} = \mathbf{A}^T \mathbf{A} = (\mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T)^T \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T = \mathbf{V} \boldsymbol{\Sigma}^2 \mathbf{V}^T = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T. \quad (6.26)$$

The diagonal matrix  $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \dots\}$  has non-negative diagonal elements  $\lambda_i$ , called eigenvalues, equal to the square of the singular values  $\sigma_i$  of the matrix  $\mathbf{A} - \lambda_i = \sigma_i^2$ . The symmetric eigenvalue decomposition is the orthogonalisation method that is also used in maximum-likelihood methods [15, 16] (see Section 6.1.3).

### 6.2.2

#### Unfolding Using the Least Squares Method

##### 6.2.2.1 Least-Squares Method Using the SVD

The least-squares solution was expressed in (6.11) by the pseudo-inverse  $\mathbf{A}^\dagger$ . Written in terms of the SVD matrices the pseudo-inverse is equal to  $\mathbf{A}^\dagger = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T$  (all singular values have to be non-zero) with  $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$ . The least-squares estimate  $\hat{\mathbf{x}}$  is thus given by

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{-1}(\mathbf{U}^T \mathbf{y}) = \sum_{j=1}^n \frac{1}{\sigma_j} (\mathbf{u}_j^T \mathbf{y}) \mathbf{v}_j = \sum_{j=1}^n \left( \frac{c_j}{\sigma_j} \right) \mathbf{v}_j . \quad (6.27)$$

The data  $\mathbf{y}$  with unit covariance matrix  $\mathbf{V}_y = \mathbf{I}$  are transformed by  $\mathbf{U}^T$  to a  $n$ -vector  $\mathbf{c} = \mathbf{U}^T \mathbf{y}$  with unit covariance matrix  $\mathbf{V}_c = \mathbf{I}$ , representing the transformed measurement. The elements  $c_j = \mathbf{u}_j^T \mathbf{y}$  of  $\mathbf{c}$ , called *Fourier coefficients*, are elements of the measurement (in contrast to the  $\sigma_j$  that describe properties of the measurement apparatus). They tend to decrease rather fast towards small values for larger indices  $j$ , if the measured distribution  $\mathbf{y}$  is smooth [17]. The coefficients  $c_j$  are statistically independent (uncorrelated) and, having a variance of 1, show the significance of the corresponding contribution  $j$  to the estimate  $\hat{\mathbf{x}}$ . The value of a Fourier coefficient  $c_j$  will follow a Gaussian  $N(0, 1)$  if the exact value is small compared to the standard deviation 1. The expression (6.27) for the estimate  $\hat{\mathbf{x}}$  shows that the contribution to the estimate  $\hat{\mathbf{x}}$  related to a single Fourier coefficient  $c_j$  is multiplied by the inverse of the singular value  $\sigma_j$ . Insignificant Fourier coefficients with small singular values  $\sigma_j$  will therefore result in large fluctuations in the unfolding result  $\hat{\mathbf{x}}$  and can render the result unacceptable.

Because of the linear transformation of the data  $\mathbf{y}$  in the solution (6.27) the calculation of the uncertainty of the estimate  $\hat{\mathbf{x}}$  is straightforward; the covariance matrix is given by

$$\mathbf{V}_x = \mathbf{A}^\dagger \mathbf{V}_y \mathbf{A}^{\dagger T} = \mathbf{V}\boldsymbol{\Sigma}^{-2}\mathbf{V}^T = \sum_{j=1}^n \left( \frac{1}{\sigma_j^2} \right) \mathbf{v}_j \mathbf{v}_j^T . \quad (6.28)$$

In other – for example iterative – methods (Section 6.5) an estimate  $\hat{\mathbf{x}}$  is determined without the construction of a transformation matrix like  $\mathbf{A}^\dagger$ , which makes the above uncertainty calculation impossible.

##### 6.2.2.2 Null Space and Truncated SVD

The matrices  $\mathbf{U}$  and  $\mathbf{V}$  of the SVD define a new basis for the measured data and for the unfolding result in a frequency space. The measured data  $\mathbf{y}$  are transformed to independent Fourier coefficients  $c_j = \mathbf{u}_j^T \mathbf{y}$  with fixed standard deviation 1 (white noise). In analogy the least-squares estimate of (6.27),  $\hat{\mathbf{x}} = \sum_j d_j \mathbf{v}_j$ , is represented by coefficients  $d_j = c_j / \sigma_j$ . These coefficients  $d_j$  are still statistically independent, but have standard deviations  $1/\sigma_j$ , increasing with index  $j$ ; this property could be called *blue noise* because the uncertainty is increasing with the frequency.

Typically the singular values  $\sigma_j$  of a response matrix  $\mathbf{A}$  decrease without a clear gap between large and small singular values. Because of rounding and other errors no singular value will be exactly zero, but taking into account potential uncertainties of the elements of matrix  $\mathbf{A}$  at least a few singular values may be effectively zero, defining the *effective rank* of the matrix  $\mathbf{A}$  by a number  $p < n$ . The singular values will, in the case of uncertain elements of the matrix  $\mathbf{A}$ , remain close to their exact values [13, 14]. Especially if the response matrix is determined by a Monte Carlo simulation there are unavoidable larger uncertainties in the elements. A tolerance  $\delta$  can be defined [14] to determine the effective rank  $p$  by  $\sigma_p > \delta \geq \sigma_{p+1}$  with

$$\delta = \epsilon \cdot \max_{1 \leq i \leq m} \sum_{j=1}^n |A_{ij}|, \quad (6.29)$$

where for instance  $\epsilon = 0.01$  if the elements  $A_{ij}$  are correct to about two digits, as is the case for typical Monte Carlo calculations.

Small singular values  $\sigma_j < \delta$  would give meaningless random contributions to the solution  $\hat{\mathbf{x}}$ . Assuming an effective rank of  $p$  (less than  $n$ ), the estimate  $\hat{\mathbf{x}}$  of (6.27) can be written in the form

$$\hat{\mathbf{x}} = \underbrace{\sum_{j=1}^p d_j \mathbf{v}_j}_{\mathbf{x}_{\text{range}} \in \mathbb{R}^p} + \underbrace{\sum_{j=p+1}^n \tilde{d}_j \mathbf{v}_j}_{\mathbf{x}_{\text{null}} \in \mathbb{R}^{n-p}} \quad (6.30)$$

with two terms,  $\mathbf{x}_{\text{range}}$  and  $\mathbf{x}_{\text{null}}$ .

The first term  $\mathbf{x}_{\text{range}}$ , with contributions  $d_j \mathbf{v}_j = (c_j/\sigma_j) \mathbf{v}_j$  for  $j = 1, \dots, p$ , is a rather well-defined element of the  $p$ -dimensional subspace of the  $\mathbb{R}^n$ , but the second term  $\mathbf{x}_{\text{null}}$  has arbitrary contributions  $\tilde{d}_j \mathbf{v}_j$ ; multiplied by the response matrix  $\mathbf{A}$ , these terms have essentially no effect on the product  $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$ :

$$\mathbf{A}\hat{\mathbf{x}} = \sum_{j=1}^p \sigma_j d_j \mathbf{v}_j + \underbrace{\sum_{j=p+1}^n \sigma_j \tilde{d}_j \mathbf{v}_j}_{\approx 0}. \quad (6.31)$$

Because the two terms in  $\hat{\mathbf{x}} = \mathbf{x}_{\text{range}} + \mathbf{x}_{\text{null}}$  (6.30) are orthogonal, the squared norm of  $\hat{\mathbf{x}}$  is the sum of the two squared norms<sup>2)</sup>

$$\|\mathbf{x}_{\text{range}} + \mathbf{x}_{\text{null}}\|^2 = \|\mathbf{x}_{\text{range}}\|^2 + \|\mathbf{x}_{\text{null}}\|^2. \quad (6.32)$$

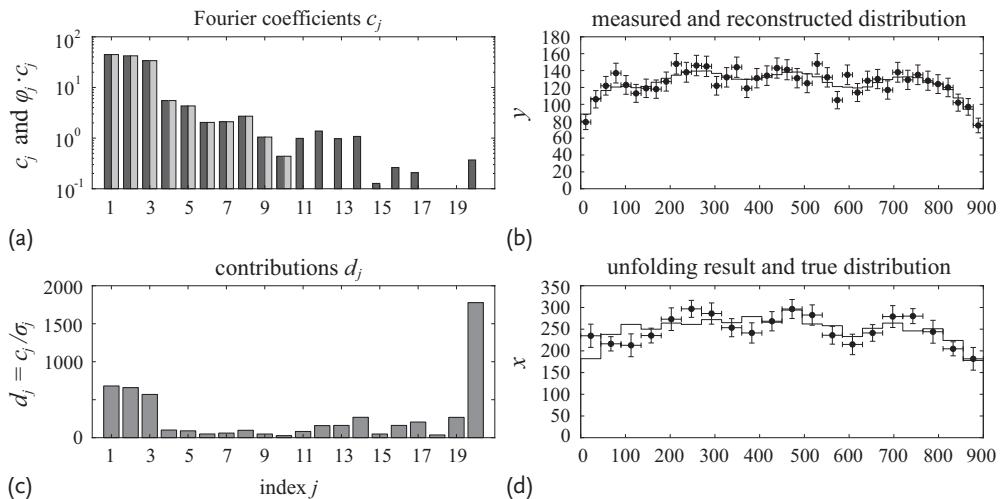
The solution recommended in textbooks (e.g. [18]) is the minimum-norm solution with  $\mathbf{x}_{\text{null}} = 0$  and  $\|\hat{\mathbf{x}}\| = \|\mathbf{x}_{\text{range}}\|$ . In this case the  $n$ -by- $n$  covariance matrix  $\mathbf{V}_x$  has a rank defect of  $n - p$  and cannot be inverted. Alternatively the dimension of estimate  $\hat{\mathbf{x}}$  can be reduced by rebinning to  $p$ , with a full-rank  $p$ -by- $p$  covariance matrix  $\mathbf{V}_x$  (see Section 6.6.4).

2) The used definition of the vector norm is  $\|\mathbf{x}\| = (\mathbf{x}^T \mathbf{x})^{1/2}$ .

The effective rank of the response matrix  $A$  defines an upper limit of the number of contributions, as illustrated in the following example.

### Example 6.2 Effective rank of the response matrix and number of contributions

In a simulation a data sample of 5000 events was generated, using numbers of bins of  $n = 20$  and  $m = 40$ , with a Gaussian response with a standard deviation of  $1/18$  of the range of the measured variable. The effective rank of  $A$  as determined from (6.29) with  $\epsilon = 0.01$  is 18. Figure 6.7a shows the Fourier coefficients  $c_j$ ; the coefficients  $c_j$  for  $j \geq 9$  are insignificant, giving a lower limit of the number of contributions. Figure 6.7c shows the coefficients  $d_j = c_j/\sigma_j$ ; there is an increase of the value of contributions  $d_j v_j$  for  $j \geq 10$ , with a very large contribution by  $d_{20}$  which would dominate the result. Truncation, ignoring the coefficients  $d_j$  after  $j = 10$ , will not bias the result. In the example, the second half of the coefficients (i.e.  $j > 10$ ) is ignored. Figure 6.7b shows the 40 bins of the measured distribution  $y$  as data points and the expected histogram  $\hat{y}$  after truncation; the agreement ( $\chi^2$ ) is acceptable. Figure 6.7d shows the result of the unfolding as data points together with the true histogram. Rebinning with the combination of pairs of neighbouring data points will reduce the uncertainties with a covariance matrix of full-rank.



**Figure 6.7** Distributions from a simulated sample of 5000 events. (a) The Fourier coefficients  $c_j$  before and after truncation; (b) the measured (data points) and reconstructed distributions  $y$  and  $\hat{y}$ ; (c) the contributions  $d_j/\sigma_j$ ; (d) the reconstructed distribution  $\hat{x}$  (data points) and the true histogram  $x$ .

As an alternative to assigning zero values to the coefficients  $\tilde{d}_j$  of the term  $x_{\text{null}}$  in the minimum-norm solution, they can also be given other, arbitrary values. One possibility is to assign plausible values to the  $\tilde{d}_j$  according to concepts like *maximum entropy* [19, 20]. In certain iterative unfolding methods (Section 6.5) an initial assumption  $x^{[0]}$  about the estimate  $\hat{x}$  is iteratively improved; the contributions cor-

responding to small singular values have a very slow convergence and the result after a few iterations may still contain large fractions of their initial contributions in  $\mathbf{x}^{[0]}$ , thus biasing the final result.

#### 6.2.2.3 Truncation and Positive Correlations

The  $\chi^2$  value  $\chi_y^2$  of the comparison of the measured distribution  $\mathbf{y}$  with the distribution  $\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}}$  that can be obtained from the unfolding result  $\hat{\mathbf{x}}$  has  $n_{\text{df}} = m - n$  degrees of freedom. The description of the measured distribution by the expected distribution is acceptable if the  $p$ -value from  $(\chi_y^2, n_{\text{df}})$  is not too small. If a single Fourier coefficient  $c_j$  is neglected, then the increase of the  $\chi_y^2$  value is equal to  $c_j^2$  for the orthogonalised solution with the SVD; at the same time the number of degrees of freedom increases by one. There will be no deterioration of the  $p$ -values if on average  $c_j^2 \approx 1$ , which is expected for insignificant coefficients.

One method to avoid the large fluctuations caused by small singular values  $\sigma_j$  is truncation, that is neglecting contributions with the smallest singular values if their Fourier coefficients  $c_j$  are statistically compatible with zero. The *truncated SVD* (or TSVD) solution is given by the first  $p$  contributions in the sum (6.30), with a corresponding reduction of the number of contributions of the covariance matrix  $\mathbf{V}_x$  according to (6.28). The rank of  $\mathbf{V}_x$  will be  $p$ , thus there is a rank defect  $n - p$ , and  $\mathbf{V}_x$  is singular. The use of only  $p$  contributions has the effect that *positive correlations* will appear in  $\mathbf{V}_x$  and the estimate  $\hat{\mathbf{x}}$  will appear rather smooth. This property is discussed in Section 6.6.4.

#### 6.2.3

##### Folding Versus Unfolding

Unfolding in the case of ill-posed problems can be avoided, if the measurement is made to test the prediction of a fixed model with the well-defined distribution  $f(t)^{\text{model}}$  without free parameters. In this case one can *fold*  $f(t)^{\text{model}}$  (i.e. the discretised version  $\mathbf{x}^{\text{model}}$ ) with the response matrix of the specific measurement to obtain the corresponding distribution  $\mathbf{y}^{\text{model}}$  (6.22), the measured distribution expected from the prediction. The statistical test of the prediction can be based on the uncertainty of the measured distribution. Conversely, the result of unfolding without a parameterisation is a *measured distribution*  $\hat{\mathbf{x}}^{\text{meas}}$  with a covariance matrix  $\mathbf{V}_x$  with correlations between elements, which have to be taken into account in a quantitative comparison with one or more model distributions  $\mathbf{x}^{\text{model}}$ .

Both folding and unfolding are influenced by small (or zero) singular values. In folding the expected measured distribution  $\mathbf{y}^{\text{model}} = \mathbf{A}\mathbf{x}^{\text{model}} = \sum_j \sigma_j (\mathbf{v}_j^T \mathbf{x}^{\text{model}}) \mathbf{u}_j$  is insensitive to contributions from the null space, and in unfolding the small singular values can deteriorate the reconstructed distribution  $\hat{\mathbf{x}} = \sum_j 1/\sigma_j (\mathbf{u}_j^T \mathbf{y}^{\text{meas}}) \mathbf{v}_j$ . If a given model  $f(t)^{\text{model}}$  has free parameters, a *parameterised unfolding* can be performed (Section 6.1.5) in order to obtain direct estimates for the model parameters from the measured distribution  $\mathbf{y}$ .

### 6.3

#### Regularisation Methods

The least squares solution by truncation (Section 6.2.2) with a sharp cut-off after a number of terms, determined by the effective rank of the response matrix  $A$ , can lead to the *Gibbs phenomenon* of oscillating components which is known from *finite Fourier sums*. As mentioned before, the singular values  $\sigma_j$  of the response matrix  $A$  decrease typically without a clear gap between large and small singular values. Some oscillations can be generated if a sharp cut-off between singular values of almost the same magnitude is used; the oscillations can be reduced by a *smooth* cut-off which is introduced in so-called *regularisation methods* with different *regularisation schemes* and which improves the solution.

##### 6.3.1

###### Norm and Derivative Regularisation

###### 6.3.1.1 Regularisation

The standard method for the solution of ill-posed problems is the regularisation method [21–23]. The expression to be minimised with respect to the unfolding result includes the least-squares and the (negative) log-likelihood expressions (6.10) and (6.15), which ensure a good description of the measured distribution. A second term  $\Omega(x)$ , often of the form  $\Omega(x) = \|Lx\|^2$  with a certain matrix  $L$ , ensures certain properties like smoothness of the unfolding result and contributes with a weight, given by a *regularisation parameter*  $\tau > 0$ :

$$\min_x (F(x) + \tau \|Lx\|^2). \quad (6.33)$$

In the regularised solution of the least-squares case (6.10) the matrix  $A^\dagger$  is replaced by the regularised matrix  $A^\#$ :

$$\hat{x} = A^\# y = [(A^T A + \tau L^T L)^{-1} A^T] y. \quad (6.34)$$

The regularisation term  $\tau L^T L$  is added to the matrix  $C = A^T A$  of the normal equations, and inserting  $y = Ax_{\text{exact}} + e$  one obtains

$$\hat{x} = A^\# A x_{\text{exact}} + A^\# e = x_{\text{exact}} + \underbrace{(A^\# A - I)x_{\text{exact}}}_{\text{systematic error}} + \underbrace{(A^\# e)}_{\text{statistical error}}. \quad (6.35)$$

The product  $\Xi \equiv A^\# A$  is called the *resolution matrix*. In the regularisation scheme the resolution matrix is not equal to the unit matrix, and thus the method has a systematic bias  $(\Xi - I)x_{\text{exact}}$ . The fact that the regularised solution  $\hat{x}$  has a potential bias, which depends on the details of the exact distribution,  $x_{\text{exact}}$ , is connected with the attempt to reduce the unnatural and unmeasurable oscillations. The *smoothing* effect of the resolution matrix gives no or small systematic errors for *smooth* exact distributions, and large systematic deviations for *unphysically oscillating* distributions. The measured distribution  $y$  has to be compared with the distri-

bution  $\hat{y}$  (see (6.22)) corresponding to the estimated reconstructed distribution  $\hat{x}$  and given by

$$\hat{y} = A\hat{x} = (AA^\#) y, \quad (6.36)$$

where the  $m$ -by- $m$  product matrix  $(AA^\#)$  is often called the *influence matrix*. The agreement between the measured data  $y$  and the vector  $\hat{y}$ , calculated in (6.36), has to be acceptable. It can – for example – be checked with  $\chi^2_y = (\hat{y} - y)^T(\hat{y} - y)$ .

The deviation of the resolution matrix  $\Xi = A^\#A$  from the unit matrix  $I$ , that is the potential bias, should avoid the unnatural properties of naive unregularised solutions. The regularisation ansatz can be used to separate the *significant* from the *insignificant* contributions of the result *without the introduction of a sizable bias*.

### 6.3.1.2 Norm Regularisation

The simplest regularisation method is *norm regularisation* with  $L = I$ . For a given value of  $\tau$  the estimate  $\hat{x}$  can be determined by standard methods of linear algebra (matrix inversion), because the condition of the combined matrix in (6.34) with the regularisation term  $L^T L = I$  is good. However, the numerical solution by the SVD is simple in this case (the regularisation matrix is diagonal) and has several advantages, especially as it allows a clear understanding of the effects of regularisation to be obtained: it is equivalent to the introduction of filter factors for the orthogonal contribution to the solution  $\hat{x}$ , depending on the singular values  $\sigma_j$  (and on the regularisation parameter  $\tau$ ). Using the SVD the solution can be written in the form

$$\hat{x} = V \underbrace{[(\Sigma^2 + \tau I)^{-1} \Sigma^2]}_{\text{filter factor matrix } F} \Sigma^{-1} \underbrace{(U^T y)}_{\text{coeff. } c} = (V F \Sigma^{-1} U^T) y, \quad (6.37)$$

where the additional matrix  $F$  is diagonal with elements equal to filter factors  $\varphi_j$  (see (6.27)). The estimate  $\hat{x}$  can be expressed by a sum:

$$\hat{x} = \sum_{j=1}^n \frac{c_j}{\sigma_j} \varphi_j v_j \quad \text{with} \quad \varphi_j = \frac{\sigma_j^2}{\sigma_j^2 + \tau} \equiv \frac{\lambda_j}{\lambda_j + \tau} \quad (6.38)$$

(the squared singular values  $\sigma_j^2$  are replaced by the eigenvalues  $\lambda_j$  in case of diagonalisation). The effect of the regularisation is thus the introduction of a filter factor  $\varphi_j$  for each term with a strength which depends on the regularisation parameter  $\tau$ . The filter factors  $\varphi_j$  appear also in the expression of the covariance matrix (compare (6.28)):

$$V_x = \sum_{j=1}^n \frac{1}{\sigma_j^2} \varphi_j^2 v_j v_j^T. \quad (6.39)$$

For  $\tau = \sigma_k^2$  the filter factor  $\varphi_k$  has a value of  $1/2$  and thus the Fourier coefficient  $c_k$  will be reduced to half the original value. For the first coefficients  $c_j$  with large singular values  $\sigma_j$  the filter factor  $\varphi_j \approx 1$  will make almost no change, but

all terms with small singular value will be reduced, avoiding the dominating influence of these terms on the result. No bias will be introduced if the selected regularisation parameter  $\tau$  is small enough to reduce only the insignificant Fourier coefficients. In the formalism one could also apply a different filter definition, for example  $\varphi_j = 1/[1 + (\tau/\sigma_j^2)^\alpha]$ , which for  $\alpha > 1$  induces a sharper transition from 1 to 0. For large values  $\alpha \gg 1$  essentially a sharp cut-off (truncation) is reached. Note that a sharp cut-off (truncation) – the simplest regularisation scheme – can lead to the Gibbs phenomenon of strongly oscillating components which is known from finite Fourier sums; this phenomenon can be avoided by a smoother cut-off. The dependence of the filter factor for various filter methods is illustrated in Figure 6.8.

The norm regularisation corresponds to the original regularisation proposal by Tikhonov [22, 23] and Philipps [21]. The regularisation parameter  $\tau$  can be interpreted as the introduction of the a priori measurement error  $s_{\text{reg}} = 1/\sqrt{\tau}$  for each component of the vector  $\mathbf{x}$ . Individual values  $s_{j,\text{reg}}$  for the different components  $x_j$  could be introduced, corresponding to a regularisation term  $\Omega(\mathbf{x}) = \sum_j x_j^2/s_j^2$ .

The scheme above can be used for unfolding problems with rather smooth solutions  $\tilde{\mathbf{x}}$ , requiring only a small number of Fourier coefficients; in other cases, especially for data with a high relative precision, some modifications are advisable. One possibility is to change the regularisation term  $\Omega(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|^2$  to a term

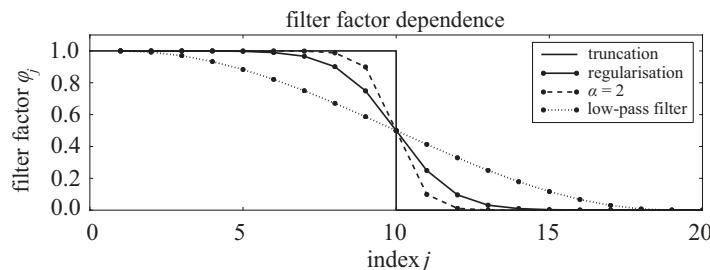
$$\Omega(\mathbf{x}) = \|\mathbf{L}(\mathbf{x} - \mathbf{x}_0)\|^2 \quad (6.40)$$

with some a priori assumption  $\mathbf{x}_0$  on the resulting vector  $\mathbf{x}$ ; this will reduce the number of significant terms.

Another possibility is to amend the Monte Carlo simulation to include a realistic a priori assumption  $f(t)^{\text{model}}$  about the function  $f(t)$  and to already include the function  $f(t)^{\text{model}}$  in the definition of the response matrix:

$$\int_{\Omega} [K(s, t) f(t)^{\text{model}}] f^*(t) dt = g(s). \quad (6.41)$$

In this way, only an almost constant correction function  $f^*(t)$  has to be determined with  $f(t) = f(t)^{\text{model}} f^*(t)$ . This option is available in unfolding methods of particle



**Figure 6.8** Dependence of the filter factor  $\varphi_j$  for various filter methods, all with the value 1/2 at  $j = 10$ , and for truncation at  $j = 10$ . The filter factor labelled *regularisation* is from

formula (6.38), the factor with  $\alpha = 2$  is from the formula given in the text. The low-pass filter is explained in Section 6.4.3.

physics [15, 16, 24]. The elements of this redefined matrix  $\mathbf{A}$  are now integers (and not conditional probabilities as before, see Section 6.1.2), the number of Monte Carlo events from bin  $j$  of  $\mathbf{x}$ , measured in bin  $i$  of  $\mathbf{y}$ .

### 6.3.1.3 Regularisation Based on Derivatives

Another regularisation scheme is based on *derivatives*, and this scheme often has advantages over the norm regularisation. Most popular are the second derivatives, but first and third derivatives are used too. The matrix  $\mathbf{L}$  is rather simple if equidistant bins are used. For example, the second derivative in bin  $j$  is approximately proportional to  $(-x_{j-1} + 2x_j - x_{j+1})$ . In the often-used matrix

$$\mathbf{L}_2^r = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n \times n} \quad (6.42)$$

the second derivatives at the inner bins are supplemented by the first derivative in the first and last bin, resulting in a symmetric matrix.

The solution (6.34) can again be obtained, for a given regularisation factor  $\tau$ , by matrix inversion. The orthogonal solution is formally equivalent to the solution (6.34), with a different definition of the singular or eigenvalues. Compared to the norm regularisation, the Fourier coefficients refer to a rotated system according to the symmetric product matrix  $\mathbf{L}^T \mathbf{L}$ . A solution using orthogonalisation methods allow a better understanding of the details and the separation of significant from insignificant contributions, but is numerically more complicated than the norm regularisation because the term  $\tau \mathbf{L}^T \mathbf{L}$  is not diagonal. The use of the SVD would require a generalised SVD version [25].

If, instead of the SVD, the symmetric eigenvalue decomposition is used, two rotations (and a scaling) are required to diagonalise simultaneously the two symmetrical matrices  $\mathbf{C} = \mathbf{A}^T \mathbf{A}$  and  $\mathbf{L}^T \mathbf{L}$  for the solution of the normal equation [5]

$$(\mathbf{C} + \tau \mathbf{L}^T \mathbf{L}) \mathbf{x} = \mathbf{b} \quad (6.43)$$

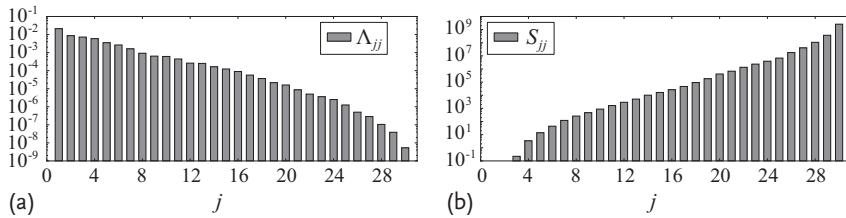
with  $\mathbf{b} = \mathbf{A}^T \mathbf{y}$ . The first diagonalisation  $\mathbf{C} = \mathbf{U}_1 \mathbf{A} \mathbf{U}_1^T$  is used to rewrite the equation in the form

$$\mathbf{U}_1 \mathbf{A}^{1/2} (\mathbf{I} + \tau \mathbf{M}) \mathbf{A}^{1/2} \mathbf{U}_1^T \mathbf{x} = \mathbf{b} \quad (6.44)$$

with the transformed regularisation matrix  $\mathbf{M} = \mathbf{A}^{-1/2} \mathbf{U}_1^T (\mathbf{L}^T \mathbf{L}) \mathbf{U}_1 \mathbf{A}^{-1/2}$ . The second diagonalisation  $\mathbf{M} = \mathbf{U}_2 \mathbf{S} \mathbf{U}_2^T$  can be used to rewrite the equation in the form

$$\mathbf{R} (\mathbf{I} + \tau \mathbf{S}) \mathbf{R}^T \mathbf{x} = \mathbf{b} \quad (6.45)$$

$$\hat{\mathbf{x}} = \mathbf{R}^{-T} \underbrace{(\mathbf{I} + \tau \mathbf{S})^{-1}}_{\text{filter factor matrix } \mathbf{F}} \underbrace{\mathbf{R}^{-1} \mathbf{b}}_{\text{coeff. } \mathbf{c}} \quad (6.46)$$



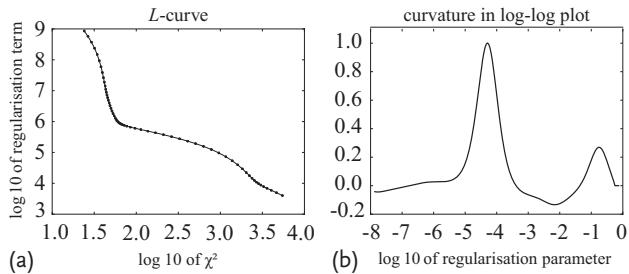
**Figure 6.9** Exemplary eigenvalues  $A_{jj}$  (a) and  $S_{jj}$  (b). The definition of the elements  $S_{jj}$  is inverse to the definition of the elements  $A_{jj}$ ; the first two elements  $S_{11}$  and  $S_{22}$  from linear contributions are zero.

using the matrix  $\mathbf{R} = \mathbf{U}_1 \mathbf{A}^{1/2} \mathbf{U}_2$  and the inverse  $\mathbf{R}^{-1} = \mathbf{U}_2^T \mathbf{A}^{-1/2} \mathbf{U}_1^T$ . The filter factor is now given by  $\varphi_j = 1/(1 + \tau S_{jj})$  with the element  $S_{jj}$  of the diagonal matrix  $\mathbf{S}$ . Figure 6.9 shows the eigenvalues  $A_{jj}$  and  $S_{jj}$  for the Example 6.3 discussed below, both with increasing frequency from the left to the right; the stronger separation of low- and high-frequency contributions by the curvature is visible. Note that the definition of the elements  $S_{jj}$  is inverse to the definition of the elements  $A_{jj}$ , and the first two eigenvalues  $S_{11}$  and  $S_{22}$ , corresponding to a constant and to a linear contribution (without a curvature), are zero.

#### 6.3.1.4 Determination/Selection of the Regularisation Parameter

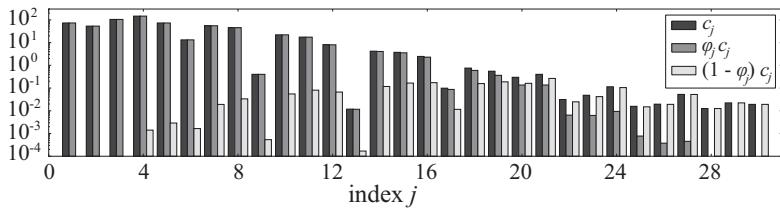
There is no generally accepted and unique method to determine the regularisation parameter  $\tau$  that is applicable to all cases. An upper limit of the regularisation parameter  $\tau$  is determined by the overall  $\chi^2$  of the agreement of the measured distribution  $\gamma$  with the predicted distribution  $\hat{\gamma}$  (6.13). The smallest  $\chi^2_\gamma$  value, with  $n_{\text{df}} = m - n$ , is obtained for  $\tau = 0$ , and this should correspond to an acceptable  $p$ -value. Each Fourier coefficient that is removed in the truncation method will increase the  $\chi^2$  value by  $c_j^2$  and  $n_{\text{df}}$  by one. As long as the coefficients  $c_j$  are compatible with zero (and the mean value of  $c_j^2$  is one), the  $p$ -value will not change significantly. In a wide range of values of the regularisation parameter  $\tau$  the  $p$ -value corresponding to  $\chi^2_\gamma$  and  $n_{\text{df}}$  will be acceptable. The  $p$ -value will decrease towards very small values if significant Fourier coefficients are removed; this defines the upper limit of  $\tau$ . In other fields than high energy physics a discrepancy principle attributed to Morozov [26] is sometimes used, where the choice of  $\tau$  is only based on the  $\chi^2_\gamma$  value.

It is recommended that one studies the dependence of several quantities like squared curvature of  $\hat{x}$  and average correlation of the elements of the covariance matrix  $\mathbf{V}_x$  on the value of the parameter  $\tau$  in repeated solutions over a wide range of  $\tau$  values that lead to acceptable  $p$ -values. A standard method mentioned in textbooks is the *L*-curve method: for each solution the  $\chi^2_\gamma$  value is plotted versus the squared curvature value (for the case where  $\mathbf{L}$  corresponds to the second derivative) in a log-log plot, as shown in Figure 6.10. The resulting shape often resembles the shape of an ‘*L*’, showing a distinct corner. For small values of  $\tau$ , where negative correlations dominate, the  $\chi^2_\gamma$  value is not large, but the squared curvature is. For



**Figure 6.10** The logarithm of the regularisation term is plotted versus the logarithm of the  $\chi^2$  value (a), for a large number of different values of the regularisation parameter  $\tau$ .

The curvature of this plot versus the logarithm of  $\tau$ , plotted in (b), is used to determine the optimal  $\tau$  value from the point with the largest curvature.



**Figure 6.11** Fourier coefficients  $c_j$ , the filtered coefficients  $\varphi_j c_j$  and their difference plotted for the example of a steeply falling distribution.

large values of  $\tau$ , where positive correlations between elements of  $\hat{x}$  dominate, the  $\chi^2_\nu$  value becomes larger and the squared curvature small. The recommendation is to select the value of  $\tau$  with the largest curvature in the log–log plot. The use of such plots for ill-conditioned least-squares problems was already mentioned in the classical book [27] on least-squares methods.

### Example 6.3 A steeply falling distribution

An example of a difficult unfolding problem is the measurement of the inclusive jet production cross section as a function of the jet transverse momentum,  $p_T$ , in collisions at very high energy, for example [28]. The distribution is steeply falling. The transverse momentum  $p_T$  as measured with the calorimeter is systematically underestimated. The corresponding bias and the accuracy of the measurement can be determined in a MC simulation. In the publication [28] corrections for the bias and the limited resolution are done in two separate steps, using essentially a bin-by-bin correction method (see Section 6.5). In a simple MC calculation a problem with similar properties is simulated, applying unfolding according to the method of Section 6.3.1.3. A pure exponential distribution is assumed with a systematic bias of the measured  $p_T$  value to smaller values up to 10% and a Gaussian smearing with a standard deviation  $\sigma(p_T)$ , given by  $\sigma(p_T)/p_T = 100\%/\sqrt{p_T \text{ (GeV)}}$ . In addition a trigger acceptance with a rapid decrease below 100 GeV is assumed. Because the

$p_T$  distribution at low values of  $p_T$  is unmeasurable, the measured and unfolded  $p_T$  ranges are restricted to the interval 64–400 GeV. A realistic model function is assumed and used for the determination of the response matrix  $\mathbf{A}$ , necessary for the unfolding with determination of  $\hat{\mathbf{x}}$  and  $\mathbf{V}_x$ . A separate acceptance correction is performed after the unfolding. The unfolding is performed in the transformed variable  $q_T = \sqrt{p_T}$ , which has the advantage of a *constant* standard deviation  $\sigma(q_T) = 0.5$ , with a back-transformation to  $p_T$  after unfolding, resulting in a bin width that increases with  $p_T$ . The Fourier coefficients without and with a filter factor are shown in Figure 6.11. The change of the coefficients by filtering is always less than the statistical uncertainty (the standard deviation has the value 1) – thus essentially no bias is introduced. The true, measured and unfolded distributions are shown in Figure 6.12; below 75 GeV the errors are larger than the cross-section values (not displayed for acceptance-corrected results). The example shows that true unfolding corrects for the bias and the limited resolution (and other effects, if present) in a single step, allowing the determination of the full covariance matrix  $\mathbf{V}_x$  with a consistent propagation of the uncertainties.

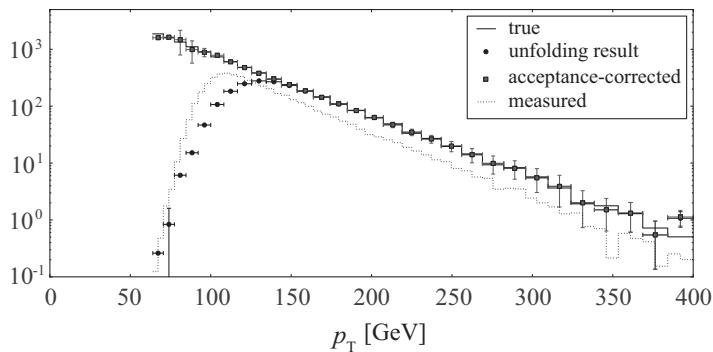


Figure 6.12 Unfolding of a steeply falling distribution. For more details see the text.

## 6.4

### The Discrete Cosine Transformation and Projection Methods

As an alternative to the SVD of the actual response matrix  $\mathbf{A}$  a fixed  $n$ -by- $p$  matrix  $\mathbf{W} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p\}$  is used in *projection* methods. This alternative with a *fixed* transformation is advisable if different versions of response matrices  $\mathbf{A}$  have to be tested, for example to check the influence of certain systematic uncertainties on the calculated response matrix. The discrete cosine transformation (DCT), introduced in Section 6.4.1, can be used to define the column vectors of the transformation matrix  $\mathbf{W}$ , which is usually not much different from the singular vectors calculated from the response matrix  $\mathbf{A}$  by SVD.

## 6.4.1

**Discrete Cosine Transformation**

The *discrete cosine transform* expresses a finite sequence of  $n$  real numbers  $f_k$ ,  $k = 0, 1, \dots, (n - 1)$ , in terms of a sum of cosine functions with different frequencies. The transformation is similar to the *discrete Fourier transform* (DFT), but using real numbers. The DCT is often used in signal processing when most of the signal information is contained in the low-frequency components, and allows a separation of significant and insignificant components. This property is used for an efficient data compression in signal and image processing. The discrete cosine transform of a vector  $\mathbf{f}$  to a vector  $\mathbf{c}$  of coefficients is performed by forming the product of vector  $\mathbf{f}$  with an orthogonal matrix  $\mathbf{U}_{\text{DCT}}$  that in the most common DCT-II variant has elements

$$U_{jk} = \begin{cases} \sqrt{1/n} & k = 0 \\ \sqrt{2/n} \cos[\pi k(j + 1/2)/n] & k = 1, 2, \dots, n - 1 \end{cases} \quad (6.47)$$

(there are eight variants with different boundary conditions) with the property  $\mathbf{U}_{\text{DCT}}^T \mathbf{U}_{\text{DCT}} = \mathbf{I}$ . The DCT is interesting for unfolding problems because of the relations

$$\mathbf{L}_2^r = \mathbf{U}_{\text{DCT}} \mathbf{A} \mathbf{U}_{\text{DCT}}^T, \quad (\mathbf{L}_2^r)^T \mathbf{L}_2^r = \mathbf{U}_{\text{DCT}} \mathbf{A}^2 \mathbf{U}_{\text{DCT}}^T \quad (6.48)$$

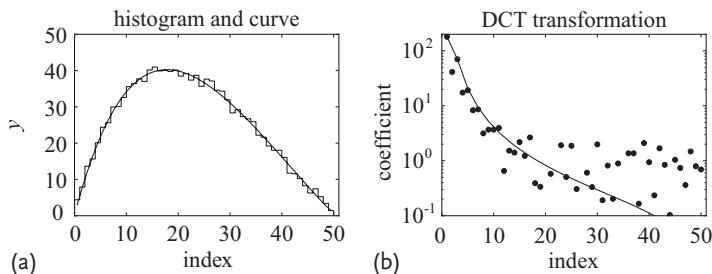
between  $\mathbf{U}_{\text{DCT}}$  and the second-derivative matrix  $\mathbf{L}_2^r$  of (6.42), which is square and symmetric. The eigenvectors of matrix  $\mathbf{L}_2^r$  (and of the product  $(\mathbf{L}_2^r)^T \mathbf{L}_2^r$ ) are the column vectors of the matrix  $\mathbf{U}_{\text{DCT}}$ , and the eigenvalues of matrix  $\mathbf{L}$ , the diagonal elements of  $\mathbf{A}$ , are given by

$$\lambda_k = 4 \sin^2 \left( \frac{k\pi}{2n} \right) \quad k = 0, 1, \dots, n - 1 \quad (6.49)$$

with  $\lambda_0 = 0$ .

The DCT has the property of *separability*: the transformation of a two-dimensional array is the one-dimensional DCT, performed in a row-column algorithm first along the rows and then along the columns (or vice versa). The result is a two-dimensional array of coefficients, selected for example as the basis for the JPEG algorithm of image compression.

The cosine functions used in the approximation (6.17) of a function  $f(t)$  and shown in Figure 6.3 are the basis functions of the DCT. A set of equidistant values  $f$  of a function  $f(t)$  is transformed to a set of coefficients  $c$ , and this process corresponds to the function approximation of (6.17). The transformation between the vector  $\mathbf{f}$  and the vector  $\mathbf{c}$  is performed by the DCT matrix  $\mathbf{U}_{\text{DCT}}$ . The vector  $\mathbf{c}$  is obtained by the transformation  $\mathbf{c} = \mathbf{U}_{\text{DCT}}^T \mathbf{f}$  and the vector  $\mathbf{f}$  can be obtained from the vector  $\mathbf{c}$  by the back-transformation  $\mathbf{f} = \mathbf{U}_{\text{DCT}} \mathbf{c}$ . The DCT transformation is rather similar to the typical transformation from the data to Fourier coefficients in the singular value decomposition.



**Figure 6.13** (a) The curve of an exact distribution and a simulated histogram with unit covariance matrix  $V_x = \mathbf{I}$ . (b) The DCT coefficients from the exact distribution (curve) and the coefficients obtained from the histogram (small filled circles).

The separation of the low- and the high-frequency contributions by the DCT is demonstrated in Figure 6.13. Figure 6.13a shows a true distribution as a curve and a simulated measurement  $f$  as a histogram with 50 bins, with a covariance matrix equal to  $\mathbf{I}$ , that is the standard deviation of each bin is 1. The coefficients obtained by the DCT are shown in Figure 6.13b: the curve shows the fast decrease of the coefficients of the underlying true distribution. The coefficients from the data, shown as points, have the same unit covariance matrix as the data because of the orthogonal transformation, representing so-called *white noise*. The plot of the coefficients clearly shows the separation of the different frequency contributions. The coefficients decrease for increasing index  $k$  and reach a level of 1, the statistical error of the frequency coefficients, at  $k \approx 20$ . The calculated coefficients  $c_k$  for high frequencies with  $k > 20$  are caused by the statistical fluctuations and are not part of the signal, and they can be replaced by zero *without* introducing a bias. The back-transformation from  $\approx 20$  non-zero coefficients to 50 histogram bins yields a *smoothed* histogram in which statistical fluctuations are reduced. After back-transformation each *individual* bin content appears to be more precise than before, it has a *smaller* variance, as calculated by a transformation of the covariance matrix, compared to the original bin. The truncation with the removal of noise thus has two effects, without introducing a bias: individual bin contents are more precise, but positive correlations are introduced between bin contents. This kind of smoothing is related to the effects of truncation (Section 6.2.2.2) and of regularisation (Section 6.3) in unfolding. It explains the surprising effect mentioned above that the value of an individual bin after unfolding can appear to be more precise than before unfolding. Local uncertainties are indeed reduced, at the cost of the introduction of positive correlations, without changing the overall information.

#### 6.4.2 Projection Methods

A fixed  $n$ -by- $p$  transformation matrix  $\mathbf{W} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p\}$  is used in *projection* methods. An appropriate choice for the column vectors of  $\mathbf{W}$  are the first  $p$  column vectors of the DCT matrix  $\mathbf{U}_{\text{DCT}}$ . A  $p$ -vector  $\bar{\mathbf{x}}$  is introduced by the replacement  $\mathbf{x} \rightarrow$

$\mathbf{W}\bar{\mathbf{x}}$ . Thus, the  $n$ -vector  $\mathbf{x}$  is replaced by  $\mathbf{W}\bar{\mathbf{x}}$  in the least-squares expression (6.10):

$$F(\bar{\mathbf{x}}) = (\mathbf{A}\mathbf{W}\bar{\mathbf{x}} - \mathbf{y})^T \mathbf{V}_y^{-1} (\mathbf{A}\mathbf{W}\bar{\mathbf{x}} - \mathbf{y}) = (\bar{\mathbf{A}}\bar{\mathbf{x}} - \mathbf{y})^T \mathbf{V}_y^{-1} (\bar{\mathbf{A}}\bar{\mathbf{x}} - \mathbf{y}). \quad (6.50)$$

In this function  $F(\bar{\mathbf{x}})$  of the  $p$ -vector  $\bar{\mathbf{x}}$  the response matrix  $\mathbf{A}$  becomes the transformed response matrix  $\bar{\mathbf{A}} = \mathbf{A}\mathbf{W}$ . The transformation is also called *preconditioning*<sup>3)</sup> (with a multiplication from the right). The vector  $\bar{\mathbf{x}}$ , obtained by the minimisation of  $F(\bar{\mathbf{x}})$ , and eventually after multiplying the components  $\bar{x}_j$  by filter factors  $\varphi_j$  (6.38), is then transformed back to the unfolding estimate by  $\hat{\mathbf{x}} = \mathbf{W}\bar{\mathbf{x}}$ .

A transformation can also be done from the left by multiplication of the equation  $\mathbf{A}\mathbf{x} = \mathbf{y}$  with a suitable matrix, for example  $\mathbf{W}^T$  from the left, and both transformations can be combined:

$$\mathbf{W}^T \mathbf{A} \mathbf{W} \bar{\mathbf{x}} = \bar{\mathbf{A}} \bar{\mathbf{x}} = \mathbf{W}^T \mathbf{y} = \bar{\mathbf{y}}. \quad (6.51)$$

The effect of the transformation of the response matrix  $\mathbf{A}$  to the almost *diagonal* transformed response matrix  $\bar{\mathbf{A}} = \mathbf{W}^T \mathbf{A} \mathbf{W}$  is similar to the effect achieved by the SVD. Regularisation with filter factors  $\varphi_j$  as well as truncation is possible. The use of the projection method with its fixed transformation matrix  $\mathbf{W}$  can be recommended to obtain stable results in tests of different versions of the response matrix  $\mathbf{A}$ , for example from different assumptions in the MC simulation used to determine the response matrix  $\mathbf{A}$ .

#### 6.4.3

##### Low-Pass Regularisation

Instead of the regularisation as described in Section 6.3 the application of a certain smoothing to the *oscillating* result  $\tilde{\mathbf{x}}$  of naive unfolding by redefining the elements  $x_j$  according to  $x_j = 1/4\tilde{x}_{j-1} + 1/2\tilde{x}_j + 1/4\tilde{x}_{j+1}$  has been proposed [29, 30]. This type of smoothing could even be optimised by an adjustment of the factor  $a_j$  in the smoothing  $x_j = a_j\tilde{x}_{j-1} + (1 - 2a_j)\tilde{x}_j + a_j\tilde{x}_{j+1}$  from the elements of the covariance matrix  $V_x$  of the naive result, but the factors  $a_j$  usually don't differ much from the simple value  $a_j = 1/4$ . This low-pass regularisation can be approximately described by the transformation  $\mathbf{x} = \mathbf{T}\tilde{\mathbf{x}}$  with a square symmetric low-pass matrix, for example for  $n = 5$ :

$$\mathbf{T} = \frac{1}{4} \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \in \mathbb{R}^{5 \times 5}. \quad (6.52)$$

This  $n$ -by- $n$  matrix  $\mathbf{T}$  has a closed form expression for the eigenvalue decomposition  $\mathbf{T} = \mathbf{S} \mathbf{A} \mathbf{S}^T$ , where the eigenvectors of  $\mathbf{T}$  are the column vectors of the orthogonal

3) Preconditioning is used to improve the convergence of iterative methods; a suitable matrix  $\mathbf{W}$  improves the condition of the system of equations.

matrix  $\mathbf{S}$  with elements  $S_{ij} = \sqrt{2/(n+1)} \sin(\pi i j / (n+1))$ . The eigenvalues are (for  $j = 1, 2, \dots, n$ )

$$\lambda_j = \frac{1}{2} + \frac{1}{2} \cos\left(\frac{\pi j}{n+1}\right) = \cos^2\left(\frac{\pi}{2} \frac{j}{n+1}\right), \quad (6.53)$$

which represents the so-called *transfer function* of the low-pass filter. If compared with the filter function  $\varphi_j$  from (6.38) it appears that no regularisation parameter is necessary. However, the number  $n$  of the components of the solution vector  $\mathbf{x}$  acts like a regularisation parameter.

## 6.5 Iterative Unfolding

Direct matrix methods (i.e. non-iterative methods) like the SVD cannot be used for problems with very large dimension parameters  $m$  and  $n$ . In such cases iterative methods for unfolding are constructed and used, where the usually sparse huge response matrix  $\mathbf{A}$  appears only in products, thus avoiding additional memory space. These iterative methods are characterised by an implicit regularisation and have a so-called *semi-convergence*. Starting from some initial assumption  $\mathbf{x}^{[0]}$ , the first iterations show substantial improvement, but the convergence becomes then rather slow, and after a very large number of iterations often a solution with large noise components similar to the naive solution is obtained. Essentially the methods have initially a strong regularisation, which gradually disappears during the iterations. The iteration has to stop early if a result is reached with an acceptable  $\chi^2_\gamma$  (discrepancy principle [26]). Because no matrix like the effective regularised inverse  $\mathbf{A}^\#$  is available, no prescription for a direct covariance matrix calculation exists. Estimates of the covariance matrix require for instance Monte Carlo methods.

**Lucy–Richardson deconvolution** An algorithm used for picture deblurring is the iterative Lucy–Richardson deconvolution [31, 32]. Here, the response matrix is assumed to be a *point spread function*, describing for example the spread of light from a single pixel over many pixels in a blurred image. The algorithm has been used for the deconvolution of images and spectra from the Hubble space telescope. The derivation of the formula for the  $k+1$  iteration step,

$$\mathbf{x}_j^{[k+1]} \equiv \frac{\mathbf{x}_j^{[k]}}{\epsilon_j} \sum_{i=1}^m \frac{y_i}{c_i} A_{ij} \quad \text{with} \quad c_i = \sum_{j=1}^n A_{ij} \mathbf{x}_j^{[k]} \quad \text{and} \quad \epsilon_j = \sum_{i=1}^m A_{ij} \quad (6.54)$$

make use of *Bayes' theorem*, which links a conditional probability to its inverse. Empirically the method converges to the maximum-likelihood solution for Poisson-distributed data.

**Landweber iteration** A standard iterative method is the Landweber iteration [33], derived from the least-squares normal equation  $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{y}$ , with a formula for the  $k + 1$  iteration step,

$$\mathbf{x}^{[k+1]} \equiv \mathbf{x}^{[k]} + \omega \mathbf{A}^T (\mathbf{y} - \mathbf{A} \mathbf{x}^{[k]}) \quad 0 < \omega < \frac{2}{\sigma_1^2} \quad k = 0, 1, 2, \dots , \quad (6.55)$$

where a relaxation factor  $\omega$  is introduced to obtain a convergent behaviour ( $\sigma_1$  is the largest singular value of  $\mathbf{A}$ ). The semi-convergence corresponds to an implicit regularisation with approximate filter factors [1, 2]

$$\varphi_j^{[k]} \approx 1 - \left(1 - \omega \sigma_j^2\right)^k \approx \begin{cases} 1 & \text{for large } \sigma_j^2 \\ k \left(\omega \sigma_j^2\right) & \text{for } \sigma_j^2 \ll 1/\omega \\ \rightarrow 0 & \text{for } \sigma_j^2 \rightarrow 0 \end{cases}, \quad (6.56)$$

and the number of iterations is the regularisation parameter. In practice only few iterations are performed during which the contributions with large singular values show a fast convergence. Contributions with small singular values will have a very slow convergence, and the components from the initial assumption are still present in the solution. An objective criterion for stopping the iteration is not known and not necessary for applications like picture deblurring, where no covariance matrix is required.

**Iterative methods in particle physics** Iterative methods are rather popular in particle physics although typically the number of parameters is rather small and there will be neither CPU-time nor memory-space problems for direct matrix methods. The *bin-by-bin correction factor method* [6, 34] with  $m = n$  and equal binning in  $\mathbf{x}$  and  $\mathbf{y}$ , which has its origin in old methods of particle physics for acceptance corrections, is used if the observed distribution is almost identical to the true distribution.

An attempt is made by iterative tuning of the MC simulation with reweighting, to perform the simulation already with the *correct* input distribution  $f(t)^{\text{model}}$ , that is that distribution which on average gives a reasonable description of the observed distributions  $\mathbf{y}$  (for example: ‘... is reweighted until the Monte Carlo samples accurately follow each of the measured  $p_{T,\text{cal}}^{\text{jet}}$  distributions’ [28]). In principle the unfolding is already accomplished by such tuning, but it is typically done in preparation for a following simple bin-by-bin unfolding  $\mathbf{x} = \mathbf{M}_x \mathbf{y}$  with a *diagonal unfolding matrix*  $\mathbf{M}_x$ , with (positive) elements  $(\mathbf{M}_x)_{ii} = x_i^{\text{MC}} / y_i^{\text{mc}}$ , determined from the tuned MC simulation with an MC model  $\mathbf{x}^{\text{model}}$  equal to the expected result as input. The correction factors for migration effects are thus based on the fixed MC input distribution. The probability for entries to fall into the same  $\mathbf{x}$  and  $\mathbf{y}$  bin is required to be large (high purity), and thus any non-linear distortion effect in the measurement process has to be corrected before.

For real unfolding in case of a truly *linear* Fredholm integral equation this tuning should not be necessary, but it may be advantageous for complex physical processes, where some non-linearity in the relation between the true distribution and the

measurement process may exist. In other iterative methods used in particle physics the unfolding matrix  $\mathbf{M}_x$  is iteratively improved and applied to the data  $\mathbf{y}$  to give an improved estimate  $\mathbf{x}^{[k+1]} = \mathbf{M}_x^{[k]} \mathbf{y}$ . A method [35] identical to Lucy–Richardson convolution (6.54), where the response matrix is assumed to be determined by a point spread function (PSF), is used in particle physics, but applied to problems with a general response matrix. The matrix  $\mathbf{M}_x$  depends on the estimate  $\mathbf{x}^{[k]}$  and does not have the properties of a regularised effective inverse  $\mathbf{A}^\#$ . Convergence of the method corresponds to the fixed-point relation  $\mathbf{x} = \mathbf{M}_x \mathbf{y} = \mathbf{M}_x \mathbf{A} \mathbf{x}$ , but the resolution matrix  $\mathbf{\Xi} = \mathbf{M}_x \mathbf{A}$  is not equal to the unit matrix  $\mathbf{I}$  and the ‘fixed-point relation’ is valid only for the used vector  $\mathbf{x}$ . Thus, there will be a non-zero systematic unfolding error  $\Delta \mathbf{x}_{\text{sys}} = (\mathbf{\Xi} - \mathbf{I})\mathbf{x}$  (see (6.35)) for any  $\mathbf{x}$  except the MC input distribution  $\mathbf{x}_{\text{model}}$ .

The unfolding matrix  $\mathbf{M}_x$  in iterative methods has only positive elements, in the case of the bin-by-bin method even only diagonal elements. Generally the methods allow a reasonable solution to be obtained with an acceptable  $\chi^2_y$ . However, often large but unknown positive correlations between the data points are present, equivalent to a strong smoothing, thus ‘...yielding unrealistically optimistic results’ [3]. Quantitative statements about bin correlations, for example by the calculation of the covariance matrix from standard error propagation, are not possible because of the properties of matrix  $\mathbf{M}_x$  (dependent on solution  $\mathbf{x}$ , with all elements positive) mentioned above.

## 6.6 Unfolding Problems in Particle Physics

### 6.6.1 Particle Physics Experiments

In particle physics, collisions of high-energy particles – *events* – are recorded and reconstructed. In the physics analysis of a large number of events, each of which is described by a large number of variables, often the distribution of a single variable or of a pair of variables is considered. The raw measured binned distribution is represented by the vector  $\mathbf{y}$ , and the uncertainties are described by the covariance matrix  $\mathbf{V}_y$ , assumed to be diagonal if only statistical fluctuations are considered. No direct interpretation or comparison with other measured data or predictions is possible if the raw measured distribution is affected by the effects mentioned in the introduction Section 6.1.1 like finite resolution and limited acceptance. In a real experiment several sources of systematic uncertainties are determined and various corrections have to be applied.

Often data corrections like background subtraction, acceptance corrections or corrections for non-linearities in some detector components are applied in several analysis steps to the raw measured distribution before unfolding, and make the consistent propagation of uncertainties more difficult, if not impossible. There is

some advantage to describing the measurement process taking into account the known systematic effects as far as meaningful within the MC simulation. The simulated data allow the response matrix to be determined, which can, in a *single step*, either be used to predict, by *folding*, the measured distribution of some model, or to perform the complex procedure of *unfolding* the raw data. Such a single-step procedure including all corrections allows the statistical properties of the data to be taken into account and results in a better control of the uncertainties. Both folding and unfolding suffer from the limited sensitivity of the measurement for contributions related to small singular values, as described in Section 6.2.1 by (6.25), and in Section 6.2.3.

**Folding** In the following specific case, the measured data distribution  $y$  can be directly interpreted using the *folding* of a model:

- Comparison with a theoretical prediction or model  $f(t)^{\text{model}}$ :

Assuming that there are *no free* parameters in the model, one can use *folding* instead of unfolding: if the prediction of the model is given by the vector  $x^{\text{model}}$ , the expected *measured* distribution can be written in the form  $y^{\text{model}} = Ax^{\text{model}}$ . For a quantitative statistical comparison between prediction and measurement one can calculate, *without* unfolding, the expression

$$\chi_y^2 = (y^{\text{model}} - y)^T V_y^{-1} (y^{\text{model}} - y) \quad (6.57)$$

and the *p*-value for  $\chi_y^2$  with  $m$  degrees of freedom in the case of  $m$  bins. The result of the comparison is either ‘compatible’, if the *p*-value is acceptable, or ‘incompatible’ otherwise.

**Unfolding** Two other cases are considered which require unfolding, that is the reconstruction of the distribution  $\hat{x}$  with the covariance matrix  $V_x$ :

- Determination of parameter values  $a$  by a fit of a theoretical parameterisation  $f(t; a)^{\text{model}}$ :

The theoretical distribution that is represented by  $x^{\text{model}}(a)$  is fitted to the reconstructed distribution  $\hat{x}$  in the minimisation with respect to the parameters  $a$  with a  $\chi^2$  expression

$$F(a) \equiv \chi_x^2 = [x^{\text{model}}(a) - \hat{x}]^T V_x^{-1} [x^{\text{model}}(a) - \hat{x}] . \quad (6.58)$$

This fit is only possible for non-singular covariance matrices  $V_x$  (see discussion in Section 6.6.4), thus requiring the use of a correspondingly small number  $n$  of bins for  $\hat{x}$ . Unfolding has the advantage that theoretical predictions developed in the *future* can be checked with the reconstructed distribution. The alternative and *preferred* method for *fixed known models* is the *parameterised unfolding* derived in Section 6.1.5, which avoids the potential problem of a singular covariance matrix.

- Comparison with other experiments:

Two experiments are considered with reconstructed distributions  $\hat{x}_1$  with  $V_{x,1}$ , and  $\hat{x}_2$  with  $V_{x,2}$ . The  $\chi^2$  expression for a quantitative comparison,

$$\chi_x^2 = (\hat{x}_1 - \hat{x}_2)^T (V_{x,1} + V_{x,2})^{-1} (\hat{x}_1 - \hat{x}_2) , \quad (6.59)$$

can only be calculated for non-singular covariance matrices.

In real experiments there may be several sources of background contributions and of systematic uncertainties. Background contributions may have statistical and scale uncertainties. Some of these uncertainties may have a significant influence on the unfolding itself. The effect of these uncertainties has to be studied and requires a detailed understanding of the measurement procedure and the physics process. The studies often require a repeated determination of the response matrix and the unfolding (or folding) under different conditions, for example with a change of a certain quantity by  $\pm 1$  standard deviations of a systematic uncertainty, giving reconstructed distributions  $\hat{x}_+$  and  $\hat{x}_-$ . From the difference between  $\hat{x}_+$  and  $\hat{x}_-$  the systematic uncertainty of the result can be estimated.

**Response matrix calculation and check** Unfolding requires an accurate calculation of the response matrix, typically in a Monte Carlo simulation. In order to avoid additional statistical uncertainties, simulations with large statistics should be used. The statistical uncertainties of the elements of the response matrix can be controlled by the tolerance  $\delta$ , defined in (6.29). It is not important to fine-tune the model  $f(t)^{\text{model}}$  used in the MC simulation, as the unfolding algorithm will make the fine adjustment.

More important and often difficult is the test of the correctness of the response matrix  $A$ . Any mistake in the response matrix will create certain systematic effects in the reconstructed distribution. The influence of a deviation between the actual measurement and the simulated measurement is in general difficult to estimate.

One example of a check of the simulation is the following. The distribution of the squared momentum transfer  $Q^2$  is unfolded using the measured distribution  $y$  of  $Q^2$  directly in the unfolding. The distribution  $\hat{y}$  of  $Q^2$ , calculated from the unfolded distribution  $\hat{x}$  of  $Q^2$  by the product  $\hat{y} = A\hat{x}$ , will agree well with the measured distribution  $y$ , because the difference is minimised in the unfolding fit. A mistake in the response matrix would not become visible. The squared momentum transfer  $Q^2$  is calculated from measured polar angles  $\vartheta$  and measured energies  $E$ . The quality of the simulation can be checked by a comparison of the directly measured distributions of  $\vartheta$  and  $E$  with the corresponding distributions from the simulation, *reweighted* according to the result of the unfolding. A mistake, for example in the alignment of the detector measuring the polar angle  $\vartheta$  or in the calibration of the energy  $E$ , should become visible in the comparison. The program `RUN` (Section 6.7) supports such checks of the simulation. The checks require the unfolding result to be represented internally by a *continuous* function (for the determination of the individual reweighting factor) and working with ntuples containing the necessary data. In the example above the ntuples have to contain  $Q^2$ ,  $\vartheta$  and  $E$ . The program

RUN also allows one to alternatively use the measured two-dimensional distribution of  $(\vartheta, E)$  in the unfolding of the squared momentum transfer  $Q^2$ .

The check of the correctness of the response matrix, that is the quality of the reconstruction and the simulation, may be the most time-consuming part of the analysis, especially if mistakes are detected and have to be corrected.

Some regions of the measured variable may be inaccessible to a measurement, for example very low transverse momenta in an experiment at high energy. The distribution cannot be reconstructed in this region, but by migration the region influences the measurement region. In this case the model  $f(t)^{\text{model}}$  should give a reasonable description of the inaccessible region in order to provide a good description of the migration.

### 6.6.2

#### **Unfolding Smooth Distributions**

In many unfolding problems there is some a priori knowledge about the general properties of the true distribution. Often the distributions are expected to be rather smooth; in those cases a regularisation based on second derivatives is justified. Two cases can be distinguished.

- *Migration problems:* Many unfolding problems in particle physics can be classified as *migration* problems. The measured values of variables differ from the true values by migration effects, without a large non-linearity in the response; the measured distribution is essentially a smoothed version of the true distribution. In addition the measured distribution is affected by a limited acceptance probability which has to be corrected for. In practice the method of bin-by-bin correction factors [34], originally developed for pure acceptance corrections, is usually applied in those simple problems with  $n = m$ .
- *Transformation problems:* Certain physical variables are inaccessible to a direct measurement, but unfolding methods with variable transformations allow these variables to be accessed. One example of a transformation problem is the determination of the differential cross-section  $d\sigma/dy$  of the scaling variable  $y$ , defined by  $y = E_{\text{hadron}}/E_{\nu,\text{in}}$ , in a neutral-current neutrino experiment [36], where an event-by-event reconstruction of the scaling variable is impossible: only the energy  $E_{\text{hadron}}$  of the hadronic shower and the radius  $r$  of the interaction with respect to the beam axis can be measured. In a MC simulation the two-dimensional distribution  $(E_{\text{hadron}}, r)$  can be determined for a given cross-section  $d\sigma/dy$ , using the known neutrino energy spectrum as a function of the radius  $r$ . Unfolding leads from the distribution  $(E_{\text{hadron}}, r)$  to the cross-section  $d\sigma/dy$ . Another example of a transformation problem was given at the end of Section 6.6.1, where the reconstruction of the distribution of the squared momentum transfer  $Q^2$  from the measured two-dimensional distribution of  $(\vartheta, E)$  was discussed.

In transformation problems it may be necessary to include additional measured variables even for a one-dimensional true variable, if these variables are corre-

lated. This is essential, as in the example of the neutrino experiment above, but will help to reduce uncertainties of the reconstructed result in other cases by the use of additionally measured information. Simple methods like the method of bin-by-bin correction factors [34] are of course useless in those cases.

For smooth and structureless distributions treated with orthogonalisation methods and second-derivative regularisation the decrease of the Fourier coefficients  $c_j$  with increasing index  $j$  is fast and the few significant coefficients are well determined. The regularisation parameter can be determined either from the spectrum of the coefficients  $c_j$  or from the  $L$ -curve. The simpler norm regularisation should be used only if the relative accuracy of the data is low; the introduction of some a priori assumption  $x_0$  on the result according to (6.40) can avoid a bias in norm and second-derivative regularisation.

Continuous distributions with *large variations* of the curvature (or second derivative), where the decrease of singular values 6.2.2 with increasing index is slow, can often be converted to smoother distributions with reduced curvature variation and faster decrease of the singular values by variable transformation, improving the unfolding procedure. A transformation can be applied to the variable of the measured and/or of the true distribution. The case of steeply falling energy and momentum distributions with a square-root transformation was discussed in an example. A transformation can improve unfolding if the transformed variable with constant bin width has a more uniform resolution and statistic. The back-transformed reconstructed distribution will have a variable bin width, well adapted to the resolution and statistic.

The numbers of bins  $m$  and  $n$  for the measured and unfolded distributions have to be chosen by the user. The number  $m$  should be large enough to avoid the deterioration of the resolution by broad bins. The number  $n$  should be selected in view of the often small effective number of degrees of freedom (see discussion in Section 6.6.4). For the regularisation methods  $n < m$  is required and in general  $n \approx m/2$  is recommended.

In case of a distribution with narrow structures (peaks and valleys) the decrease of the Fourier coefficients  $c_j$  with increasing index  $j$  will be slower; a larger number of Fourier coefficients would be needed to avoid a bias in the reconstruction of significant structures. In case of very narrow peaks a special parameterised unfolding with an explicit parameterisation of the peaks should be preferred, using parameterised unfolding (see Section 6.1.5).

### 6.6.3

#### **Unfolding Non-smooth Distributions**

In some unfolding problems of particle physics observed events have to be assigned to different event classes, where some ‘mixing’ occurs, that is some mis-assignment because of limited detector resolution. An example from particle physics is the classification of events into *charged-current* and *neutral-current* events, or the assignment of events to classes of ‘ $N$ ’-jet events. In these cases, the

variables  $s$  and  $t$  of the direct and the inverse process are integers. Regularisation based on derivatives can not be used in case of very few bins (classes) or is meaningless in class assignment problems with a given fixed number of classes. The norm regularisation is still possible and is recommended. Often no standard unfolding program can be used because of special experimental conditions, for example difficult background contributions. Specialised code may be necessary, eventually using a standard fit program (e.g. MINUIT [10]). The exercise provided in Section 6.8 given below can even be solved with a pencil alone.

Another example worth studying in detail is the multiplicity distribution of charged particles, where the acceptance probability and resolution of each charged particle depends on the momentum distribution; unfolding the multiplicity distribution using a sample of MC events for the response function therefore requires one to check at least the correct simulation of the momentum distribution.

#### 6.6.4

##### Presentation of Regularisation Results

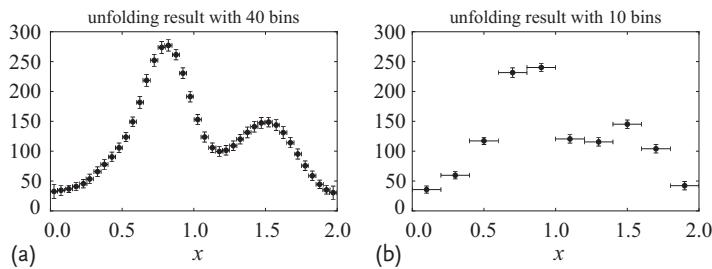
The result of regularised unfolding is an  $n$ -vector  $\hat{\mathbf{x}}$ , representing the ‘true’ function  $f(t)$ , together with an  $n$ -by- $n$  covariance matrix  $\mathbf{V}_x$ , where the number  $n$  is user-defined. Because of the limited resolution leading to migration between bins there is an unavoidable correlation between the bin contents of the reconstructed distribution  $\hat{\mathbf{x}}$ , which is quantified in the off-diagonal elements of the covariance matrix  $\mathbf{V}_x$ . The selection of an adequate value of the regularisation parameter  $\tau$  in regularisation methods (Section 6.3) allows one to find a solution with reduced correlations.

For the presentation of the result usually a larger value of  $n$  is preferred although correlations increase with  $n$ , and often a value  $n$  much larger than the effective number of degrees of freedom of the measurement is selected. The effective number of degrees of freedom can be estimated with different methods. One method is by diagonalisation of  $\mathbf{V}_x$ , counting the number  $k$  of (significantly) non-zero eigenvalues, resulting in an effective rank  $k$  of the covariance matrix. Another method is based on an estimate of the effective number of degrees of freedom from the sum of the filter factors:

$$n_{\text{df}} \approx \sum_{j=1}^n \varphi_j = \sum_{j=1}^n \frac{\sigma_j^2}{\sigma_j^2 + \tau} \equiv \sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau}. \quad (6.60)$$

This value has been used in the unfolding program `RUN` [15, 16]. A method to avoid the singular-matrix problem is to present the unfolding result with a number of data points  $n$  equal or not much larger than  $n_{\text{df}}$ , such that the effective rank  $k$  of the matrix is close to  $n$ .

If a larger value  $n$  is used, calculating  $n$  bin contents from an effective rank  $k$  with  $k < n$ , there will be large positive correlations between the elements of vector  $\hat{\mathbf{x}}$ , which give the plot of the unfolded data a misleadingly *smooth* appearance. The plot will be difficult to interpret and impossible to compare quantitatively with predictions, because the covariance matrix  $\mathbf{V}_x$  is singular with rank defect  $\approx n - k$ . A



**Figure 6.14** Plots of the same unfolding result with (a) 40 correlated reconstructed data points and with (b) 10 data points. The data points in (a) were taken from [24].

fit of a parameterisation is of course possible with the original data, as described in Section 6.1.5, but not with the unfolded data.

Figure 6.14 illustrates the problem of selecting a reasonable  $n$  of bins. In Figure 6.14a, where the data points with approximately  $n_{\text{df}} \approx 10$  were taken from [24], the unfolded data are shown in 40 bins. Figure 6.14b shows the same result with ten unfolded bins. Both results, together with their covariance matrices, are statistically correct and equivalent. The plot with 40 data points will give the impression of a more accurate result, but the comparison with the full-rank data on the right shows the true information content. Note that combining four data points into one will not reduce the error by a factor 1/2, but less, if the data are positively correlated. In particle physics there seems to be no general agreement on the presentation of results from unfolding.

## 6.7

### Programs Used for Unfolding in High Energy Physics

Several general programs, based on regularisation methods, are used in particle physics; they are described below. An overview is also given in [37]. Several programs and applications were discussed in the workshop on unfolding of the Physstat2011 conference [38]. In general the matrix  $\mathbf{L}$  from second derivatives is used for regularisation, but other choices are possible.

- |    |  |
|----|--|
| LR | (1972) The iterative Lucy–Richardson algorithm is based on the maximum-likelihood method for Poisson-distributed input data, and can be applied to a large number of input data (e.g. picture deblurring, two-dimensional); no covariance matrix is calculated. Regularisation is obtained by limiting the number of iterations as mentioned in Section 6.5. The response function is a PSF: it describes the migration to neighbouring data points. A program called BAYES UNFOLDING (1995) [35] is available for use in particle physics; it is based on the same iterative algorithm as the Lucy–Richardson method. |
|----|--|

RUN	(1984) [15, 16] The general program, developed originally for neutral-current neutrino experiments, reconstructs one-dimensional distributions from one-dimensional to three-dimensional measured distributions. While all other programs require histogram-like input of the measured distributions and the response matrix, the input to RUN are ntuples for data, MC simulation and background. This allows the use of B-splines in the discretisation and helps to avoid the discontinuities of histograms. In addition unfolding with transformed distributions is easily done. On the basis of the maximum-likelihood method for Poisson-distributed input data it uses second-derivative regularisation with orthogonalisation and a user-defined effective number of significant coefficients. The program includes options to check the response matrix calculation as explained at the end of Section 6.6.1. A C++ version of the code, named TRUEE, has been implemented [39].
GURU	(1996) [24] On the basis of the SVD the least-squares solution is determined using second-derivative regularisation. The SVD allows for a well-founded estimate of the regularisation parameter. The regularisation parameter is a user-defined effective number of significant coefficients, recognised with the SVD.
TUNFOLD	(2010) [40] This flexible program allows the reconstruction of multi-dimensional distributions by the method of least squares; in the case of more than one dimension the bins have to be reordered to fit in one dimension. The user can choose different regularisation matrices, but no orthogonalisation is done. The regularisation parameter is determined by the <i>L</i> -curve method. The program provides methods to do systematic error propagation and to do unfolding with background subtraction.

The programs GURU and TUNFOLD both allow user-defined regularisation parameter values. If applied to a one-dimensional problem with regularisation by second derivatives, the results of both least-squares methods should be identical if the same value is used for the regularisation parameter. The direct (i.e. non-iterative) methods allow the calculation of the full covariance matrix by propagation of the uncertainties. In addition there are iterative methods in use which were already mentioned in Section 6.5; in these methods the covariance matrix has to be estimated using MC methods.

In order to simplify the application of the unfolding programs the framework ROOUNFOLD [41] within the ROOT software package [42] has been developed. It includes, among others, interfaces to the programs GURU and TUNFOLD. The BAYES UNFOLDING program mentioned above is also included, as are simple methods that use direct matrix inversion (Section 6.1.2). Finally, the bin-by-bin correction factor method [34] is available to allow the comparison with the other methods.

## 6.8 Exercise

### Exercise 6.1 An unfolding problem with two measured values

An unfolding problem with two measured and two true values is considered. The true and the measured values are assumed to be

$$\text{true } \mathbf{x} = \begin{pmatrix} 25 \\ 16 \end{pmatrix}, \quad \text{measured } \mathbf{y} = \begin{pmatrix} 20 \\ 20 \end{pmatrix}$$

(the numbers are assumed to be event numbers). The response matrix  $\mathbf{A}$  is given for two detectors of different resolution:

$$\mathbf{A}_a = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix}, \quad \mathbf{A}_b = \begin{pmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{pmatrix}.$$

From the measured data  $\mathbf{y}$  the unfolded data  $\mathbf{x}$  including uncertainties and correlations between the two components have to be determined by matrix or other methods.

## References

- 1 Hansen, P.C. (1997) *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM monographs on mathematical modeling and computation, Society for Industrial and Applied Mathematics.
- 2 Hansen, P.C. (2010) *Discrete Inverse Problems – Insight and Algorithms*, Fundamentals of Algorithms, Society for Industrial and Applied Mathematics.
- 3 Kaipio, J. and Somersalo, E. (2004) Statistical and computational inverse problems, in *Applied Mathematical Science*, vol. 160, Springer.
- 4 Vogel, C.R. (2002) *Computational Methods for Inverse Problems*, SIAM Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics.
- 5 Blobel, V. and Lohrmann, E. (1998) *Statistische und numerische Methoden der Datenanalyse*, Teubner.
- 6 Cowan, G. (1998) *Statistical Data Analysis*, Oxford University Press.
- 7 de Boor, C. (1978) *A Practical Guide to Splines*, Springer.
- 8 Kaipio, J. and Somersalo, E. (2007) Statistical inverse problems: discretization, model reduction and inverse crimes. *J. Comput. Appl. Math.*, **198**, 493.
- 9 Blobel, V. (2002) An unfolding method for high-energy physics experiments. *Proc. Adv. Stat. Techn. Part. Phys.* Durham, arXiv:hep-ex/0208022.
- 10 James, F. and Roos, M. (1975) Minuit – A system for function minimization and analysis of the parameter errors and correlations. *Comput. Phys. Commun.*, **10**, 343.
- 11 Gagunashvili, N. (2011) Parametric fitting of data obtained from detectors with finite resolution and limited acceptance. *Nucl. Instrum. Methods A*, **635**, 86.
- 12 Wilkinson, J.H. and Reinsch, C. (1971) *Handbook for Automatic Computation*, vol. 2, Linear Algebra, Springer.
- 13 Bjork, A. (1996) *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics.
- 14 Golub, G.H. and van Loan, C.F. (1983) *Matrix Computations*, John Hopkins University Press.

- 15** Blobel, V. (1984) Unfolding methods in high energy physics experiments, Proc. 1984 CERN School of Computing, Aiguablava, Spain. CERN-85-09 and DESY 84-118.
- 16** Blobel, V. (1996) The *RUN* manual: regularized unfolding for high-energy physics experiments. OPAL Technical Note TN361.
- 17** Courant, R. and Hilbert, D. (1924) *Methoden der mathematischen Physik I*, Springer.
- 18** Press, W.H. et al. (1992) *Numerical recipes – the art of scientific computing*, Cambridge University Press.
- 19** Jaynes, E. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620.
- 20** Schmeling, M. (1994) The method of reduced cross-entropy – a general approach to unfold probability distributions. *Nucl. Instrum. Methods A*, **340**, 400.
- 21** Phillips, D.L. (1962) A technique for the numerical solution of certain integral equations of the first kind. *J. Assoc. Comput. Mach.*, **9**, 84.
- 22** Tikhonov, A. (1963) On the solution of improperly posed problems and the method of regularization. *Sov. Math.*, **5**, 1035.
- 23** Tikhonov, A. and Arsenin, V. (1977) *Solutions of Ill-Posed Problems*, Wiley.
- 24** Höcker, A. and Kartvelishvili, V. (1996) SVD approach to data unfolding. *Nucl. Instrum. Methods A*, **372**, 469.
- 25** Paige, C. (1985) The general linear model and the generalized singular value decomposition. *Linear Algebra Appl.*, **70**, 269.
- 26** Morozov, V. (1966) On the solution of functional equations by the method of regularization. *Sov. Math. Dokl.* (**7**), 414.
- 27** Lawson, C.L. and Hanson, R.J. (1974) *Solving Least Squares Problems*, Prentice Hall.
- 28** CDF Collab., Abulencia, A. et al. (2007) Measurement of the inclusive jet cross section using the  $k_t$  algorithm in  $p\bar{p}$  collisions at  $\sqrt{s} = 1.96$  TeV with the CDF II detector. *Phys. Rev. D*, **75**, 092006.
- 29** Takiya, C. et al. (2004) Minimum variance regularization in linear inverse problems. *Nucl. Instrum. Methods A*, **523**, 186.
- 30** Takiya, C. et al. (2007) Variances, covariances and artifacts in image deconvolution. *Nucl. Instrum. Methods A*, **580**, 1466.
- 31** Richardson, W.H. (1972) Bayesian-based iterative method for image restoration. *J. Opt. Soc. Am.*, **62**, 55.
- 32** Lucy, L.B. (1974) An iterative technique for the rectification of observed distribution. *Astron. J.*, **79**, 745.
- 33** Landweber, L. (1951) An iteration formula for Fredholm integral equations of the first kind. *Am. J. Math.*, **73**, 615.
- 34** Cowan, G. (2002) A survey of unfolding methods for particle physics. *Conf. Proc., C0203181*, 248.
- 35** D'Agostini, G. (1995) A multidimensional unfolding method based on Bayes' theorem. *Nucl. Instrum. Methods A*, **362**, 487.
- 36** Jonker, M. et al. (1981) Experimental study of differential cross sections  $d\sigma/dy$  in neutral current neutrino and antineutrino interactions. *Phys. Lett. B*, **102**, 62.
- 37** Albert, J. et al. (2007) Unfolding of differential spectra in the MAGIC experiment. *Nucl. Instrum. Methods A*, **583**, 494.
- 38** Lyons, L. and Prosper, H.B. (eds) (2011) *PHYSTAT2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN, Geneva, Switzerland, CERN-2011-006.
- 39** Milke, N., Doert, M., Klepser, S., Mazin, D., Blobel, V., and Rhode, W. (2013) Solving inverse problems with the unfolding program TRUEE: Examples in Astroparticle physics. *Nucl. Instrum. Meth. A*, **697**, 133.
- 40** Schmitt, S. (2010) TUnfold, an algorithm for correcting migration effects in high energy physics. *J. Instr.*, **7**, T10003, [www.desy.de/~sschmitt/tunfold.html](http://www.desy.de/~sschmitt/tunfold.html) (last accessed on 7 March 2013), arXiv: 1205, 6201 [physics.data-an], DESY 12-129.

- 41** Adye, T. (2011) *RooUnfold: ROOT Unfolding Framework*, Rutherford Appleton Laboratory, <http://hepunix.rl.ac.uk/~adye/software/unfold/RooUnfold.html> (last accessed 7 March 2013).
- 42** Brun, R. and Rademakers, F. (1997) ROOT: An object oriented data analysis framework. *Nucl. Instrum. Methods A*, **389**, 81.

## 7

### Constrained Fits

*Benno List*

#### 7.1 Introduction

In physics, several quantities measured in an experiment are often related by equations. As an example, one may assume that a number of particle tracks measured by a detector should come from a common origin, the primary vertex, or that the energies of all hadronic jets found in an  $e^+ e^-$  collision event should add up to the centre-of-mass energy of the accelerator while the sums of the  $x$ ,  $y$  and  $z$  momenta should be zero. The measurements are assumed to be randomly distributed about the true values, and a model predicts relationships that should hold between these true values.

These relationships can be the result of a complex theoretical calculation, or can be quite simple, for instance if one has measured the same quantity several times one expects the true value for all measurements to be the same.

To get the most accurate result out of a set of measurements, one may seek a set of estimates of the true quantities that are as close as possible to the measured values and fulfil a given set of constraints at the same time. This is called a *constrained fit*. One could say that because of the constraints we expect that the different measurements ‘draw each other’ towards the true values and thus improve the overall accuracy.

The determination of unmeasured quantities, such as the energy and direction of a neutrino, is an additional motivation to perform a constrained fit. For instance, the three independent components of a neutrino’s three-momentum at an  $e^+ e^-$  collider could be calculated from any three of the four equations provided by energy and momentum conservation. To determine the three-momentum without any arbitrariness, and as accurately as possible at the same time, a constrained fit can be employed.

To quantify ‘as close as possible’, one needs a *measure of distance* between the estimated parameters and the measurements, and to express the constraints one needs a set of equations.

In the following, we denote the  $M$  measurements by a vector  $\mathbf{t} = (t_1, \dots, t_M)$ . They are assumed to be independently Gaussian distributed about the true values  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$  with errors<sup>1)</sup>  $\delta \mathbf{t} = (\delta t_1, \dots, \delta t_M)$ . The measurements could for example be track parameters, or the energies, polar and azimuthal angles of jets measured by a detector.

A popular choice for quantifying the distance between the measurements  $t_i$  and the estimated parameters  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_M)$  is given by

$$\chi^2 = (\mathbf{t} - \hat{\boldsymbol{\theta}})^T \mathbf{V}_t^{-1} (\mathbf{t} - \hat{\boldsymbol{\theta}}), \quad (7.1)$$

where  $\mathbf{V}_t$  is the *covariance matrix* of the measurements. The quantity  $\chi^2$  is related to the likelihood function  $L$  of a multivariate Gaussian function by  $\chi^2 = -2 \ln L$  (see Chapter 2 or the discussion in this chapter in Section 7.2.1). More generally, any likelihood function can be regarded as a measure of ‘distance’. In the following we will call the function to be minimised simply the *objective function*  $f$ .

If we have  $K$  constraints, they can be expressed by  $K$  equations of the form

$$c_k(\hat{\theta}_1, \dots, \hat{\theta}_M, \hat{\xi}_1, \dots, \hat{\xi}_U) = 0, \quad (7.2)$$

with  $k$  running from 1 to  $K$ . The constraints are functions of the estimates  $\hat{\theta}_1, \dots, \hat{\theta}_M$  of the measured parameters, and additionally of the unmeasured parameters  $\hat{\xi}_1, \dots, \hat{\xi}_U$  (see Table 7.1 for an overview of the notation used in this chapter).

**Table 7.1** Notation used for constrained fit problems.

Symbol	Meaning
$\mathbf{t} = (t_1, \dots, t_M)^T$	Vector of measured values (length $M$ )
$\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)^T$	Vector of true values (length $M$ )
$\boldsymbol{\xi} = (\xi_1, \dots, \xi_U)^T$	Vector of unmeasured quantities (length $U$ )
$\mathbf{x} = (\boldsymbol{\theta}^T, \boldsymbol{\xi}^T)^T$	Measured and unmeasured quantities (length $N = M + U$ )
$\mathbf{c}(\boldsymbol{\theta}, \boldsymbol{\xi}) = (c_1, \dots, c_K)^T$	Vector of constraint functions (length $K$ )
$\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)^T$	Lagrange multipliers (length $K$ )
$\mathbf{X} = (\mathbf{x}^T, \boldsymbol{\lambda}^T)^T$	Vector of all fit quantities and Lagrange multipliers
$\chi^2$	chi-square function
$f(\boldsymbol{\theta})$	Objective function to be minimised
$\mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}) = f + \boldsymbol{\lambda}^T \mathbf{c}$	Lagrange function
$f_x = \partial f / \partial \mathbf{x}$	Derivatives of the objective function
$g_x = \partial \mathcal{L} / \partial \mathbf{x}$	Derivatives of the Lagrangian
$\mathbf{Y} = \partial \mathcal{L} / \partial \boldsymbol{\lambda}$	Vector of derivatives of the Lagrangian
$\mathbf{A}^T = \partial \mathbf{c} / \partial \mathbf{x}$	Jacobian matrix of the constraints
$\mathbf{L} = \partial^2 \mathcal{L} / \partial \mathbf{x} \partial \mathbf{x}^T$	Hessian matrix of the Lagrangian
$\mathbf{p} = \mathbf{x}^{(v+1)} - \mathbf{x}^{(v)}$	Full step in Newton iteration
$\phi(\mathbf{x}; \mu)$	Merit function with scale parameter $\mu$
$\mathbf{x}^*, \boldsymbol{\lambda}^*$	Solution values of $\mathbf{x}$ and $\boldsymbol{\lambda}$

1) The theory of estimators usually assumes that the errors are known. In practice, errors are often estimated from the data themselves.

The mathematical formulation of our problem is: find a set of values  $(\hat{\theta}_1, \dots, \hat{\theta}_M, \hat{\xi}_1, \dots, \hat{\xi}_U)$  that minimise the objective function  $f(\hat{\theta}_1, \dots, \hat{\theta}_M)$  and simultaneously fulfil the  $K$  constraint equations  $c_k(\hat{\theta}_1, \dots, \hat{\theta}_M, \hat{\xi}_1, \dots, \hat{\xi}_U) = 0$ .

### Example 7.1 Particle decay

Consider the case where a particle of mass  $m$  decays into two massless particles, for instance the decay  $\pi^0 \rightarrow \gamma\gamma$  of a  $\pi^0$  meson into two photons with energies  $E_1$  and  $E_2$ . Given the opening angle  $\psi$  between the two photons, which can be measured quite precisely in a detector, the invariant mass  $m_{12}$  of the two photons is given by  $m_{12}^2 = 2E_1 E_2 (1 - \cos \psi)$ . We are interested in the energies  $\hat{E}_1$  and  $\hat{E}_2$  that result in exactly the  $\pi^0$  mass and minimise the  $\chi^2$  given by

$$\chi^2 = (\hat{E}_1 - E_1^{\text{meas}}, \hat{E}_2 - E_2^{\text{meas}})^T \mathbf{V}^{-1} (\hat{E}_1 - E_1^{\text{meas}}, \hat{E}_2 - E_2^{\text{meas}})$$

with the covariance matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where  $\sigma_1$ ,  $\sigma_2$  and  $\rho$  are the uncertainties and correlation coefficient of the measured energies  $E_1^{\text{meas}}$ ,  $E_2^{\text{meas}}$ , respectively. The inverse of the covariance matrix  $\mathbf{V}$  is given by

$$\mathbf{V}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{pmatrix},$$

and the constraint is expressed by the constraint function

$$c(\hat{E}_1, \hat{E}_2) = 2\hat{E}_1 \hat{E}_2 (1 - \cos \psi) - m_{\pi^0}^2 \quad (7.3)$$

with the nominal  $\pi^0$  mass  $m_{\pi^0}$ .

A general solution to problems of this kind is provided by the method of *Lagrange multipliers*, discussed in Section 7.3. If the objective function and the constraints are not too complicated, an analytic solution of the problem is possible, which is discussed in Section 7.4. In more complicated cases, an iterative solution is necessary. Section 7.5 gives an introduction to the problems and possible solutions that are encountered in that situation; this section can be omitted at first reading.

However, it is often possible to use the constraint equations to eliminate parameters of the problem and arrive at a reduced set of parameters that fulfil the constraints. This method of elimination transforms the difficult problem of minimisation with constraints into a ‘simple’ minimisation problem for which very powerful and well-tested algorithms exist. We will discuss this approach in the next section.

## 7.2

### Solution by Elimination

Consider a case where we have  $M$  measurements collected in a vector  $\mathbf{t}$ , and  $K$  constraints  $c_k(\hat{\boldsymbol{\theta}}) = 0$  that shall be fulfilled by the best estimates  $\hat{\boldsymbol{\theta}}$  of the true values  $\boldsymbol{\theta}$ . If the  $K$  constraints are independent, there should be a set of  $E = M - K$  parameters  $\boldsymbol{\eta}$  and a mapping  $\boldsymbol{\eta} \rightarrow \boldsymbol{\theta}$  such that all constraints are fulfilled, regardless of the value of  $\boldsymbol{\eta}$ . It may be possible to choose all or some of the parameters  $\eta_e$  ( $e = 1, \dots, M$ ) such that they are identical to some of the  $\theta_m$  ( $m = 1, \dots, M$ ) parameters, but this is not necessary.

If it is possible to find such a set of  $E$  parameters  $\boldsymbol{\eta}$ , our minimisation problem may be substantially simplified: instead of searching for a minimum in an  $M$ -dimensional space, subject to  $K$  constraints, we are left with an unconstrained minimisation problem in  $M - K$  dimensions.

There are caveats to this approach, though: finding a parameterisation  $\boldsymbol{\eta}$  that fits a particular set of constraints can be difficult, and the resulting mapping  $\boldsymbol{\eta} \rightarrow \boldsymbol{\theta}$  may be highly non-linear, which could lead to convergence problems in the minimisation process.

Since generally the minimisation in  $\boldsymbol{\eta}$ -space will be performed by some iterative algorithm, another problem arises: to find good starting values for the iteration. In essence, a good set of starting values  $\boldsymbol{\eta}_{\text{start}}$  would be a set of values that gives mapped values  $\boldsymbol{\theta}_{\text{start}}$  ‘close’ to the measured values, where ‘close’ means that the objective function  $f$  is ‘small’ (though not minimal, because that would already be the solution to our problem). In the worst case, the problem of starting values brings us almost back to square one.

However, in many cases it is indeed possible to find a parameterisation where the constraints can be fulfilled easily and where good starting values or even the real solution can be found, and then the elimination approach will probably be much faster than a solution with Lagrange multipliers. One example of such a situation is the problem of constraining many helix tracks measured in a tracking chamber to a common vertex [1].

#### Example 7.1 Particle decay (continued)

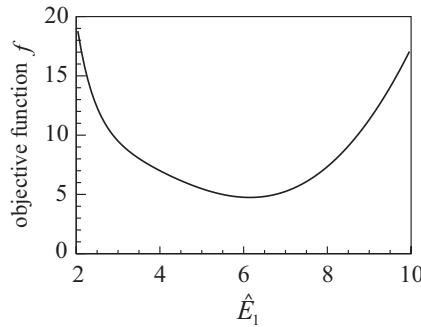
Coming back to the example, we observe that the constraint (7.3) can be used to eliminate the constraint from the problem by setting

$$\hat{E}_2 = \frac{m_{\pi^0}^2}{2(1 - \cos \psi)\hat{E}_1} .$$

In that case,  $\chi^2$  is a function of a single variable  $\hat{E}_1$  only, and we have reduced a constrained minimisation problem with two unknowns into a relatively simple minimisation problem of a single variable. Figure 7.1 shows the resulting  $\chi^2$  for an artificial example where we have set the measured values to be  $E_1^{\text{meas}} = 6$ ,  $E_2^{\text{meas}} = 7$ ,  $\sigma_1 = 1$ ,  $\sigma_2 = 1.5$ , and  $\rho = -0.5$ , and we use  $m_{\pi^0}^2/(2(1 - \cos \psi)) = 25$ .

To keep the notation of the example simple, we ignore units and treat the energies as dimensionless numbers. We will use this parameter set throughout this chapter.

Numerically, the solution  $E_1^*$  is given by  $E_1^* = 6.148\ 28$ , which corresponds to  $E_2^* = 4.066\ 18$  and  $\chi^2 = 4.743\ 25$ .



**Figure 7.1** Illustration of Example 7.1 discussed in the text: value of the objective function  $f = \chi^2$  as a function of  $\hat{E}_1$  after elimination of the second variable  $\hat{E}_2$ .

### 7.2.1

#### Statistical Interpretation

We have developed above the following picture for a constrained fit problem: we assume there exist  $M$  measurements  $\mathbf{t} = (t_1, \dots, t_M)$  that are random variables, distributed about  $M$  true values  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$  with a known probability density function (pdf), which we assume to be Gaussian, with a covariance matrix  $\mathbf{V}_t$ . The true values are subject to  $K$  constraints, expressed by equations  $c_k(\boldsymbol{\theta}) = 0$  for  $k = 1, \dots, K$ .

By solving the constraint equations, we can obtain  $E = N - K$  parameters  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_E)$  and a mapping  $\boldsymbol{\eta} \rightarrow \boldsymbol{\theta}(\boldsymbol{\eta})$  such that each value of  $\boldsymbol{\eta}$  leads to a vector  $\boldsymbol{\theta}$  that fulfills all constraint equations.

Finding the best estimate  $\hat{\boldsymbol{\theta}}$  for the parameters is performed by minimising the  $\chi^2$  given by (7.1). If the measured values are indeed normally distributed about the true values  $\boldsymbol{\theta}$  with the covariance matrix  $\mathbf{V}_t$ , then this  $\chi^2$  is linked to the likelihood by

$$\chi^2 = -2 \ln L(\mathbf{t}|\hat{\boldsymbol{\theta}}), \quad (7.4)$$

or, vice versa,

$$L(\mathbf{t}; \hat{\boldsymbol{\theta}}) = e^{-\frac{\chi^2}{2}}. \quad (7.5)$$

Therefore, minimising this  $\chi^2$  is equivalent to maximising the likelihood (under the assumption of Gaussian errors). This is a purely frequentist approach: no assumption is made about the distribution of the *true* parameters  $\boldsymbol{\theta}$ , only about the distribution of the *measured* values  $\mathbf{t}$  about the true parameters.

The theory of maximum-likelihood estimators, proposed by R.A. Fisher [2, 3], tells us that this is indeed a good choice.

Since we maximise a likelihood, we may change the parameterisation from a set  $\boldsymbol{\theta}$  to a different set of parameters  $\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$  without introducing a Jacobian determinant (see [4] or [5] for a derivation of the transformation properties of pdfs, and [6] for a discussion of the Jacobian). If we choose a parameter set where  $t$  follows a multivariate Gaussian distribution, the minimal  $\chi^2$  of the constrained fit is expected to be distributed according to a  $\chi^2$  distribution with  $K$  degrees of freedom.<sup>2)</sup> Therefore, a  $\chi^2$  test can be used to evaluate the goodness-of-fit, which is very important (cf. Section 3.8).

### 7.3

#### The Method of Lagrange Multipliers

From here on, we will be concerned with the general problem of how to solve a minimisation problem in the presence of equality constraints. Although this problem arises in a statistical context, the mathematical approach to this kind of question is of a more general nature and independent of whether we aim to minimise a  $\chi^2$ , a likelihood function or any other function. Therefore, we will in the following denote the objective function that is to be minimised by the symbol  $f$  rather than the statistically motivated  $\chi^2$ . The theory outlined in the following may be extended by introducing inequality constraints of the form  $c(\boldsymbol{\theta}) \leq 0$  in addition to the equality constraints of the form  $c(\boldsymbol{\theta}) = 0$ ; we refer the reader to the more advanced literature for this problem.

The method of eliminating variables by the application of constraints that was developed in the preceding section has the advantage of transforming a minimisation problem with  $N$  parameters and  $K$  constraints into an unconstrained minimisation problem in a space with  $N - K$  dimensions. The disadvantages are that finding the  $N - K$  parameters that solve the constraints in general requires some algebraic computations that are hard to automatise, and that the choice of good starting values for the minimisation can be non-trivial.

The method of Lagrange multipliers, invented by Joseph Louis Lagrange in 1788 [7] based on earlier work by Leonhard Euler (for a history, see [8]), provides a way to transform the constrained minimisation problem into one where a stationary point of a single function  $\mathcal{L}$ , the *Lagrange function*, is to be found.

Although Lagrange's method is (deceptively) easy to formulate, it also has its disadvantages: first, it introduces a new parameter, the Lagrange multiplier, for each constraint, and thus *increases* the number of parameters by  $K$ , instead of *reducing* it. Second, the solution of the problem is no longer given by the *minimum* of a scalar function  $f$ , but rather by a *stationary point* of the new function  $\mathcal{L}$ ; thus, the first derivatives are zero at the desired solution, but the second derivatives are positive in all directions.

2) This is strictly true only for linear constraints!

## 7.3.1

**Lagrange Multipliers**

In the following, we sketch the proof for the method of Lagrange multipliers. Unfortunately, this entails the use of some concepts that seem daunting at the beginning.

**Example 7.2 Method of Lagrange multipliers in three dimensions**

We will therefore try to give a less precise but hopefully more accessible description in parallel by assuming that we are faced with a minimisation problem with three variables and one constraint.

Consider the problem to find the minimum of an objective function  $f(\mathbf{x})$  in  $N$ -dimensional space (i.e.  $\mathbf{x} \in \mathbb{R}^N$ ), subject to  $K$  constraints given in the form  $c_k(\mathbf{x}) = 0$  with  $k = 1, \dots, K$ . We assume that the objective function  $f$  and the constraint functions  $c_k$  are continuously differentiable in the neighbourhood of the solution  $\mathbf{x}^*$ .

We call the points  $\mathbf{x}$  that satisfy all constraints *feasible points*, and the set  $\Omega = \{\mathbf{x} | c_k(\mathbf{x}) = 0\}$  the *feasible set*. The desired solution  $\mathbf{x}^*$  clearly belongs to the feasible set and is characterised by  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  for all feasible points  $\mathbf{x} \in \Omega$ .

**Example 7.2 Method of Lagrange multipliers in three dimensions (continued)**

In our three-dimensional example with a single constraint, the feasible set is just a two-dimensional surface, and all points on that surface fulfil the single constraint equation  $c(\mathbf{x}) = 0$ . The solution  $\mathbf{x}^*$  must lie on that surface, and the objective function  $f$  is nowhere smaller on the surface than at  $\mathbf{x}^*$ .

Next, we define the *feasible direction*  $\mathbf{d}$ , which is a vector of unit length  $\|\mathbf{d}\| = 1$  that points into a direction where all constraints remain zero. If we expand the constraints into a Taylor series around  $\mathbf{x}^*$  we get

$$c_k(\mathbf{x}^* + \epsilon \mathbf{d}) = c_k(\mathbf{x}^*) + \epsilon \mathbf{d}^T \nabla c_k + \mathcal{O}(\epsilon^2) = 0 + \epsilon \mathbf{d}^T \nabla c_k + \mathcal{O}(\epsilon^2), \quad (7.6)$$

which shows that all feasible directions must satisfy

$$0 = \mathbf{d}^T \nabla c_k, \quad k = 1, \dots, K, \quad (7.7)$$

that is all feasible directions must be perpendicular to the gradients of all constraints. Conversely, all feasible points  $\mathbf{x} \in \Omega$  in the close vicinity of  $\mathbf{x}^*$  may be approximated as  $\mathbf{x} = \mathbf{x}^* + \epsilon \mathbf{d}$ .

**Example 7.2 Method of Lagrange multipliers in three dimensions (continued)**

In our three-dimensional example, the gradient  $\nabla c$  at  $\mathbf{x}^*$  is the direction in which the constraint grows fastest. If we move away from  $\mathbf{x}^*$  without getting off the surface where the constraint  $c = 0$  holds, the direction in which we moved is called a feasible direction.

So, for  $\mathbf{x}^*$  to be a local minimum of the objective function  $f(\mathbf{x})$  in the feasible set  $\mathcal{Q}$ , we must have

$$f(\mathbf{x}^*) \leq f(\mathbf{x} + \epsilon \mathbf{d}) = (\mathbf{x}^*) + \epsilon \mathbf{d}^\top \nabla f(\mathbf{x}^*) + \mathcal{O}(\epsilon^2), \quad (7.8)$$

and therefore

$$0 = \mathbf{d}^\top \nabla f(\mathbf{x}^*) \quad (7.9)$$

must hold for any feasible direction  $\mathbf{d}$ , that is  $\nabla f(\mathbf{x}^*)$  must be perpendicular to all feasible directions. Thus, in the vicinity of the solution  $\mathbf{x}^*$ , the feasible set (which may be a line, a surface or a hypersurface) is tangential to the space where  $f(\mathbf{x})$  is constant.

#### Example 7.2 Method of Lagrange multipliers in three dimensions (continued)

Going back to the three-dimensional example, this means that  $f(\mathbf{x}^*)$  can only be a local minimum on the constraint surface if the projection of the gradient  $\nabla f$  of the objective function onto all directions  $\mathbf{d}$  in the tangential plane is zero, that is, if  $\nabla f$  is perpendicular to the tangential plane.

It follows that  $\nabla f(\mathbf{x}^*)$  must be a linear combination of the constraint gradients, that is numbers  $\lambda_k$  must exist such that

$$0 = \nabla f(\mathbf{x}^*) + \sum_{k=1}^K \lambda_k \nabla c_k(\mathbf{x}^*). \quad (7.10)$$

The numbers  $\lambda_k$  are called the *Lagrange multipliers*. This is a *first-order necessary condition* for  $\mathbf{x}^*$  to be a local minimum of  $f$ .

#### Example 7.2 Method of Lagrange multipliers in three dimensions (continued)

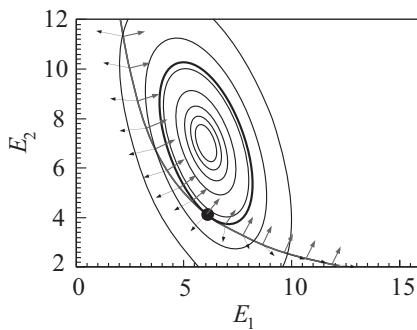
In the three-dimensional case, we see that the gradient  $\nabla f$  of the objective function, which is orthogonal to the tangential plane of the feasible directions, must be parallel to the gradient of the constraint  $\nabla c$  that defines the constraint surface, that is  $\nabla f = -\lambda \nabla c$ . However, if there are more constraints than one, ‘orthogonal to the surface that is orthogonal to all constraint gradients’ no longer means ‘parallel to the constraint gradient’, but ‘a linear combination of all constraint gradients’. This is what (7.10) says.

Figure 7.2 illustrates the method of Lagrange multipliers in the case of the  $\pi^0 \rightarrow \gamma\gamma$  example (two-dimensional case) introduced earlier.

One should note that the Lagrange multipliers are uniquely defined if, and only if, the  $K$  gradient vectors  $\nabla c_k(\mathbf{x}^*)$  are non-zero and linearly independent. This condition is known as *linear independence constraint qualification* (LICQ) [9].

Now we introduce the Lagrange function

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{k=1}^K \lambda_k c_k(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{c}(\mathbf{x}) \quad (7.11)$$



**Figure 7.2** Sketch of the method of Lagrange multipliers: At the solution, which lies on the line where the constraint function is zero, the gradients of the constraint (arrows pointing to the right or up) and of the objective function (arrows pointing left or down) are antiparallel.

and seek a stationary point of  $\mathcal{L}$ , which is characterised by

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_n} &= \frac{\partial f}{\partial x_n} + \sum_{k=1}^K \lambda_k \frac{\partial c_k}{\partial x_n} = 0 \quad n = 1, \dots, N, \\ \frac{\partial \mathcal{L}}{\partial \lambda_k} &= c_k = 0 \quad k = 1, \dots, K.\end{aligned}\tag{7.12}$$

We see that the conditions for a stationary point of  $\mathcal{L}$  are precisely the first-order *necessary conditions* for a minimum of  $f$ .

As usual in the study of minimisation problems, we need to consider the second derivatives in order to obtain *sufficient* conditions in addition to the *necessary* ones. Again we consider that the desired solution  $\mathbf{x}^*$  belongs to the feasible set  $\Omega$  of points for which the constraints are fulfilled, and is characterised by  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  for all feasible points  $\mathbf{x} \in \Omega$ .

Now, since the feasible points are defined by  $\mathbf{c}(\mathbf{x}) = 0$ , it follows that the Lagrangian is equal to  $f$  on the feasible set:  $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x})$  for all  $\mathbf{x} \in \Omega$ , and hence  $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$  for all feasible points  $\mathbf{x} \in \Omega$ .

We make a Taylor expansion<sup>3)</sup> of  $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  in  $\mathbf{x}$  around the solution  $\mathbf{x}^*$ :

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) = \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + (\mathbf{x} - \mathbf{x}^*)^T \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \mathbf{L}^* (\mathbf{x} - \mathbf{x}^*) + \dots\tag{7.13}$$

with the second derivative matrix

$$\mathbf{L}^* = \nabla_{\mathbf{x}\mathbf{x}}^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*).\tag{7.14}$$

Now we observe that for the Lagrangian, in contrast to the objective function  $f$ , the first-order terms all vanish due to the first-order necessary conditions, so that we

3) The subscript ‘x’ in  $\nabla_{\mathbf{x}}$  indicates that derivatives are taken only with respect to  $\mathbf{x}$ , not with respect to  $\boldsymbol{\lambda}$ .

are left with

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) = \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^\top \mathbf{L}^* (\mathbf{x} - \mathbf{x}^*) + \dots \quad (7.15)$$

Since  $\mathbf{x}^*$  must be a local minimum of  $f$ , which means  $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) \geq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ , the *second-order necessary condition* follows

$$\mathbf{d}^\top \mathbf{L}^* \mathbf{d} \geq 0 \quad (7.16)$$

for all feasible directions  $\mathbf{d}$  satisfying (7.7). It turns out that the stricter condition

$$\mathbf{d}^\top \mathbf{L}^* \mathbf{d} > 0 \quad (7.17)$$

is in fact a *sufficient* condition. For a more rigorous derivation of Lagrange multipliers see Section 9.1 of [10].

### Example 7.1 Particle decay (continued)

Returning to the  $\pi^0 \rightarrow \gamma\gamma$  example, we can write out the objective function as

$$f = \chi^2 = (\hat{E}_1 - 6)^2 / 1.0^2 - 2 \cdot (-0.5)(\hat{E}_1 - 6)(\hat{E}_2 - 7) / (1.0 \cdot 1.5) + (\hat{E}_2 - 7)^2 / 1.5^2.$$

We see that at the solution  $(E_1^*, E_2^*) = (6.148\ 28, 4.066\ 18)$ , the gradient of  $f$  is given by  $\nabla f(E_1^*, E_2^*) = (-2.212\ 43, -3.345\ 32)^\top$ , while the gradient of the constraint  $c(E_1, E_2) = E_1 E_2 - 25$  is  $\nabla c(E_1^*, E_2^*) = (4.066\ 18, 6.148\ 28)^\top$ , so with  $\lambda^* = 0.544\ 10$  we have indeed  $\mathbf{0} = \nabla f(E_1^*, E_2^*) + \lambda^* \nabla c(E_1^*, E_2^*)$ .

The feasible set in the example is just given by all points on the hyperbola  $E_1 E_2 - 25 = 0$ , and the feasible direction vectors are the tangential vectors at the solution point  $(E_1^*, E_2^*)$  (see Figure 7.2).

#### 7.3.2

##### Unmeasured Parameters

In physics applications one often encounters the situation that some of the parameters  $x_n$  are not measured, and thus do not contribute to the overall  $\chi^2$ . In the context of the Lagrange method this means that the objective function  $f(\mathbf{x})$  does not depend on these unmeasured parameters:

$$\frac{\partial f}{\partial x_n} = 0. \quad (7.18)$$

In the following we will assume that we have  $M$  measured quantities  $\theta_m$ ,  $m = 1, \dots, M$ , and  $U$  unmeasured quantities  $\xi_u$ ,  $u = 1, \dots, U$ ; we collect these quantities in vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\xi}$  and can combine them in the vector  $\mathbf{x}$  such that the measured quantities come first, followed by the unmeasured quantities:

$$\mathbf{x}^\top = (\boldsymbol{\theta}^\top, \boldsymbol{\xi}^\top). \quad (7.19)$$

The vector  $\mathbf{x}$  has altogether  $N = M + U$  rows.

The minimisation problem will only have a unique solution if the unmeasured parameters are determined by the constraint functions, therefore we need  $K \geq U$  independent constraints in the presence of unmeasured parameters.

### Example 7.3 W decay

Unmeasured quantities inevitably occur in particle collisions with neutrinos in the final state. As an example, consider the leptonic decay of a  $W$ -boson to a positron and a neutrino ( $W^+ \rightarrow e^+ \nu_e$ ) in an event at a hadron collider where the  $W$  is produced together with a number of jets. Assuming that all particles are treated as massless and are parameterised with their transverse energies  $E_i^T$ , pseudo-rapidities  $\eta_i$  and azimuthal angles  $\phi_i$ , two constraints arise from transverse momentum conservation:

$$0 = c_1(\mathbf{x}) = \sum_j E_j^T \cos \phi_j + E_e^T \cos \phi_e + E_\nu^T \cos \phi_\nu , \quad (7.20)$$

$$0 = c_2(\mathbf{x}) = \sum_j E_j^T \sin \phi_j + E_e^T \sin \phi_e + E_\nu^T \sin \phi_\nu . \quad (7.21)$$

A third constraint could come from the assumption that the invariant mass of the positron–neutrino system must be the  $W$  mass,  $M_W$ :

$$\begin{aligned} 0 &= c_3(\mathbf{x}) \\ &= M_W^2 - (E_e^T \cosh \eta_e + E_\nu^T \cosh \eta_\nu)^2 + (E_e^T \cos \phi_e + E_\nu^T \cos \phi_\nu)^2 \\ &\quad + (E_e^T \sin \phi_e + E_\nu^T \sin \phi_\nu)^2 + (E_e^T \sinh \eta_e + E_\nu^T \sinh \eta_\nu)^2 . \end{aligned} \quad (7.22)$$

In this problem, the vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\xi}$  of measured and unmeasured quantities would be given by

$$\boldsymbol{\theta}^T = (E_e^T, \eta_e, \phi_e, E_{j1}^T, \eta_{j1}, \phi_{j1}, \dots) , \quad (7.23)$$

$$\boldsymbol{\xi}^T = (E_\nu^T, \eta_\nu, \phi_\nu) . \quad (7.24)$$

## 7.4

### The Lagrange Multiplier Problem with Linear Constraints and Quadratic Objective Function

Before we consider the Lagrange multiplier problem given by the set of equations (7.12) for arbitrary objective functions and constraints, we begin with the simpler case of linear constraints and a quadratic objective function  $f$ , which has an analytic solution that also serves as a starting point for the solution of the general case.

If the objective function  $f$  is a quadratic function of  $\mathbf{x}$ , its derivative vector  $\partial f / \partial \mathbf{x}$  is a linear function of  $\mathbf{x}$ , and we can write down a Taylor expansion of the constraints around any set of starting values  $\mathbf{x}^{(0)}$ :

$$\mathbf{c}(\mathbf{x}) = \mathbf{c}(\mathbf{x}^{(0)}) + \mathbf{A}(\mathbf{x} - \mathbf{x}^{(0)}) = \mathbf{c}^{(0)} + \mathbf{A}(\mathbf{x} - \mathbf{x}^{(0)}), \quad (7.25)$$

where we have introduced the Jacobian matrix  $\mathbf{A}^T = \partial \mathbf{c} / \partial \mathbf{x}$  of the constraints:

$$A_{nk} = \frac{\partial c_k}{\partial x_n}. \quad (7.26)$$

Since by definition the objective function  $f$  depends only on  $\boldsymbol{\theta}$ , the Taylor expansion involves only the  $\boldsymbol{\theta}$  part of  $\mathbf{x}$ :

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{f}_{\boldsymbol{\theta}}^{(0)} + \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}), \quad (7.27)$$

where  $\mathbf{f}_{\boldsymbol{\theta}}^{(0)} = \partial f / \partial \boldsymbol{\theta}(\boldsymbol{\theta}^{(0)})$  and  $\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \partial^2 f / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)(\boldsymbol{\theta}^{(0)})$ . Using this Taylor expansion, (7.12) can be rewritten as:

$$\begin{pmatrix} -\mathbf{f}_{\boldsymbol{\theta}}^{(0)} \\ \mathbf{0} \\ -\mathbf{c}^{(0)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{0} & \mathbf{A}_{\boldsymbol{\theta}} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{\boldsymbol{\xi}} \\ \mathbf{A}_{\boldsymbol{\theta}}^T & \mathbf{A}_{\boldsymbol{\xi}}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)} \\ \boldsymbol{\xi} - \boldsymbol{\xi}^{(0)} \\ \lambda \end{pmatrix}. \quad (7.28)$$

Assuming that  $\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}$  is indeed non-singular and invertible, this system can be solved by Gaussian block-elimination (after interchanging the second and third rows):

$$\begin{pmatrix} -\mathbf{f}_{\boldsymbol{\theta}}^{(0)} \\ \mathbf{r} \\ -\mathbf{A}_{\boldsymbol{\xi}} \mathbf{S}^{-1} \mathbf{r} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{A}_{\boldsymbol{\theta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} & -\mathbf{A}_{\boldsymbol{\xi}}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{\boldsymbol{\xi}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\xi}}^T \end{pmatrix} \begin{pmatrix} \Delta \boldsymbol{\theta} \\ \lambda \\ \Delta \boldsymbol{\xi} \end{pmatrix}, \quad (7.29)$$

where we use the definitions

$$\mathbf{S} = \mathbf{A}_{\boldsymbol{\theta}}^T (\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} \mathbf{A}_{\boldsymbol{\theta}}, \quad (7.30)$$

$$\mathbf{r} = \mathbf{c}^{(0)} - \mathbf{A}_{\boldsymbol{\theta}}^T (\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} \mathbf{f}_{\boldsymbol{\theta}}^{(0)}. \quad (7.31)$$

Note that  $\mathbf{S}$  is symmetric, and if  $\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}$  is positive definite and  $\mathbf{A}_{\boldsymbol{\theta}}$  has full column rank, also  $\mathbf{S}$  is positive definite;<sup>4)</sup> the same applies to  $\mathbf{A}_{\boldsymbol{\xi}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\xi}}^T$ .

Backward insertion then leads to the final solution:

- solve for  $\Delta \boldsymbol{\xi}$ :  $-\mathbf{A}_{\boldsymbol{\xi}} \mathbf{S}^{-1} \mathbf{r} = \mathbf{A}_{\boldsymbol{\xi}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\xi}}^T \Delta \boldsymbol{\xi}$ ,
- solve for  $\lambda$ :  $\mathbf{r} + \mathbf{A}_{\boldsymbol{\xi}}^T \Delta \boldsymbol{\xi} = \mathbf{S} \lambda$ ,
- solve<sup>5)</sup> for  $\Delta \boldsymbol{\theta}$ :  $-\mathbf{f}_{\boldsymbol{\theta}}^{(0)} - \mathbf{A}_{\boldsymbol{\theta}} \lambda = \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} \Delta \boldsymbol{\theta}$ ,
- calculate  $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)} + \Delta \boldsymbol{\theta}$  and  $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)} + \Delta \boldsymbol{\xi}$ .

4) Therefore, the inverse of  $\mathbf{S}$  may be calculated using Cholesky decomposition [11].

5) Often,  $\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$  is explicitly known, in which case the solution is simply  $\Delta \boldsymbol{\theta} = -\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} (\mathbf{f}_{\boldsymbol{\theta}}^{(0)} + \mathbf{A}_{\boldsymbol{\theta}} \lambda)$ .

If there are no unknowns, the solution simplifies to:

- solve for  $\lambda$ :  $r = S\lambda$ ,
- solve for  $\Delta\theta$ :  $-f_\theta^{(0)} - A_\theta\lambda = F_{\theta\theta}\Delta\theta$ ,
- calculate  $\theta = \theta^{(0)} + \Delta\theta$ .

The advantage of this approach is that it reduces the solution of a single linear system of  $(N + K)$  equations into three smaller sets with  $U$ ,  $K$  and  $M$  equations, respectively, which leads to a significant speed gain as the solution of a symmetric linear set of equations with  $n$  unknowns needs  $n^3/3$  floating point operations (FLOPs).

The definition of  $S$  has one caveat: we have noted that  $S$  is positive definite if  $A_\theta$  has full *column rank*.<sup>6</sup> This is, however, a significantly more stringent assumption than the conventional assumption (LICQ) that the full Jacobian matrix  $A$  has full column rank. If any constraints are present that depend only on unmeasured parameters, then a full column of  $A_\theta$  will vanish, and  $A_\theta$  no longer has full column rank, making  $S$  singular, and  $S^{-1}$  does not exist.

This situation can be changed by observing that at the stationary point, that is the solution to the iteration problem, we have  $0 = A_\xi\lambda$ . Thus, we can decide to seek the solution to  $0 = c^{(0)} + A_\xi^T A_\xi \lambda$  instead of only  $0 = c^{(0)}$ . This changes  $S$  to

$$S = A_\theta^T (F_{\theta\theta})^{-1} A_\theta^{(0)} + A_\xi^T A_\xi . \quad (7.32)$$

In cases with constraints that depend only on unmeasured quantities, this procedure makes  $S$  non-singular. Everything else proceeds as before.

#### 7.4.1

##### Error Propagation

One of the main motivations behind performing a constrained fit is to obtain fitted parameters with errors that are reduced compared to the original measurement. Thus, the calculation of the covariance matrix of the fitted quantities is an important task. Before we approach this problem, we should be clear about one thing: in a constrained fit problem, the *Hessian matrix* (see Section 2.4 or [6])  $\nabla_{xx}^2 f$  of the objective function  $f$ , or of the Lagrange function  $\mathcal{L}$ , is *not* the inverse covariance matrix of the fitted quantities. If anything, the Hessian matrix may be the inverse covariance matrix of the measured input values to the fit. The covariance matrix of the fit result is considerably more complicated to obtain, as we shall see.

Indeed, it turns out that the inverse of the covariance matrix of the fitted parameters generally does not even exist: imposing the constraints means that one or several fit parameters can be calculated from the others, and therefore the error on certain functions (namely, the constraint functions) of the parameters is exactly zero. Translated into the properties of the covariance matrix this tells us that some of the eigenvalues of the covariance matrix are zero, and that the covariance

6) The column rank of a matrix is the number of linearly independent columns.

matrix does not have an inverse. However, subsets of the fitted parameters can be independent and have an invertible covariance matrix.

The key to the calculation of the covariance matrix of the fit parameters is the observation that the fit parameters  $\boldsymbol{\theta}$  and  $\xi$  are functions of the measured values  $t$ . The prescription ‘find the minimum of  $f(\boldsymbol{\theta}, t)$  subject to the constraints  $c(\boldsymbol{\theta}, \xi)$ ’ defines a function that maps the measured values  $t$  to the fit results  $\boldsymbol{\theta}$  and  $\xi$ . In the case of linear constraints and a quadratic objective function  $f$ , this mapping function can even be explicitly given and used for the error propagation. However, as we will see later when we consider the more general fit problem, an analytic solution is not necessary to perform the error propagation.

For now, we consider the case  $f = (\boldsymbol{\theta} - t)^T \mathbf{W}^{-1} (\boldsymbol{\theta} - t)$  with a symmetric weight matrix  $\mathbf{W}^{-1}$ , which could be the inverse of the covariance matrix of the measurements  $t$ , in which case  $f$  would simply be the usual  $\chi^2 = (\boldsymbol{\theta} - t)^T \mathbf{V}^{-1} (\boldsymbol{\theta} - t)$ . We first calculate the covariance matrix of the fit result for the case where no unmeasured quantities are present. We choose  $\boldsymbol{\theta}^{(0)} = \mathbf{0}$ , which gives:

$$f_{\boldsymbol{\theta}}^{(0)} = -2\mathbf{W}^{-1}t, \quad (7.33)$$

$$\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} = 2\mathbf{W}^{-1}, \quad (7.34)$$

$$\mathbf{r} = c^{(0)} + \mathbf{A}_{\boldsymbol{\theta}}^T t, \quad (7.35)$$

$$\mathbf{S} = \frac{1}{2}\mathbf{A}_{\boldsymbol{\theta}}^T \mathbf{W} \mathbf{A}_{\boldsymbol{\theta}}, \quad (7.36)$$

$$\lambda = \mathbf{S}^{-1} (c^{(0)} + \mathbf{A}_{\boldsymbol{\theta}}^T t), \quad (7.37)$$

$$\boldsymbol{\theta} = \left(1 - \frac{1}{2}\mathbf{W} \mathbf{A}_{\boldsymbol{\theta}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\theta}}^T\right) t - \frac{1}{2}\mathbf{W} \mathbf{A}_{\boldsymbol{\theta}} \mathbf{S}^{-1} c^{(0)}. \quad (7.38)$$

Standard error propagation gives us the covariance matrix  $\mathbf{V}_{\boldsymbol{\theta}}$  as a function of the covariance matrix  $\mathbf{V}_t$  of the measured quantities:

$$\mathbf{V}_{\boldsymbol{\theta}} = \left(1 - \frac{1}{2}\mathbf{W} \mathbf{A}_{\boldsymbol{\theta}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\theta}}^T\right)^T \mathbf{V}_t \left(1 - \frac{1}{2}\mathbf{W} \mathbf{A}_{\boldsymbol{\theta}} \mathbf{S}^{-1} \mathbf{A}_{\boldsymbol{\theta}}^T\right). \quad (7.39)$$

#### Example 7.4 Averaging as a constrained-fit problem

We consider the case where we have two measurements  $t_1$  and  $t_2$  of the same parameter  $\theta$  with a covariance matrix  $\mathbf{V}_t$  given by

$$\mathbf{V}_t = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

This covariance matrix is also used as weight matrix  $\mathbf{W} = \mathbf{V}_t$  that defines the objective function  $f$ , so that it is just the usual  $\chi^2$ . Now we average the two measurements by imposing the constraint  $c(\theta_1, \theta_2) = \theta_1 - \theta_2$ , which leads to

$$\mathbf{A} = \mathbf{A}_{\boldsymbol{\theta}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The result is:

$$\mathbf{c}^{(0)} = (0) , \quad (7.40)$$

$$\mathbf{f}_{\theta}^{(0)} = -2 \begin{pmatrix} t_1/\sigma_1^2 \\ t_2/\sigma_2^2 \end{pmatrix} , \quad (7.41)$$

$$\mathbf{F}_{\theta\theta} = \begin{pmatrix} 2\sigma_1^{-2} & 0 \\ 0 & 2\sigma_2^{-2} \end{pmatrix} , \quad (7.42)$$

$$\mathbf{r} = (t_1 - t_2) , \quad (7.43)$$

$$\mathbf{S} = \frac{1}{2} (\sigma_1^2 + \sigma_2^2) , \quad (7.44)$$

$$\lambda = \frac{2}{\sigma_1^2 + \sigma_2^2} (t_1 - t_2) , \quad (7.45)$$

$$\left( 1 - \frac{1}{2} \mathbf{W} \mathbf{A}_\theta \mathbf{S}^{-1} \mathbf{A}_\theta^\top \right) = \begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} & \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \\ \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} & \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \end{pmatrix} , \quad (7.46)$$

$$\boldsymbol{\theta} = \begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} t_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} t_2 \\ \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} t_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} t_2 \end{pmatrix} , \quad (7.47)$$

$$\mathbf{V}_\theta = \begin{pmatrix} \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} & \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \\ \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} & \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \end{pmatrix} , \quad (7.48)$$

$$\boldsymbol{\theta} = \begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} t_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} t_2 \\ \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} t_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} t_2 \end{pmatrix} . \quad (7.49)$$

So we see that the solution to this constrained fit problem is given by

$$\theta_1 = \theta_2 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} t_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} t_2 ,$$

which is the well-known weighted average result for the combination of two measurements with different errors. The error propagation also yields the expected result for the errors of  $\theta_1$  and  $\theta_2$ , which are 100% correlated.

### 7.4.2

#### Error Propagation in the Presence of Unmeasured Quantities

The set of equations (7.12) constitutes an implicit definition of the fitted quantities  $x_n$  (and  $\lambda_k$ ) as a function of the measurements  $t_m$ . Assuming that we know the covariance matrix  $\mathbf{V}_t$  of the measurements, the covariance matrix of the fitted values

$\mathbf{V}_x$  is given from standard error propagation by

$$(V_x)_{nn'} = \sum_{m,m'} \frac{\partial x_n}{\partial t_m} (V_t)_{mm'} \frac{\partial x_{n'}}{\partial t_{m'}} . \quad (7.50)$$

To evaluate  $\partial x_n / \partial t_m$ , we make a Taylor expansion for  $t$  around the result of the measurement  $t^0$  and write (7.28) as:

$$\begin{aligned} \begin{pmatrix} -f_{\theta}^{(0)} \\ \mathbf{0} \\ -c^{(0)} \end{pmatrix} + \begin{pmatrix} \mathbf{F}_{\theta t} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} (t - t^0) &= \begin{pmatrix} \mathbf{F}_{\theta\theta} & \mathbf{0} & \mathbf{A}_{\theta} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{\xi} \\ \mathbf{A}_{\theta}^T & \mathbf{A}_{\xi}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)} \\ \boldsymbol{\xi} - \boldsymbol{\xi}^{(0)} \\ \lambda \end{pmatrix} \\ &= \mathbf{M} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)} \\ \boldsymbol{\xi} - \boldsymbol{\xi}^{(0)} \\ \lambda \end{pmatrix} \end{aligned} \quad (7.51)$$

with the formal solution

$$\begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)} \\ \boldsymbol{\xi} - \boldsymbol{\xi}^{(0)} \\ \lambda \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} -f_{\theta}^{(0)} \\ \mathbf{0} \\ -c^{(0)} \end{pmatrix} - \mathbf{M}^{-1} \begin{pmatrix} \mathbf{F}_{\theta t} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} (t - t^0) . \quad (7.52)$$

With Gaussian block elimination one can calculate  $\mathbf{M}^{-1}$  with the result:

$$\mathbf{M}^{-1} = \begin{pmatrix} (\mathbf{M}^{-1})_{\theta\theta} & (\mathbf{M}^{-1})_{\theta\xi} & (\mathbf{M}^{-1})_{\theta\lambda} \\ (\mathbf{M}^{-1})_{\xi\theta} & (\mathbf{M}^{-1})_{\xi\xi} & (\mathbf{M}^{-1})_{\xi\lambda} \\ (\mathbf{M}^{-1})_{\lambda\theta} & (\mathbf{M}^{-1})_{\lambda\xi} & (\mathbf{M}^{-1})_{\lambda\lambda} \end{pmatrix} \quad (7.53)$$

with

$$\begin{aligned} \mathbf{T} &= \mathbf{A}_{\xi} \mathbf{S}^{-1} \mathbf{A}_{\xi}^T , \\ (\mathbf{M}^{-1})_{\theta\theta} &= \mathbf{F}_{\theta\theta}^{-1} - \mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_{\theta} \mathbf{S}^{-1} \mathbf{A}_{\theta}^T \mathbf{F}_{\theta\theta}^{-1} + \mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_{\theta} \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} \mathbf{A}_{\theta}^T \mathbf{F}_{\theta\theta}^{-1} , \\ (\mathbf{M}^{-1})_{\theta\xi} &= -\mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_{\theta} \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} , \\ (\mathbf{M}^{-1})_{\theta\lambda} &= \mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_{\theta} \mathbf{S}^{-1} - \mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_{\theta} \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} , \\ (\mathbf{M}^{-1})_{\xi\theta} &= -\mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} \mathbf{A}_{\theta}^T \mathbf{F}_{\theta\theta}^{-1} , \\ (\mathbf{M}^{-1})_{\xi\xi} &= \mathbf{T}^{-1} , \\ (\mathbf{M}^{-1})_{\xi\lambda} &= \mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} , \\ (\mathbf{M}^{-1})_{\lambda\theta} &= \mathbf{S}^{-1} \mathbf{A}_{\theta}^T \mathbf{F}_{\theta\theta}^{-1} - \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} \mathbf{A}_{\theta}^T \mathbf{F}_{\theta\theta}^{-1} , \\ (\mathbf{M}^{-1})_{\lambda\xi} &= \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} , \\ (\mathbf{M}^{-1})_{\lambda\lambda} &= -\mathbf{S}^{-1} + \mathbf{S}^{-1} \mathbf{A}_{\xi}^T \mathbf{T}^{-1} \mathbf{A}_{\xi} \mathbf{S}^{-1} . \end{aligned}$$

Using this, we finally get:

$$\begin{aligned}\frac{\partial \boldsymbol{\theta}}{\partial t} &= -(\mathbf{M}^{-1})_{\theta\theta} \mathbf{F}_{\theta t} \\ &= -[1 - \mathbf{F}_{\theta\theta}^{-1} \mathbf{A}_\theta \mathbf{S}^{-1} (1 - \mathbf{A}_\xi^T \mathbf{T}^{-1} \mathbf{A}_\xi \mathbf{S}^{-1}) \mathbf{A}_\theta^T] \mathbf{F}_{\theta\theta}^{-1} \mathbf{F}_{\theta t},\end{aligned}\quad (7.54)$$

$$\begin{aligned}\frac{\partial \xi}{\partial t} &= -(\mathbf{M}^{-1})_{\xi\theta} \mathbf{F}_{\theta t} \\ &= \mathbf{T}^{-1} \mathbf{A}_\xi \mathbf{S}^{-1} \mathbf{A}_\theta^T \mathbf{F}_{\theta\theta}^{-1} \mathbf{F}_{\theta t}.\end{aligned}\quad (7.55)$$

We note that in the usual case where the objective function  $f$  is a quadratic function of  $\boldsymbol{\theta} - \mathbf{t}$ , we have  $\mathbf{F}_{\theta t} = -\mathbf{F}_{\theta\theta}$ , so that  $\mathbf{F}_{\theta\theta}^{-1} \mathbf{F}_{\theta t} = -1$ . This simplification applies in particular when  $f$  is just a  $\chi^2$  expression. However, the more general form that we have derived can also be applied to other cases.

### Example 7.5 Three-body decay

In this example we consider an event with two highly energetic jets, the energies  $t_1$  and  $t_2$  which are measured with uncertainties  $\sigma_1$  and  $\sigma_2$ , and a photon that hits a calorimeter crack so that its energy is not measured. Assuming that the particles' azimuthal angles are known with negligible error, we have three unknowns  $\theta_1$ ,  $\theta_2$  and  $\xi_1$ , two measurements  $t_1$ ,  $t_2$  of the energies and two constraints from momentum conservation in the transverse plane ( $\phi_1$ ,  $\phi_2$ ,  $\phi_3$  are the azimuthal angles of the jets and the photon):

$$\begin{aligned}c_1 &= \cos \phi_1 \cdot \theta_1 + \cos \phi_2 \cdot \theta_2 + \cos \phi_3 \cdot \xi_1, \\ c_2 &= \sin \phi_1 \cdot \theta_1 + \sin \phi_2 \cdot \theta_2 + \sin \phi_3 \cdot \xi_1.\end{aligned}$$

This results in

$$\begin{aligned}\mathbf{A}_\theta &= \begin{pmatrix} \cos \phi_1 & \sin \phi_1 \\ \cos \phi_2 & \sin \phi_2 \end{pmatrix}, \\ \mathbf{A}_\xi &= (\cos \phi_3 \quad \sin \phi_3), \\ \mathbf{S} &= \frac{1}{2} \begin{pmatrix} \sigma_1^2 \cos^2 \phi_1 + \sigma_2^2 \cos^2 \phi_2 & \sigma_1^2 \sin \phi_1 \cos \phi_1 + \sigma_2^2 \sin \phi_2 \cos \phi_2 \\ \sigma_1^2 \sin \phi_1 \cos \phi_1 + \sigma_2^2 \sin \phi_2 \cos \phi_2 & \sigma_1^2 \sin^2 \phi_1 + \sigma_2^2 \sin^2 \phi_2 \end{pmatrix}.\end{aligned}$$

For the rest of the exercise we use the following numbers:

$$\begin{aligned}t_1 &= 123, & \sigma_1 &= 5.5, & \phi_1 &= 0.63, \\ t_2 &= 154, & \sigma_2 &= 6.2, & \phi_2 &= 2.55, \\ & & & & \phi_3 &= -1.44.\end{aligned}$$

Numerically, we get the system of equations (7.51):

$$\begin{pmatrix} 8.132 \\ 8.012 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.06612 & 0.0000 & 0 & 0.8080 & 0.5891 \\ 0.0000 & 0.05203 & 0 & -0.8301 & 0.5577 \\ 0 & 0 & 0 & 0.13042 & -0.9915 \\ 0.8080 & -0.8301 & 0.13042 & 0 & 0 \\ 0.5891 & 0.5577 & -0.9915 & 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \xi_1 \\ \lambda_1 \\ \lambda_2 \end{pmatrix}$$

with the inverse matrix

$$\mathbf{M}^{-1} = \begin{pmatrix} 7.280 & 8.519 & 9.118 & 0.5858 & 0.07705 \\ 8.519 & 9.970 & 10.670 & -0.6361 & -0.08367 \\ 9.118 & 10.670 & 11.420 & -0.0097 & -1.00989 \\ 0.5858 & -0.6361 & -0.0097 & -0.0437 & -0.00575 \\ 0.07705 & -0.08367 & -1.00989 & -0.00575 & -0.00076 \end{pmatrix}$$

and the solution

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \xi_1 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 127.46 \\ 149.16 \\ 159.64 \\ -0.3328 \\ -0.0438 \end{pmatrix}.$$

For the derivatives we get

$$\frac{\partial}{\partial \mathbf{t}} \begin{pmatrix} \boldsymbol{\theta} \\ \xi \end{pmatrix} = \begin{pmatrix} (\mathbf{M}^{-1})_{\theta\theta} \\ (\mathbf{M}^{-1})_{\theta\xi} \end{pmatrix} \mathbf{F}_{\theta t} = \begin{pmatrix} 0.4813 & 0.4432 \\ 0.5632 & 0.5187 \\ 0.6028 & 0.5551 \end{pmatrix}.$$

The final solution and the covariance matrix are

$$\begin{aligned} \theta_1 &= 127.46 \pm 3.82 \\ \theta_2 &= 149.16 \pm 4.47 \quad \text{and} \quad \mathbf{V}_{\theta,\xi} = \begin{pmatrix} 14.559 & 17.038 & 18.235 \\ 17.038 & 19.939 & 21.340 \\ 18.235 & 21.340 & 22.839 \end{pmatrix}. \\ \xi_1 &= 159.64 \pm 4.78 \end{aligned}$$

One can see that the fitted values are more precise than the measured input values. A closer inspection of the covariance matrix shows that all correlation coefficients are unity, that is all three fitted values are a function of only a single variable. This is expected because the three quantities are connected by two constraints, which leaves exactly one degree of freedom.

## 7.5

### Iterative Solution of the Lagrange Multiplier Problem

In the previous section we discussed the Lagrange multiplier problem for the case of a quadratic objective function  $f$  and linear constraints, which is a problem with an analytic solution.

While a quadratic objective function is indeed a very common case in high energy physics applications, where often a  $\chi^2$  expression is minimised, the constraints are often non-linear. Non-linear constraints make it generally necessary to revert to iterative solutions, which we will discuss in the following.

Although the basic ingredients of an iterative algorithm are not much harder to understand than the analytic solution for the linear case, writing a program that

solves a variety of problems without running into numerical difficulties, with good convergence properties and reasonable speed can become a quite demanding task. This section, which can be skipped by the more casual reader, cannot give an exhaustive guide to writing such a program. Instead, it aims at giving an introduction to the difficulties that arise in the solution of such problems, in the hope that it may help to use existing fit programs more efficiently and better understand why they sometimes fail. For the more adventurous character that wants to write a new fitter or improve an existing one, this section may serve as an initial guide to problems he/she may encounter and to the vast amount of literature on the subject of constrained optimisation.

The iterative solution of the system of equations given by the Lagrange multiplier method has four distinct parts:

- finding appropriate starting values,
- choosing a step direction,
- choosing a step length, and
- detecting convergence.

After the starting values have been determined, the necessary number of iteration steps are performed until convergence is detected. These steps are considered in more detail in the following.

### 7.5.1

#### Choosing a Direction

Consider the Lagrange function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = f(\boldsymbol{\theta}) + \sum_{k=1}^K \lambda_k \cdot c_k(\boldsymbol{\theta}, \boldsymbol{\xi}), \quad (7.56)$$

for which we seek a stationary point,<sup>7)</sup> that is

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \theta_m} = \frac{\partial f}{\partial \theta_m} + \sum_{k=1}^K \lambda_k \cdot \frac{\partial c_k}{\partial \theta_m} \quad m = 1, \dots, M, \\ 0 &= \frac{\partial \mathcal{L}}{\partial \xi_u} = \sum_{k=1}^K \lambda_k \cdot \frac{\partial c_k}{\partial \xi_u} \quad u = 1, \dots, U, \\ 0 &= \frac{\partial \mathcal{L}}{\partial \lambda_k} = c_k \quad k = 1, \dots, K. \end{aligned} \quad (7.57)$$

When we arrange the parameters into a vector  $\mathbf{X}$  with  $P = M + U + K$  components,

$$\mathbf{X}^T = (\theta_1, \dots, \theta_M, \xi_1, \dots, \xi_U, \lambda_1, \dots, \lambda_K), \quad (7.58)$$

7) Note that  $f$  does not depend on the ‘unmeasured’ parameters  $\boldsymbol{\xi}$ .

and the derivatives into a vector

$$\mathbf{Y}^T = \left( \frac{\partial \mathcal{L}}{\partial \theta_1}, \dots, \frac{\partial \mathcal{L}}{\partial \theta_M}, \frac{\partial \mathcal{L}}{\partial \xi_1}, \dots, \frac{\partial \mathcal{L}}{\partial \xi_U}, c_1, \dots, c_K \right)^T, \quad (7.59)$$

we can write this system of equations succinctly as:

$$\mathbf{Y} = \mathbf{0}. \quad (7.60)$$

If this system of equations cannot be solved analytically, an obvious choice for an iterative procedure is given by the Newton–Raphson method [12]:

$$\Delta \mathbf{X} = \mathbf{X}^{(\nu+1)} - \mathbf{X}^{(\nu)} = - \left( \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} \right)^{-1} \mathbf{Y}, \quad (7.61)$$

where the superscript  $(\nu)$  indicates that the values are those of iteration<sup>8)</sup>  $\nu$ . In other terms, we seek the solution to the system of equations given by

$$\begin{pmatrix} -\mathbf{g}_\theta \\ -\mathbf{g}_\xi \\ -\mathbf{c} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{\theta\theta} & \mathbf{L}_{\theta\xi} & \mathbf{A}_\theta \\ \mathbf{L}_{\theta\xi}^T & \mathbf{L}_{\xi\xi} & \mathbf{A}_\xi \\ \mathbf{A}_\theta^T & \mathbf{A}_\xi^T & 0 \end{pmatrix} \begin{pmatrix} \Delta \boldsymbol{\theta} \\ \Delta \boldsymbol{\xi} \\ \Delta \boldsymbol{\lambda} \end{pmatrix}, \quad (7.62)$$

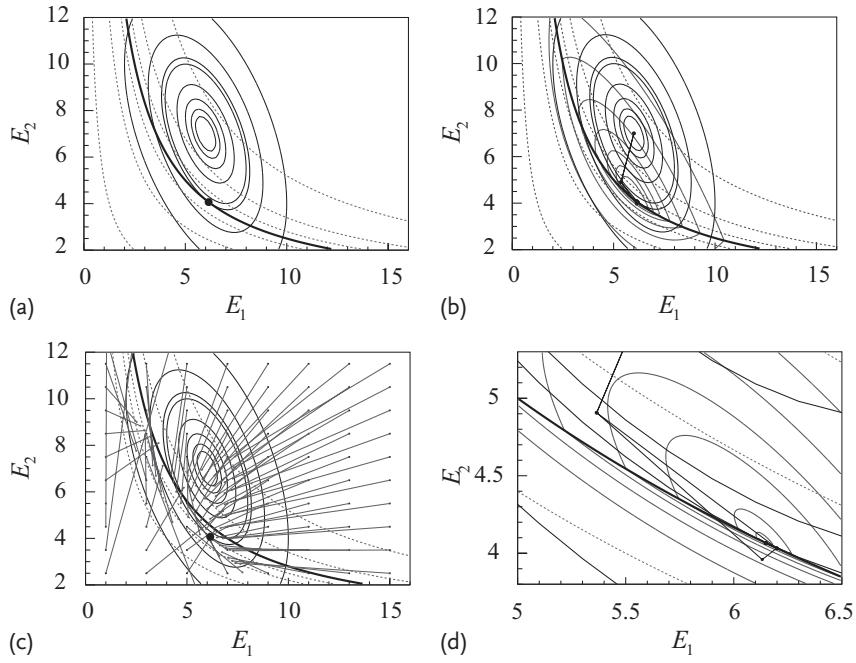
where we use the notation

$$\begin{aligned} \mathbf{g}_\theta &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} &= \frac{\partial f}{\partial \boldsymbol{\theta}} + \sum_{k=1}^K \lambda_k^{(\nu)} \cdot \frac{\partial c_k}{\partial \boldsymbol{\theta}} &= \mathbf{f}_\theta + \mathbf{A}_\theta \boldsymbol{\lambda}^{(\nu)}, \\ \mathbf{g}_\xi &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} &= \sum_{k=1}^K \lambda_k^{(\nu)} \cdot \frac{\partial c_k}{\partial \boldsymbol{\xi}} &= \mathbf{A}_\xi \boldsymbol{\lambda}^{(\nu)}, \\ \mathbf{L}_{\theta\theta} &= \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \frac{\partial^2 f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \sum_{k=1}^K \lambda_k^{(\nu)} \cdot \frac{\partial^2 c_k}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}, \\ \mathbf{L}_{\theta\xi} &= \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\xi}^T} &= \sum_{k=1}^K \lambda_k^{(\nu)} \cdot \frac{\partial^2 c_k}{\partial \boldsymbol{\theta} \partial \boldsymbol{\xi}^T}, \\ \mathbf{L}_{\xi\xi} &= \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} &= \sum_{k=1}^K \lambda_k^{(\nu)} \cdot \frac{\partial^2 c_k}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T}, \\ \mathbf{A}_\theta &= \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\theta}^T} &= \frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}, \\ \mathbf{A}_\xi &= \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\xi}^T} &= \frac{\partial \mathbf{c}}{\partial \boldsymbol{\xi}}. \end{aligned}$$

We can write this as

$$\mathbf{Y} = \mathbf{M}(\mathbf{X}^{(\nu+1)} - \mathbf{X}^{(\nu)}). \quad (7.63)$$

8) In the following, the superscript will only be used when values at different iterations  $\nu$  and  $\nu + 1$  occur; if no iteration index is given, values are assumed to be evaluated for iteration  $\nu$ .



**Figure 7.3** Illustration of the solution to the first example,  $\pi^0 \rightarrow \gamma\gamma$  with Lagrange multipliers. (a) Isolines of constant objective function  $f$  (at values of 0.25, 0.5, 1, 2, 4, **4.743**) and constraint function  $c_1$  (at values  $-20, -10, -3, \mathbf{0}$  (bold line), 3, 9 and 27) in the plane spanned by  $E_1$  and  $E_2$ ; (b) the path followed by the minimisation algorithm, starting at the unconstrained mini-

mum off at  $E_1 = 6, E_2 = 7$ , overlaid with the isolines of the merit function  $\phi_1$ ; (c) arrows depicting the Newton steps (without step size control) starting at various values of  $E_1$  and  $E_2$ ; (d) a close-up version of (b) around the final solution. The step with the small dot at the corner involves a quadratic correction to avoid the Maratos effect (discussed in Section 7.5.2.5).

An equivalent way to write this system of equations is

$$\begin{pmatrix} -f_\theta \\ \mathbf{0} \\ -c \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{\theta\theta} & \mathbf{L}_{\theta\xi} & \mathbf{A}_\theta \\ \mathbf{L}_{\xi\theta}^T & \mathbf{L}_{\xi\xi} & \mathbf{A}_\xi \\ \mathbf{A}_\theta^T & \mathbf{A}_\xi^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Delta\theta \\ \Delta\xi \\ \lambda^{(v+1)} \end{pmatrix}, \quad (7.64)$$

where  $\lambda^{(v+1)}$  is calculated directly, rather than via the update relationship  $\lambda^{(v+1)} = \lambda^{(v)} + \Delta\lambda$ . In this form, the calculation of the left-hand side is simplified.

Let us look more closely at the matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{L}_{\theta\theta} & \mathbf{L}_{\theta\xi} & \mathbf{A}_\theta \\ \mathbf{L}_{\xi\theta}^T & \mathbf{L}_{\xi\xi} & \mathbf{A}_\xi \\ \mathbf{A}_\theta^T & \mathbf{A}_\xi^T & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{L} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{pmatrix} \quad (7.65)$$

and its submatrices  $\mathbf{L}$  and  $\mathbf{A}$ . The matrices  $\mathbf{L}_{\theta\theta}$  and  $\mathbf{L}_{\xi\xi}$  are obviously symmetric, and therefore  $\mathbf{L}$  and  $\mathbf{M}$  are symmetric as well. We assume for the time being that

$\mathbf{M}$  is non-singular, that is (7.62) has a unique solution.  $\mathbf{L}_{\theta\theta}$  will be typically non-singular and thus invertible, because of the properties of the objective function. However, in the presence of unmeasured parameters, many or all elements of  $\mathbf{L}_{\theta\xi}$  and  $\mathbf{L}_{\xi\xi}$  can be zero (in particular if all  $\lambda_k$  are zero), so in the presence of unmeasured parameters,  $\mathbf{L}$  is a symmetric indefinite matrix, which is not invertible. This severely limits the application of algorithms that are based on block elimination, as described in Section 7.4.

Figure 7.3c shows the step calculated by the Newton–Raphson method for a grid of starting points. We see that in the vicinity of the solution, the steps given by the method point in the right direction and have a reasonable length, but for starting points too far away from the final solution the Newton step may be much too long and either overshoot the solution considerably or in fact point towards an unfavourable direction altogether. But before we turn to the issue of step-length control, which addresses this problem, let us discuss the case that the matrix  $\mathbf{M}$  does not have full rank and is therefore not invertible.

#### 7.5.1.1 Coping with a Rank-Deficient Matrix $\mathbf{M}$

It may happen that the matrix  $\mathbf{M}$  does not have full rank, which means that the systems of (7.62) or (7.64) do not have unique solutions, or that  $\mathbf{M}$  is nearly singular. This can occur during the iterative solution, in particular if non-optimal starting values are used, or if the constraints do not hold any information about one parameter. For instance, if a momentum vector is parameterised with polar coordinates  $(p, \theta, \phi)$ , then for  $\theta = 0$  all derivatives of the cartesian components with respect to  $\phi$  will vanish.

In such a case, the non-existing inverse  $\mathbf{M}^{-1}$  may be replaced by the Moore–Penrose pseudo-inverse  $\mathbf{M}^+$  [13, 14]. Consider the eigenvector decomposition of  $\mathbf{M}$ ,

$$\mathbf{M} = \mathbf{O}\Sigma\mathbf{O}^T, \quad (7.66)$$

with an orthogonal matrix  $\mathbf{O}$  and a diagonal matrix  $\Sigma = \text{diag}(\sigma_1 \dots \sigma_p)$ , where the eigenvalues are ordered such that  $|\sigma_1| \geq |\sigma_2| \geq \dots \geq |\sigma_p|$ . Then, the pseudo-inverse  $\mathbf{M}^+$  is given by ([11], Section 5.5.4):

$$\mathbf{M}^+ = \mathbf{O}\Sigma^+\mathbf{O}^T \quad (7.67)$$

with

$$\Sigma^+ = \text{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0\right). \quad (7.68)$$

This pseudo-inverse can be calculated without performing a complete eigenvector decomposition, which is computationally very expensive, by means of the QR decomposition [11], where  $\mathbf{M}$  is decomposed as  $\mathbf{M} = \mathbf{QR}$  with an orthogonal matrix  $\mathbf{Q}$  and an upper triangular matrix  $\mathbf{R}$ .

### 7.5.1.2 Normalisation of Parameters and Constraint Functions

The Moore–Penrose pseudo-inverse has an important minimisation property: consider any system of equations of the form

$$\mathbf{L}\mathbf{x} = \mathbf{b} \quad (7.69)$$

with a quadratic  $n \times n$  matrix  $\mathbf{L}$  of rank  $r < n$ . Then the Moore–Penrose pseudo-inverse  $\mathbf{L}^+$  gives us the solution

$$\mathbf{x}_{\text{LS}} = \mathbf{L}^+ \mathbf{b} \quad (7.70)$$

which has the smallest Euclidean norm  $\|\mathbf{x}_{\text{LS}}\|_2$  of all possible solutions of (7.69), if the system is consistent, that is it has an exact solution,<sup>9)</sup> see Section 5.5 of [11].

This property is intuitively desirable, because it means that in a situation where the system of equations does not specify how far a step in the iteration procedure should go in a certain direction, we simply do not go along that direction at all.

However, recall that in the iteration problem at hand, the vector  $\mathbf{X}$  is given by

$$\mathbf{X}^T = (\theta_1, \dots, \theta_M, \xi_1, \dots, \xi_U, \lambda_1, \dots, \lambda_K). \quad (7.71)$$

From a physical point of view, the parameters  $\theta_m$ ,  $\xi_u$  and  $\lambda_k$  all have different units (since  $\chi^2$  and hence  $\mathcal{L}$  are dimensionless, the dimension of  $\lambda_k$  must be the inverse of the dimension of the corresponding constraint  $c_k$ ). In such a situation, the norm of  $\mathbf{x}$  is completely meaningless.

This points to a deeper problem: as long as the components of  $\mathbf{x}$  have different physical units (say, GeV and cm), there is no way to say whether a step  $\Delta\mathbf{x}$  is sufficiently small or whether a constraint violation is sufficiently small to indicate convergence. For example,  $c(\mathbf{x}) = 0.01$  can be quite small if constraint  $c$  is a sum of jet momenta measured in GeV, or large if  $c$  is a vertex constraint indicating that some tracks are apart by  $0.01 \text{ cm} = 100 \mu\text{m}$ . For the same reason, the eigenvalues of matrix  $\mathbf{M}$  are meaningless, and  $\mathbf{M}$  may be ill-conditioned, meaning that the smallest eigenvalues are much smaller than the largest ones, simply because we have mixed quantities with very different units.

However, we know the typical scale on which the measured parameters  $\theta_m$  should vary, which is given by the measurement errors  $\delta\theta_m$ . Likewise, for unmeasured parameters we can define error estimates  $\delta\xi_u$  which indicate the expected accuracy, for example  $\delta\xi = 0.0001 \text{ cm}$  for a vertex constraint or  $\delta\xi = 1 \text{ GeV}$  for a neutrino momentum. Using standard error propagation, we can easily calculate an error estimate  $\delta c$  for each constraint:

$$(\delta c)^2 = \sum_{m,m'=1}^M \frac{\partial c}{\partial \theta_m} V_{m,m'} \frac{\partial c}{\partial \theta_{m'}} + \sum_u^U \frac{\partial c}{\partial \xi_u} (\delta \xi_u)^2 \frac{\partial c}{\partial \xi_u}. \quad (7.72)$$

These error estimates can be used to define a diagonal matrix<sup>10)</sup>

$$\mathbf{D} = \text{diag}(\delta\theta_1, \dots, \delta\theta_M, \delta\xi_1, \dots, \delta\xi_U, (\delta c_1)^{-1}, \dots, (\delta c_K)^{-1}) \quad (7.73)$$

9) If the system (7.69) is not consistent, then  $\mathbf{x}_{\text{LS}}$  will minimise the Euclidean norm of the residual  $\|\mathbf{L}\mathbf{x}_{\text{LS}} - \mathbf{b}\|_2$ .

10) If we divide the constraints  $c_k$  by their respective errors  $\delta c_k$ , then in turn we have to divide the Lagrange multipliers  $\lambda_k$  by  $(\delta c_k)^{-1}$ .

and to rewrite our equations as

$$\mathbf{D} \mathbf{Y} = -\mathbf{DMDD}^{-1}(\mathbf{X}^{(\nu+1)} - \mathbf{X}^{(\nu)}) \quad (7.74)$$

or

$$\tilde{\mathbf{Y}} = -\tilde{\mathbf{M}}(\tilde{\mathbf{X}}^{(\nu+1)} - \tilde{\mathbf{X}}^{(\nu)}) \quad (7.75)$$

with

$$\tilde{\mathbf{Y}} = \mathbf{D} \mathbf{Y}, \quad \tilde{\mathbf{M}} = \mathbf{DMD}, \quad \tilde{\mathbf{X}} = \mathbf{D}^{-1} \mathbf{X}. \quad (7.76)$$

We see that now all components of  $\tilde{\mathbf{Y}}$ ,  $\tilde{\mathbf{M}}$  and  $\tilde{\mathbf{X}}$  are dimensionless. The components of

$$\Delta \tilde{\mathbf{X}} = \tilde{\mathbf{X}}^{(\nu+1)} - \tilde{\mathbf{X}}^{(\nu)} \quad (7.77)$$

now indicate the step length in units of the uncertainties of the components of  $\tilde{\mathbf{X}}$ , which is certainly a sensible measure. Also, the last  $K$  components of  $\tilde{\mathbf{Y}}$  are now the values of the constraint functions, divided by the uncertainty on the constraints coming from the measurements.

### 7.5.2

#### Controlling the Step Length

Iterative methods, such as the Newton–Raphson method, tend to converge very quickly once the current estimate is sufficiently close to the solution. However, the region of convergence is generally surrounded by a region where the steps predicted by the Newton method are too large and overshoot the true solution. Without controlling – that is, reducing – the step length, the method will not converge for starting points in this region. Therefore, it is crucial for a successful algorithm to control whether a step actually constitutes an improvement, and to thus extend the region of convergence.

In constrained fit problems there are two, often conflicting, indicators of ‘improvement’:

- the reduction of the objective function  $f$ , and
- the reduction of the violation of the constraints.

To decide whether a given step constitutes an improvement, these two measures have to be combined.

Here, we will describe the following approach: we define a merit function  $\phi(\mathbf{x})$  that indicates whether a step  $\alpha \mathbf{p}$  with  $\mathbf{p} = \mathbf{x}^{(\nu+1)} - \mathbf{x}^{(\nu)}$  is advantageous, and perform a line search along  $\mathbf{x} + \alpha \mathbf{p}$  (where the scalar  $\alpha$  is varied) until a step is found that improves the current solution. This approach needs several ingredients:

- a choice of the merit function and its parameters,

- a line search algorithm to find better values of  $\alpha$  efficiently, and
- a stopping condition that determines whether a given step is sufficiently good to be accepted.

To achieve good performance, it is necessary that the value of  $\alpha$  found is not too small, because many unnecessarily small steps obviously slow down convergence. More precisely, in the region around the true solution where the Newton method would converge without step-size control, the value of  $\alpha$  should be equal to one. We will return to this point later when the Maratos effect is discussed. Also, the line search itself should be fast, which means that the merit function should be computationally inexpensive to evaluate and the line search algorithm should find a good value of  $\alpha$  with few iterations (if any).

#### 7.5.2.1 Merit Function

One method is to define a merit function that becomes minimal at the true solution. One popular choice is the so-called  $\ell_1$  *penalty function* [10]:

$$\phi_1(\mathbf{x}; \mu) = f(\mathbf{x}) + \mu \|\mathbf{c}(\mathbf{x})\|_1. \quad (7.78)$$

Here,  $\|\mathbf{c}\|_1 = \sum_k |c_k|$  denotes the *Manhattan norm* or *taxicab norm* of vector  $\mathbf{c}$ .

It turns out [15] that for sufficiently large values of  $\mu > \mu^*$ , this penalty function is *exact*, which means that the solution of the constrained minimisation problem is the global minimum of the merit function  $\phi_1$ . A possible choice of  $\mu^*$  is given by

$$\mu^* = \max(\lambda_k^*), \quad (7.79)$$

where  $\lambda_k^*$  denotes the values of the Lagrange multipliers at the true solution. Figure 7.3b,d shows isolines of the merit function for the example problem.

Of course, this raises the following question: if  $\phi_1$  has a global minimum at the desired solution of the constrained minimisation problem, why don't we simply employ any off-the-shelf minimisation procedure on  $\phi_1$  and be done with it? One problem lies in the definition of  $\mu^*$ , which is unknown until the solution has been found. Choosing an arbitrarily large value of  $\mu$  is not a good strategy, because it will lead to a very ill-conditioned minimisation problem, where the solution lies at the bottom of a valley with very steep sides and a curved floor. Iterative procedures tend to converge very slowly in such a case. Moreover,  $\phi_1$  is not differentiable at all points where any constraint is zero, that is precisely in the subspace where we seek the minimum of  $f$ , which precludes the use of all minimisation algorithms that rely on derivatives. Figure 7.3b,d shows this nicely: all isolines of the merit function have kinks on the line where the constraint function is zero.

The *directional derivative*  $D(\phi_1(\mathbf{x}), \mathbf{p})$ , defined as

$$D(\phi_1(\mathbf{x}), \mathbf{p}) = \lim_{\epsilon \rightarrow 0} \frac{\phi_1(\mathbf{x} + \epsilon \mathbf{p}) - \phi_1(\mathbf{x})}{\epsilon} \quad (7.80)$$

of  $\phi_1$  is given by

$$D(\phi_1(\mathbf{x}), \mathbf{p}) = \nabla f^\top \mathbf{p} - \mu \|\mathbf{c}\|_1 \quad (7.81)$$

if the direction  $\mathbf{p}$  satisfies  $\mathbf{A}\mathbf{p} = -\mathbf{c}$ .

If  $\mathbf{p}$  and  $\boldsymbol{\lambda}^{(v+1)}$  solve the system

$$\begin{pmatrix} \nabla_{xx}^2 \mathcal{L} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{p} \\ \boldsymbol{\lambda}^{(v+1)} \end{pmatrix} = - \begin{pmatrix} \nabla f \\ \mathbf{c} \end{pmatrix}, \quad (7.82)$$

where  $\nabla_{xx}^2 \mathcal{L}$  is the second derivative of the Lagrangian, then the inequality [15, 16]

$$D(\phi_1(\mathbf{x}), \mathbf{p}) \leq -\mathbf{p}^T \nabla_{xx}^2 \mathcal{L} \mathbf{p} - (\mu - \|\boldsymbol{\lambda}^{(v+1)}\|_\infty) \|\mathbf{c}\|_1 \quad (7.83)$$

holds.<sup>11)</sup> This means that the directional derivative is negative, implying that  $\mathbf{p}$  is a descent direction for  $\phi_1(\mathbf{x})$  if

- $\mu$  is large enough, namely  $\mu > \|\boldsymbol{\lambda}^{(v+1)}\|_\infty$ , and
- $\nabla_{xx}^2 \mathcal{L}$  is positive definite (more precisely,  $\mathbf{d}^T \nabla_{xx}^2 \mathcal{L} \mathbf{d} > 0$  for all feasible directions  $\mathbf{d}$  for which  $\mathbf{A}\mathbf{d} = \mathbf{0}$ ).

The second condition is important, because it is not always fulfilled by the second-derivative matrix  $\mathbf{L}$  of the Lagrangian (remember the definition  $\mathbf{L} = \nabla_{xx}^2 f + \sum_{k=1}^K \lambda_k \nabla_{xx}^2 c_k$ ). While we can generally assume that the Hessian matrix  $\nabla_{xx}^2 f$  of the objective function  $f$  is positive definite (at least for directions fulfilling  $\mathbf{A}\mathbf{d} = \mathbf{0}$ ), this is not generally true for the Hessian matrices  $\nabla_{xx}^2 c_k$  of the constraints, or for  $\nabla_{xx}^2 \mathcal{L}$ . See Section 18.3 of [15] and Section 17.1 of [16] for a more in-depth treatment of this subject.

From (7.83) it follows immediately that if  $\nabla_{xx}^2 \mathcal{L}$  is positive definite, that is  $\mathbf{p}^T \nabla_{xx}^2 \mathcal{L} \mathbf{p} > 0$ , the merit function will decrease in direction  $\mathbf{p}$  if

$$\mu > \|\boldsymbol{\lambda}^{(v+1)}\|_\infty. \quad (7.84)$$

Another possible choice is

$$\mu \geq \frac{\nabla f \cdot \mathbf{p}}{(1 - \rho) \|\mathbf{c}\|_1} \quad (7.85)$$

with some value  $0 < \rho < 1$ ; using (7.81) this leads to  $D(\phi_1(\mathbf{x}), \mathbf{p}) \leq -\rho\mu \|\mathbf{c}\|_1$ . However, when the constraints are already fulfilled to a good degree, that is when  $\|\mathbf{c}\|_1$  is small, the resulting value of  $\mu$  becomes quite large. See Section 18.3 of [15] for more possibilities to choose  $\mu$ .

### 7.5.2.2 Line Searches

The purpose of the line search is to find a value of  $\alpha$  such that the merit function  $\phi(\mathbf{x} + \alpha \mathbf{p})$  is better by a sufficient margin than the merit function  $\phi(\mathbf{x})$  at the starting point. One may be tempted to try and find the true minimum of the merit function along the ray given by  $\mathbf{p}$ , that is to minimise the function  $q(\alpha) = \phi(\mathbf{x} + \alpha \mathbf{p})$ . However, this can be quite time-consuming, and in practice it turns out that a

<sup>11)</sup> The norm  $\|\mathbf{x}\|_\infty$  is simply the largest absolute value of all components of  $\mathbf{x}$ .

full minimisation of  $\phi$  often costs more time than it saves compared to algorithms where the search is stopped when an acceptable value of  $\alpha$  has been found, rather than the optimum.

Since the merit function decreases in the vicinity of the current value of  $x$ , the derivative  $q'$  with respect to  $\alpha$  is  $q'(\alpha) < 0$ , which means that we will always find positive values of  $\alpha$  that fulfil the stopping conditions. Since we need not go beyond the full Newton step, we restrict the search to the range  $0 < \alpha \leq 1$ . We observe that  $q(\alpha)$  is a continuous function, but generally not continuously differentiable, that is we expect  $q(\alpha)$  to have ‘kinks’ at places where the constraint functions change sign.

Before we discuss practical algorithms for the line search, we have to define what an ‘acceptable’ step is.

#### 7.5.2.3 Stopping Conditions

Various stopping conditions have been proposed in the literature. The three conditions discussed in the following are illustrated in Figure 7.4.

**The Armijo condition** [17], also called the *sufficient-decrease condition*, is the simplest, and a very popular, stopping condition. A step is considered acceptable if

$$q(\alpha) \leq q(0) + c_1 \alpha q'(0) \quad (7.86)$$

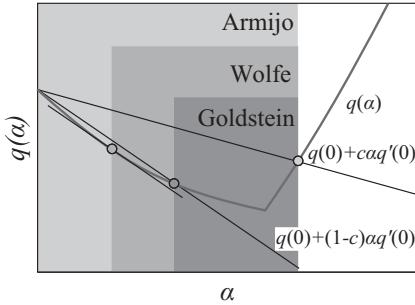
or

$$\phi(x + \alpha p) \leq \phi(x) + c_1 \alpha \nabla \phi^T p, \quad (7.87)$$

with some constant  $c_1$  that satisfies  $0 < c_1 < 1$ . Contrary to the naive expectation it turns out that  $c_1$  can be chosen to be quite small, for instance  $c_1 = 10^{-4}$ . At least,  $c_1$  should satisfy  $c_1 < 1/2$ , because otherwise a step size that is smaller than the optimum one would be chosen for a purely quadratic function  $q(\alpha)$ .

**The Wolfe conditions** [18, 19] add a second requirement to the Armijo condition, namely the curvature condition that

$$q'(\alpha) \geq c_2 q'(0) \quad (7.88)$$



**Figure 7.4** Allowed regions in a line search given by the Armijo, Goldstein and Wolfe conditions for the merit function  $q(\alpha)$  as a function of the scale parameter  $\alpha$ .

or

$$\nabla\phi(\mathbf{x} + \alpha\mathbf{p})^T\mathbf{p} \geq c_2\nabla\phi^T\mathbf{p} \quad (7.89)$$

with  $c_1 < c_2 < 1$ . The curvature condition forbids steps that are so short that  $q(\alpha)$  is still falling too fast, so that a larger value of  $\alpha$  promises a significantly better value of  $q$ .

**The Goldstein conditions** [20, 21] are an alternative to the Wolfe conditions and demand that

$$q(0) + (1 - c)\alpha q'(0) \leq q(\alpha) \leq q(0) + c\alpha q'(0) \quad (7.90)$$

or

$$\phi(\mathbf{x}) + (1 - c)\alpha\nabla\phi(\mathbf{x})^T\mathbf{p} \leq \phi(\mathbf{x} + \alpha\mathbf{p})^T\mathbf{p} \leq \phi(\mathbf{x}) + c\alpha\nabla\phi(\mathbf{x})^T\mathbf{p}, \quad (7.91)$$

with  $0 < c < 1/2$ . Again, in addition to the sufficient-decrease condition, steps are rejected that are too small. Compared to the Wolfe conditions, the Goldstein conditions do not require the knowledge of the gradient  $q'(\alpha)$  at the test point. Conversely, the true minimiser of  $q(\alpha)$  may fail to fulfil the Goldstein conditions.

#### 7.5.2.4 Practical Choice of the Step Length

If the Armijo condition is chosen as a stopping condition, a simple bisection algorithm is sufficient to find an allowed value of  $\alpha$ : start with  $\alpha = 1$ ; stop, if  $q(\alpha)$  fulfils the stopping condition, otherwise set  $\alpha = \beta\alpha$ , with some constant  $0 < \beta < 1$ .

In the case of the Wolfe or Goldstein conditions, a violation of (7.88) or (7.90) indicates that the chosen value of  $\alpha$  is too small; therefore, a bracketing algorithm is needed which repeatedly searches a new  $\alpha$  between a left and a right value  $\alpha_L < \alpha < \alpha_R$ , for instance by setting  $\alpha = (1 - \beta)\alpha_L + \beta\alpha_R$ . If  $\alpha$  violates (7.86), set  $\alpha_R = \alpha$  and iterate. Stop if the new  $\alpha$  fulfils the stopping conditions. If  $\alpha$  violates (7.88) or (7.90), set  $\alpha_L = \alpha$  and iterate. Care must be taken during the initialisation: before the iterations can start, a sufficiently large  $\alpha_R$  must be found such that for  $\alpha = \alpha_R$  (7.88) or (7.90) are fulfilled, which is not guaranteed for  $\alpha_R = 1$ , in which case  $\alpha_R$  must be increased before the iterations can start.

Choosing the new value of  $\alpha$  in the interval  $\alpha_L < \alpha < \alpha_R$  by just subdividing it according to a fixed fraction  $\beta$  is simple, but not necessarily the most efficient way. If the Wolfe conditions are applied, naturally the values  $q_{L,R}$  and derivatives  $q'_{L,R}$  of  $q$  are known at the interval borders  $\alpha_L$  and  $\alpha_R$ . From these four values, an interpolating cubic polynomial can be extracted and its minimum can be found by the following stepwise calculation (see [16], Section 3.2):

$$p = q'_R + q'_L - 3\frac{q_R - q_L}{\alpha_R - \alpha_L}, \quad (7.92)$$

$$D = p^2 - q'_R q'_L, \quad (7.93)$$

$$\beta = \frac{\sqrt{D} - p + q'_L}{2\sqrt{D} + q'_L - q'_R} \quad (7.94)$$

and set  $\alpha = (1 - \beta)\alpha_L + \beta\alpha_R$ . If  $D < 0$ , either use  $D = 0$ , or revert to a default value of  $\beta$ .

In summary, the choice of step length is crucial for the performance of the algorithm and needs careful treatment. More on this subject can be found in the advanced literature: Sections 3 and 17 of [16], Section 3 of [15], and Section 2.5 of [10]. Figure 7.3b illustrates how the example problem is solved iteratively, starting the iterations from the measured values.

#### 7.5.2.5 The Maratos Effect

A big advantage of Newton-type iteration methods is that once  $\mathbf{x}^{(v)}$  is close enough to the desired solution  $\mathbf{x}^*$ , these methods converge *super-linearly* (for a proof, see Section 5.3 of [12]). This means that for vectors  $\mathbf{x}^{(v)}$  and  $\mathbf{x}^{(v+1)}$  from subsequent iteration steps, the relation

$$\lim_{v \rightarrow \infty} \frac{\|\mathbf{x}^{(v+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(v)} - \mathbf{x}^*\|} = 0 \quad (7.95)$$

holds [15], whereas linear convergence is characterised, for sufficiently large  $v$ , by the much weaker condition

$$\frac{\|\mathbf{x}^{(v+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(v)} - \mathbf{x}^*\|} \leq \mu \quad (7.96)$$

with  $0 \leq \mu < 1$ . In fact, only the Newton method gives super-linear convergence, which is known as the *Dennis–Moré theorem* [22, 23] (see also Section 6.2 of [10] and Section 4.6 of [16]). However, super-linear convergence is only achieved if full steps with  $\alpha = 1$  are taken, otherwise the convergence is only linear and thus much slower.

Maratos discovered [24] that under certain unfavourable conditions the use of merit functions in the step-length determination will give rise to  $\alpha < 0$ , and thus the convergence will be only linear, even very close to the desired solution. The problem occurs if a constraint is fulfilled to a good approximation and the constraint function is curved. Under these conditions, the iteration step, which is tangential to the  $c = 0$  surface, will invariably lead to an increase of the constraint violation  $|c|$ . If this is not balanced by a sufficient decrease of the objective function  $f$ , the merit function  $\phi$  will have a minimum very close to the last iteration point (i.e. at a small value of  $\alpha$ ), and will rise very rapidly beyond that minimum. To achieve better convergence,  $\alpha = 1$  should be used close to the desired solution. However, it is not trivial to identify when close is close enough.

Basically, three approaches exist to overcome the Maratos effect (see [15], Section 15.5):

- Use a merit function that does not suffer from the Maratos effect. Such a merit function has been given by Fletcher [10, 25] and is known as *Fletcher's augmented Lagrangian*.
- The use of second-order corrections [26–29].

- Allowing intermediate steps that increase the merit function [30–32]. This approach is called the *non-monotone strategy* or *watchdog strategy*.

Fletcher's augmented Lagrangian is quite expensive to calculate, and therefore of minor interest.

In the method of second-order corrections, an additional correction step  $\hat{p}$  is calculated with the aim of reducing the constraint violation after taking the full Newton step  $p$  by searching the minimum-norm solution to the equation

$$\mathbf{A}\hat{p} + c(\mathbf{x} + p) = 0. \quad (7.97)$$

This is achieved using the Moore–Penrose pseudo-inverse  $\mathbf{A}^+$  of matrix  $\mathbf{A}$ , which is given by  $\mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$  if  $\mathbf{A}$  has full row rank:

$$\hat{p} = -\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}c(\mathbf{x} + p). \quad (7.98)$$

The calculation of the second-order correction involves the evaluation of the constraints at the point  $c(\mathbf{x} + p)$ , which is necessary in any case for the calculation of the merit function, and the solution of a  $K$ -dimensional system of equations, plus some matrix multiplications. It is thus considerably cheaper than calculating an additional Newton step. Figure 7.3d illustrates the effect of a quadratic correction in the example problem.

A useful strategy is thus to first try the full Newton step, and if that fails, to try the full step with a second-order correction. If the merit function does not decrease sufficiently even with the second-order correction, the step size has to be reduced. Second-order correction steps may improve the merit function also for reduced step sizes, but Nocedal (Section 15.5 of [15]) recommends in this case to proceed directly with the next Newton iteration.

An alternative approach to the method of second-order corrections is to allow a number of relaxed Newton steps, where the merit function is allowed to increase, then do a line search at the last relaxed step, and only do a line search along the first Newton direction if the relaxed steps fail to produce a sufficient decrease of the merit function. This approach, called the *non-monotone strategy* or *watchdog strategy* [15], needs a certain amount of book-keeping in its implementation, but often yields good practical performance.

The Maratos effect is also discussed in [10, 16].

### 7.5.3

#### Detecting Convergence

Determining whether the iterative algorithm has converged to a viable solution of the constrained minimisation problem has two parts:

- A decision when to stop iterating.
- A check whether a true minimum has been found.

To decide when to stop the iterations, two questions need to be answered:

- a) Is it expected that the next steps will change the result significantly?
- b) Is the current result good enough?

#### 7.5.3.1 Stopping the Iterations

The first question ‘Is it expected that the next steps will change the result significantly?’ can be answered by considering whether the length of the last step has changed any of the fitted variables  $\boldsymbol{\theta}, \xi$  by a significant amount. Mathematically speaking, this can be expressed by the condition

$$\|\Delta\boldsymbol{\theta}\| + \|\Delta\xi\| < \epsilon \quad (7.99)$$

with a suitably small value of  $\epsilon$  and any vector norm  $\|\cdot\|$ , where  $\Delta\boldsymbol{\theta}$  and  $\Delta\xi$  are the vectors of the full Newton step calculated in the last iteration. Mainly, the choice is whether to use the *Euclidean norm*  $\|\cdot\|_2$ , which corresponds to demanding that the root mean square (RMS) of the steps in the  $N$  different variables is smaller than  $\epsilon/\sqrt{N}$ , or the *maximum norm*  $\|\cdot\|_\infty$ , which demands that none of the variables has changed by more than  $\epsilon$ . From a physics point of view one needs to remember that a value of  $\epsilon = 10^{-3}$  may mean very different things:  $10^{-3}$  GeV is a very small energy in a problem where 100-GeV jet energies are fitted, while  $10^{-3}$  cm is a quite large value in the context of a vertex-fitting problem. Therefore, it is recommended that all elements of  $\Delta\boldsymbol{\theta}$  and  $\Delta\xi$  need to be scaled with their expected resolutions before the norm is evaluated. In that case, a value of  $\epsilon = 10^{-2}$  means that the variables have changed (either on average or maximally) by about  $0.01\sigma$ , which is certainly reasonable in most high energy physics applications.

Another criterion is the change of the overall  $\chi^2$  in the last step. While at the minimum the change of a variable by  $0.01\sigma$  will lead to a change of  $10^{-4}$  in the  $\chi^2$ , far away from the minimum the same step will lead to a much larger change in  $\chi^2$ . The value of  $\chi^2$  is dimensionless and already includes a scaling with errors. Therefore, a criterion

$$\Delta\chi^2 < \epsilon \quad (7.100)$$

for the change  $\Delta\chi^2$  in the last iteration is reasonable. However, (7.100) does not take into account changes in the unmeasured variables  $\xi$  and thus is alone not a sufficient convergence criterion if the problem contains unmeasured variables.

The second question ‘Is the present result good enough?’ can be reformulated as ‘Are the constraints fulfilled well enough?’, which leads to the condition

$$\|c(\boldsymbol{\theta}, \xi)\| < \epsilon \quad (7.101)$$

on the vector of constraints  $c$ . Again, the same caveats apply as for (7.99), namely that the elements of  $c$  need to be scaled by some expected resolution in order to make the condition physically meaningful.

### 7.5.3.2 Verifying Convergence

After the iterations have been stopped because a suitable set of the stopping conditions discussed in the previous section has been fulfilled, it still has to be verified that the (approximate) solution that has been found corresponds to a minimum rather than a stationary point or even a maximum. This is necessary because the solution determined by the iterations is a solution only to the first-order necessary conditions (7.12), but does not guarantee that *sufficient* conditions are met, in particular (7.17).

To check the second-order sufficient condition (7.17), we first need a basis of the set of feasible directions  $\mathbf{d}$ . These directions are characterised by (7.7), or using the Jacobian matrix  $\mathbf{A}^T$  defined in (7.26), by

$$\mathbf{0} = \mathbf{A}\mathbf{d}. \quad (7.102)$$

Thus, the feasible directions  $\mathbf{d}$  span the null-space of  $\mathbf{A}$ . If we calculate the full *singular value decomposition* (SVD) of  $\mathbf{A}$  (see also Section 6.2.1),

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T, \quad (7.103)$$

then the last column vectors that correspond to singular values  $\sigma_i = 0$  form a basis of the null-space of  $\mathbf{A}$ . The submatrix  $\mathbf{Z}$  formed from these column vectors, has the property  $\mathbf{A}\mathbf{Z} = \mathbf{0}$ . It also has full column rank. If all constraints are linearly independent (LICQ, see Section 7.3.1),  $\mathbf{Z}$  is a  $N \times (N - K)$  matrix with orthonormal column vectors. Therefore, every feasible direction vector  $\mathbf{d}$  can be written as  $\mathbf{d} = \mathbf{Z}\hat{\mathbf{d}}$ , with a coefficient vector given by  $\hat{\mathbf{d}} = \mathbf{Z}^T\mathbf{d}$ . The condition (7.17)  $\mathbf{d}^T\mathbf{L}^*\mathbf{d} > 0$  then becomes

$$\hat{\mathbf{d}}^T \mathbf{Z}^T \mathbf{L}^* \mathbf{Z} \hat{\mathbf{d}} > 0, \quad (7.104)$$

which means that the transformed matrix  $\hat{\mathbf{L}}^* = \mathbf{Z}^T \mathbf{L}^* \mathbf{Z}$  has to be positive definite.

An easy way to check whether  $\hat{\mathbf{L}}^*$  is positive definite is to try to calculate its *Cholesky decomposition* [6, 11]  $\hat{\mathbf{L}}^* = \mathbf{C}\mathbf{C}^T$ , which works only for positive semi-definite matrices, and check that all diagonal entries of the lower triangular matrix  $\mathbf{L}$  are non-zero.

#### 7.5.4

##### Finding Initial Values

In the previous sections, we have discussed an iterative procedure to solve the system of equations derived from the Lagrange multiplier method. As for most iterative methods, the performance of this procedure, in terms of the number of necessary iterations and in terms of the overall fraction of problems where a valid solution is found at all, depends heavily on the choice of the initial values.

For problems without unmeasured quantities the choice of initial values is in fact rather simple: just take the solution of the unconstrained minimisation problem. In particular in those applications where we seek to improve the result of a measurement by imposing constraints between the measured observables, intuition

tells us that the solution to the constrained problem cannot be too far away from the solution to the unconstrained problem, which is simply given by the original measurements. Only if the measurement (e.g. an observed event) does not fit the hypothesis (e.g. the assumption that the event comes from a certain particle reaction) that defines the constraints, will the constrained problem have a solution that is ‘far away’ from the original measurements.

If unmeasured quantities (such as a neutrino’s momentum) are present, the situation becomes much more difficult, because the unconstrained problem tells us nothing about the start values for the unmeasured quantities. Since these are determined by the constraints, it makes sense to try and determine starting values for the  $U$  unmeasured quantities such that  $U$  out of the total  $K$  constraints are fulfilled. But since there are more constraints than unmeasured quantities, there are several possibilities for which constraints to consider, and they will lead to different initial values. Also for a given selection of constraints there may be multiple solutions, or even none.

There is no simple solution to that problem if the constraints are non-linear. In general, it will be necessary to develop a solution for each specific set of constraints and unmeasured quantities, which possibly needs a certain amount of testing to find a method that gives the best starting value for the iterative procedure.

### 7.5.5

#### Error Calculation

After the solution to the minimisation problem has been found, in statistical applications we are often interested in the uncertainties of the fitted parameters. It turns out that the same reasoning that we employed in the analytically solvable problem of Section 7.4.1 applies here as well: the system of equations (7.57) implicitly defines a set of functions that map the measurements  $t$  to the fitted values  $\boldsymbol{\theta}$ ,  $\xi$ . In Section 7.4.1 we actually had an explicit solution for that function at hand, which is no longer true in the case of non-linear constraints. However, an explicit solution is not necessary to evaluate the derivatives that are needed for the error propagation: Equation 7.64 is equivalent to (7.28), and with the same methods as used before, one arrives at the same result (7.54) and (7.55).

## 7.6

### Further Reading and Web Resources

The reader who is interested in delving deeper into the subject of constrained optimisation may refer to the following books:

- The book by Nocedal and Wright [15] gives an excellent overview over numerical optimisation methods; it also covers inequality-constrained problems and other approaches to search for the minimum, in particular interior-point and trust-region methods.

- The book by Bonnans, Gilbert, Lemaréchal and Sagastizábal [16] is also an excellent text.
- The book by Fletcher [10] is significantly older, but also interesting.
- A classic text on numerical methods is the book by Stoer and Bulirsch [12], which includes a discussion of the Newton method in several dimensions.

Other useful material can be found at:

- Paul Avery's website [33] which contains several excellent write-ups on kinematic fitting of tracks and particle decay chains, and his `KWFIT` library.
- Volker Blobel has written the software package `APLCON` [34] for constrained fit problems.

## 7.7

### Exercises

#### Exercise 7.1 Particle decay

For the particle-decay example in Section 7.3.1, write down the explicit set of three equations given by  $\mathbf{0} = \nabla f(E_1^*, E_2^*) + \lambda^* \nabla c(E_1^*, E_2^*)$ .

- From these equations, derive the system of equations for one iteration of the Newton–Raphson method.
- Write a program (or spreadsheet) that calculates for a given set of parameters  $E_1$ ,  $E_2$  and  $\lambda$  the values after one step and after  $n$  steps of the Newton iteration, where  $n$  is a number between 1 and 10.
- Try varying the start values of  $E_1$  and  $E_2$  (set the initial value of  $\lambda$  to 0) and check whether one, two, or five Newton steps increase or decrease the distance  $d = \sqrt{(E_1 - E_1^*)^2 + (E_2 - E_2^*)^2}$  to the desired solution, where  $E_1^*$  and  $E_2^*$  are the values of the solution, as given in the example.
- Try to apply step-size control to your program.
- Try to understand for which initial values of  $E_1$ ,  $E_2$  your program fails, and why, and try to improve the algorithm.

#### Exercise 7.2 W decay

Equations (7.20) to (7.22) of the W decay example in Section 7.3.2 constitute three equations for the three unmeasured quantities  $E_{T,\nu}$ ,  $\phi_\nu$  and  $\eta_\nu$ .

- Derive the explicit solution of this system of equations.
- Is the solution unique and guaranteed to exist for  $E_{T,\nu}$ ,  $\phi_\nu$  and  $\eta_\nu$ ?
- This kind of problem occurs typically within a larger fit problem (for instance in an event where a  $t\bar{t}$  pair decays to four jets, an electron, and a neutrino). Using some of the constraints to calculate starting values for the unknown parameters

is often a useful strategy. How would you proceed if for some of the parameters there are no, or more than one solutions?

### Exercise 7.3 Linear fit

In the spirit of the averaging example in Section 7.4.1, formulate a linear fit as a constrained fit problem. Start with the equation  $y = a \cdot x + b$ , where  $a$  and  $b$  are the two unknown parameters of the straight line fit. Assume that the true values  $\hat{y}_i$  at the given values of  $x_i$  must lie on the straight line, while the measurements  $y_i^{\text{meas}}$  scatter around the true values.

- Write down the objective function (using least squares) and the constraints of the problem.
- Write down the system of equations that arise from the Lagrange multiplier method for this problem.
- Look up the standard solution to this problem and show that it solves the Lagrange multiplier formulation of this problem (and calculate the  $\lambda$  values).
- Consider how the problem changes if one wants to use a different penalty function for the distance between measured and true values of  $y_i$ , if one wants to admit also uncertainties of the abscissa values  $x_i$ , or if one wants to use a more complicated fit function than a straight line. Would that be easy or difficult to implement in a framework that solves constrained fit problems?

### References

- Lutz, G. (1993) Topological vertex search in collider experiments. *Nucl. Instrum. Methods A*, **337**, 66.
- Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London A*, **222**, 309.
- Fisher, R.A. (1925) Theory of statistical estimation. *Math. Proc. Camb. Philos. Soc.*, **22**, 700.
- James, F. (2006) *Statistical Methods in Experimental Physics*, 2nd edn, World Scientific.
- Stuart, A. and Ord, J.K. (1994) Distribution theory, in *Kendall's Advanced Theory of Statistics*, vol. 1, 6th edn, John Wiley & Sons.
- Gentle, J.E. (2007) *Matrix Algebra*, Springer Texts in Statistics, Springer.
- Lagrange, J. (1788) *Méchanique Analytique*, Desaint.
- Fraser, C. (1992) Isoperimetric problems in the variational calculus of Euler and Lagrange. *Hist. Math.*, **19**, 4.
- Kuhn, H.W. and Tucker, A.W. (1951) Nonlinear programming, in *Proc. Second Berkeley Symp. Math. Stat. Prob.*, 1950, University of California Press, p. 481.
- Fletcher, R. (1987) *Practical Methods of Optimization*, 2nd edn, John Wiley & Sons.
- Golub, G.H. and Van Loan, C.F. (1996) *Matrix Computations*, 3rd edn, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press.
- Stoer, S. and Bulirsch, R. (2010) *Introduction to Numerical Analysis*, Texts in Applied Mathematics, vol. 12, 3rd edn, Springer.

- 13 Moore, E.H. (1920) On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, **26**, 394.
- 14 Penrose, R. (1955) A generalized inverse for matrices. *Proc. Camb. Philos. Soc.*, **51**, 406.
- 15 Nocedal, J. and Wright, S.J. (2006) *Numerical Optimization*, 2nd edn, Springer Series in Operations Research and Financial Engineering, Springer.
- 16 Bonnans, J.F. et al. (2006) *Numerical Optimization, Theoretical and Practical Aspects*, 2nd edn, Universitext, Springer.
- 17 Armijo, L. (1966) Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math.*, **16**, 1.
- 18 Wolfe, P. (1969) Convergence conditions for ascent methods. *SIAM Rev.*, **11**, 226.
- 19 Powell, M.J.D. (1976) Some global convergence properties of a variable metric algorithm for minimization without exact line searches, in *Nonlinear Programming (Proc. Sympos., New York, 1975)*, vol. IX, Am. Math. Soc., New York, 1976, pp. 53, SIAM-AMS Proc.
- 20 Goldstein, A.A. (1965) On steepest descent. *J. Soc. Ind. Appl. Math. Ser. A Control*, **3**, 147.
- 21 Goldstein, A.A. and Price, J.F. (1967) An effective algorithm for minimization. *Num. Math.*, **10**, 184.
- 22 Broyden, C.G., Dennis, Jr., J.E., and Moré, J.J. (1973) On the local and superlinear convergence of quasi-Newton methods. *J. Inst. Math. Appl.*, **12**, 223.
- 23 Dennis, Jr., J.E. and Moré, J.J. (1974) A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comput.*, **28**, 549.
- 24 Maratos, N. (1978) Exact penalty function algorithms for finite dimension-  
al and control optimization problems, Ph.D. thesis, Department of Control Theory, Imperial College London.
- 25 Fletcher, R. (1973) An exact penalty function for nonlinear programming with inequalities. *Math. Program.*, **5**, 129.
- 26 Coleman, T.F. and Conn, A.R. (1982) Nonlinear programming via an exact penalty function: asymptotic analysis. *Math. Program.*, **24** (2), 123.
- 27 Fletcher, R. (1982) Second order corrections for nondifferentiable optimization, in *Numerical analysis (Dundee, 1981)*, Lecture Notes in Math., vol. 912 (ed. G.A. Watson), Springer, p. 85.
- 28 Gabay, D. (1982) Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization. *Math. Program. Stud.*, **16**, 18.
- 29 Mayne, D.Q. and Polak, E. (1982) A superlinearly convergent algorithm for constrained optimization problems. *Math. Program. Stud.*, **16**, 45.
- 30 Chamberlain, R.M. et al. (1982) The watchdog technique for forcing convergence in algorithms for constrained optimization. *Math. Program. Stud.*, **16**, 1.
- 31 Grippo, L., Lampariello, F., and Lucidi, S. (1986) A nonmonotone line search technique for Newton's method. *SIAM J. Num. Anal.*, **23** (4), 707.
- 32 Conn, A.R., Gould, N.I.M., and Toint, P.L. (2000) *Trust Region Methods*, MPS-SIAM Series on Optimization, SIAM.
- 33 Avery, P. (1999) Fitting Theory Write-ups and References. [www.phys.ufl.edu/~avery/fitting.html](http://www.phys.ufl.edu/~avery/fitting.html) (last accessed 2013).
- 34 Blobel, V. (2010) Constrained Least Squares. [www.desy.de/~blobel/wwwcondl.html](http://www.desy.de/~blobel/wwwcondl.html) (last accessed 2013).

## 8

### How to Deal with Systematic Uncertainties

Rainer Wanke

#### 8.1 Introduction

Every physicist learned about systematic uncertainties<sup>1)</sup> in the university lab courses. Typical examples of such systematics include the scale-reading accuracies of rulers, voltmeters, or something similar. These systematic uncertainties enter the computation, undergo the laws of error propagation, and are finally quoted as a second error after the statistical error (if the experiment had a statistical error at all).

However, somehow these ‘systematic errors’ appear from nowhere, and many questions remain, most of which have not been answered in a satisfactory way: What is the correct size of the systematic error? Were all the systematic uncertainties considered or were some of them forgotten? Are there correlations between these systematics, and if yes, how are they taken into account? Such questions usually result in a general uneasiness about systematics, and many students just write down what they are told – without a real understanding of the issue.

This uneasiness about systematic uncertainties often continues when doing particle physics analyses. These analyses are usually very complex, involving complicated fits or other statistical subtleties. Finally, however, the analysis is hopefully finished with a result including a statistical error. Only the systematic uncertainties still need to be evaluated, a task which is usually shifted to the very end of an analysis. At this stage people often discover that it is not simple to estimate a systematic uncertainty. Moreover, unlike in the case of statistical uncertainties, there seem to be no clear recipes for the detection and estimation of systematics uncertainties. What can be done?

Indeed there are, in general, no clear recipes for the determination of systematic uncertainties. Most estimations of systematic errors are the result of a mixture of

1) We will use the terms ‘systematic uncertainties’, ‘systematic errors’, or just ‘systematics’ as synonyms throughout this chapter. Being pedantic, only ‘systematic uncertainty’ would be correct, but since all these terms are likewise used in real life we will keep it like this. ‘Systematic effects’ and ‘systematic problems’, however, are not systematic uncertainties, but the causes which lead to them.

knowledge, experience, common sense, and sometimes intuition; there are no formulae to follow. Therefore, this chapter – rather than being stuffed with advanced mathematics – tries to offer general advice and to discuss a few ‘do’s and don’ts’ of systematic errors, together with several examples from particle physics analyses. It aims at giving a general feeling on how to detect, to estimate, and possibly to avoid systematic uncertainties.

## 8.2

### What Are Systematic Uncertainties?

Often, systematic uncertainties are neither very clearly defined nor well separated from statistical uncertainties. Sometimes the two uncertainties even become mixed – for example in trigger efficiencies, which may be partially determined from data statistics. Moreover, in many analyses, so-called ‘external errors’ from external input values, ‘theory errors’ from theoretical input or other error sources are separated from the ‘experimental’ systematic uncertainty. In many cases, therefore, results are quoted with three, four, or even more uncertainties attached. While this may emphasise the impact of different sources on the total uncertainty, a long chain of errors does not enhance the readability of a result. Different contributions to the systematics are therefore more clearly presented in a dedicated table.

A standard definition of systematic uncertainties is the following:

“Systematic uncertainties are all uncertainties that are not directly due to the statistics of the data.”

With this definition, also statistical uncertainties of trigger efficiencies, measured from data, and detector acceptances, determined from Monte Carlo (MC) simulation, are considered as systematic errors. This may seem strange (and indeed people often do include these effects into the statistical error), but it appears justified when considering that these uncertainties may still be reduced after the data-taking by further Monte Carlo production or by smarter methods of determining a trigger efficiency.<sup>2)</sup>

In this chapter, however, we will use a pragmatic definition of systematic uncertainties, which better fits the purpose of this chapter:

“Systematic uncertainties are measurement errors which are not due to statistical fluctuations in real or simulated data samples.”

This means – as an example – that here we treat neither trigger efficiency errors, determined from data statistics, nor detector acceptance errors, determined from Monte Carlo statistics, as systematic uncertainties – simply because they only underlie the laws of statistics and can fully be treated as statistical uncertainties. In

2) Of course, one could also reduce the statistical error of the data afterwards by for example loosening selection criteria. The breakdown into statistical and systematic errors therefore is always somewhat arbitrary and a matter of personal taste.

the final result, these errors may then either be given separately or be added to the systematic uncertainty, as explained above.

With this definition, we can write down a list of typical sources of systematic errors and biases in high energy physics that should always be borne in mind by a data analyst:

- badly known detector acceptances or trigger efficiencies;
- incorrect detector calibrations;
- badly known detector resolutions;
- badly known background;
- uncertainties in the simulation or underlying theoretical models;
- uncertainties on input parameters like cross sections, branching fractions, lifetimes, the luminosity, and so on (often called ‘external uncertainties’);
- computational and software errors;
- personal biases towards a specific outcome of an analysis;
- other usually unknown effects on the measurement.

Obviously, in particular the last three sources of systematics are difficult to assess, but also the other items on the list are sometimes difficult to find and to estimate. In the next Section 8.3 we will therefore try to provide strategies to detect unknown systematics. These can then be estimated using various methods, some of which are introduced in Section 8.4. The best strategy, of course, is to try to avoid systematics from the very beginning of an analysis, and some appropriate methods are presented at the end of this chapter in Section 8.5.

### 8.3

#### Detection of Possible Systematic Uncertainties

The first step towards correctly estimated systematic uncertainties is of course their detection. There are basically two approaches. The first method is to seriously ponder about all potential sources of systematics and their possible impact on the analysis. Since this deductive approach needs an overview of the complete analysis in order to detect potential problems in the details, we will refer to it as the ‘top-down approach’ in the following.

The other method works the other way around and is thus named the ‘bottom-up approach’ below: potential systematic problems are detected by performing cross-checks, for example comparing data and Monte Carlo distributions, to look for inconsistencies in the analysis and to conclude from them the source of the problem.

##### 8.3.1

###### Top-Down Approach

The main point for this approach is to think about all possible sources of potential systematics, that is to think of virtually everything that could have gone wrong.

The procedure is somewhat similar to a detective investigating a murder: which persons (i.e. quantities, effects and methods) had a reason and the means to carry out the murder (i.e. influence the result)? Those then become the prime suspects and are accurately investigated in the following.

It may seem quite impossible to think of every possible systematic effect, and it probably is in most cases. However, even if the complete task should be unfeasible, the list given in the previous section is certainly a good starting point. Probably you can also think of several additional potential sources of systematics in your particular analysis.

For the remaining potential problems it is important to open your mind: look around and try to find out what problems were encountered by other people doing similar analyses. And: Talk to other people! Explain your analysis to as many different people as possible. This means of course to give presentations in your group, but also to explain your analysis in great detail to other members of your group or even to outsiders. The point is that you have to reach a sort of ‘meta-level’ and view your analysis from far above. In the everyday routine of an analysis, fighting with huge data samples, large MC productions and the usual computer problems, you can very quickly lose sight of the analysis in its entirety. Most of the time, you do not think about the actual physics involved. Therefore, every now and then, you should step back from the daily routine and try to take a more general view on your work. This is most easily achieved by explaining your work to other people. By doing so, you might discover many potential problems – but you will also be led to solutions and new ideas.

### 8.3.2

#### **Bottom-Up Approach**

The bottom-up approach is used in addition to the top-down approach in order to detect systematic errors that were not considered in the previous step. The idea is to scrutinously check your analysis for internal consistency. For this, the main method is usually to compare data and Monte Carlo-simulated distributions (as many as possible) and to look for significant differences.

In addition to data-to-MC comparisons, one can also divide the available dataset into several subsamples with different data-taking conditions (magnet polarities, instantaneous luminosities, detector configurations, etc.) or from different time periods. Are the partial results for all of these subsamples consistent with each other?

Finally, very often variations of selection criteria (‘cut variations’) are performed. This method usually involves the least work: instead of producing a lot of plots to compare data distributions to simulation, selection criteria can easily be varied and may also show inconsistencies in the analysis.

Continuing the ‘detective theme’ from the last section: the bottom-up approach is similar to convicting the culprit by just analysing the clues left at the crime scene. Every single detail may be important. It is therefore absolutely necessary to look at as many potential clues as possible and to consider everything which looks unusual

such as, for example, cigarette stubs in a non-smoker household or different energy spectra for positively and negatively charged pions.

### 8.3.3

#### Examples for Detecting Systematics

In the following we will have a look at a few examples of how to find potential systematics. While the first two use the top-down approach, the others should give ideas on how to look for potential analysis problems.

##### 8.3.3.1 Background Systematics

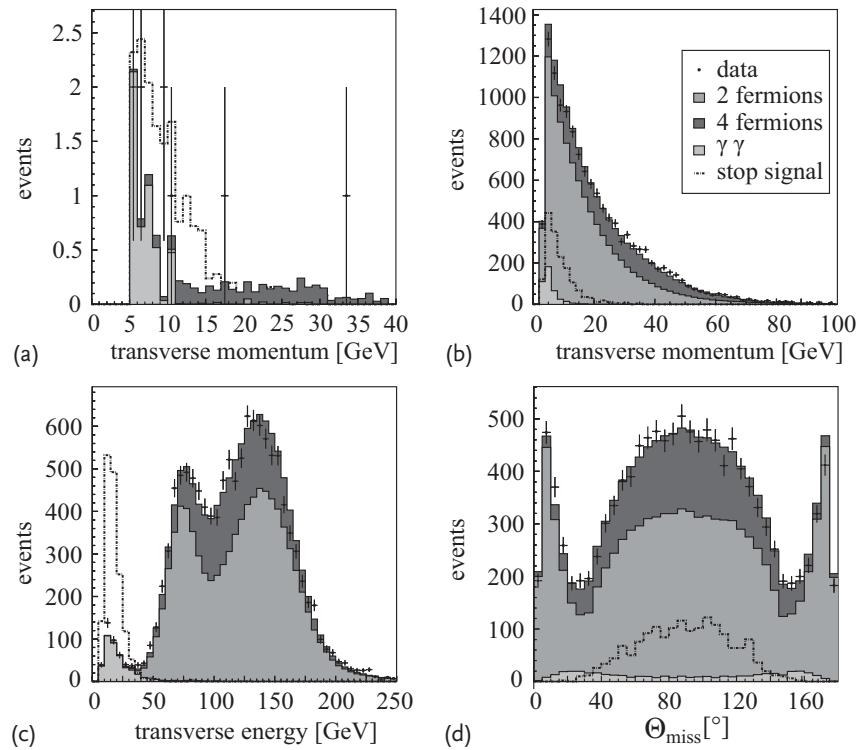
Especially in searches, the estimation of possible backgrounds to a signal may be difficult. Sometimes only a few events are left in the dataset after applying all selection criteria.

##### Example 8.1 A search for stop-quark pairs with DELPHI

An example of background systematics is shown in Figure 8.1: shown are distributions from a search for supersymmetric stop-quarks  $\tilde{t}$  by the DELPHI collaboration [1]. The signature of a possible  $e^+ e^- \rightarrow \tilde{t}\tilde{t}$  pair production with following decays  $\tilde{t} \rightarrow c\tilde{\chi}_1^0$  are two acoplanar jets and missing energy in the detector. Possible Standard Model backgrounds are 2-jet, 4-jet, and two-photon events. One of the variables most sensitive to a stop-quark signature is the total transverse momentum of the event, which is shown together with the background expectation from MC simulations in Figure 8.1a after applying the final set of selection cuts.

How can one make sure that the background is correctly estimated? Indeed, this is a difficult and non-trivial task: in order to reduce the background as much as possible, only the very tails of the background remain, which leak into the signal region. These are, however, not necessarily well described by the simulation since, for instance, resolution effects may be enhanced. One possibility for detecting possible systematics – used by the DELPHI collaboration – is to enhance the background by releasing some of the selection criteria and to compare the larger, background-dominated sample with the simulation as shown in Figure 8.1b–d. Two things are important for this: first, the dependencies of the background on the released selection criteria need to be well understood – otherwise the comparison between the data and the background-enhanced sample would be meaningless. Second, as always, as many distributions as possible should be compared to look for unexpected behaviour of the data with respect to the simulation. Looking again at the DELPHI analysis, it can be seen that the simulated distributions of the transverse momentum and of the polar angle of the missing momentum agree well with the data (Figure 8.1b,d), while in the lower range of the transverse energy (Figure 8.1c) the data disagree by about 15% from the simulation. Here the usefulness of looking at several distributions becomes obvious: while no discrepancy can be observed in the distribution of transverse momenta (b), there may be a problem

with the  $\gamma\gamma$  background (shown as the lightest histogram) – the one remaining underneath the potential stop signal in the final analysis (a)! Starting from this observation, one can now investigate the reason for the disagreement of this specific background contribution and, as the DELPHI collaboration did, attach a systematic uncertainty to the result of the search (see Section 8.4.2.1).

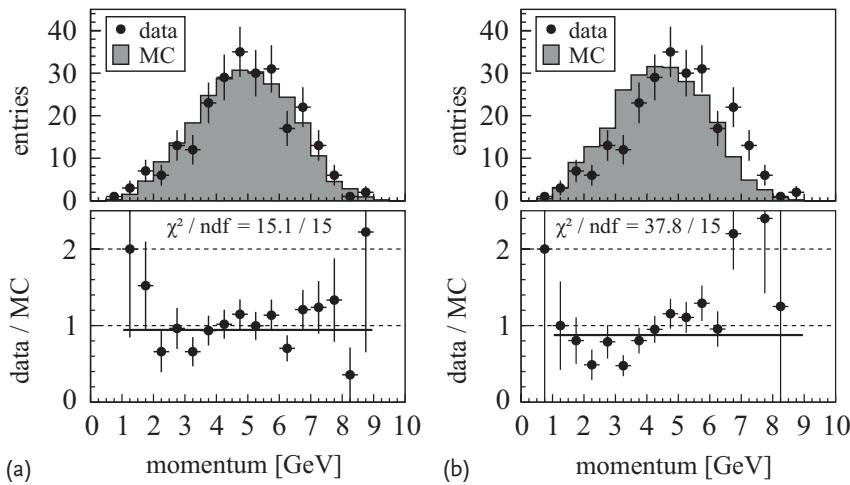


**Figure 8.1** Stop-quark search by the DELPHI collaboration at LEP. (a) Total transverse momentum with all selection criteria applied. (b-d) Distributions of the total transverse momentum, the total transverse energy and the angle of the missing momentum with only

loose cuts against background. The data are shown as points with error bars. The expected background contributions from different sources and the potential stop signal are also shown as shaded and dashed histograms, respectively. (Adapted from [1]).

### 8.3.3.2 Detector Acceptances

Systematics due to detector acceptance are usually a result of a poor Monte Carlo description of the detector, in particular of detector inefficiencies or misalignments. In most cases such error sources are not precisely known (otherwise one could correct for them). If there are no strong hints towards a specific problem, there is essentially only the bottom-up approach left to discover them. This means: compare data and MC distributions for as many variables as possible! For the comparison, both data and simulation, normalised to the data, are plotted into



**Figure 8.2** Comparison of hypothetical momentum distributions in data and simulation. (a) Good description of the data by the simulation. (b) Bad data description.

the same histogram as in the example shown in Figure 8.2a. It is important that all selection criteria except the one on the plotted variable are applied. Otherwise, the understanding of the distribution is limited. Also, one should try to avoid logarithmic y-axis scales, or at least look at both linear and logarithmic plots.

As in the upper plot of Figure 8.2b, possible problems can often immediately be seen. The next step is then to divide real data by the normalised Monte Carlo data. Also here, deviations are usually immediately obvious. In addition, a  $\chi^2$  test (see Section 3.8.1) can be performed by fitting a constant line. In the example of Figure 8.2a data and simulation agree well with  $\chi^2/\text{ndf} = 15.1/15$ , corresponding to a probability of 44.4% to have a  $\chi^2 \geq 15.1$  when using data from the simulation, while the example in Figure 8.2b has  $\chi^2/\text{ndf} = 37.8/15$ , corresponding to a probability of 0.1%.<sup>3)</sup> Still, to look at the divided distributions alone, is *not* enough in order to really understand the systematic problem. While the lower plot of Figure 8.2b shows some problematic behaviour, only the physics distributions in the top part of the figure can reveal the source of the problem: in our example, data and MC distributions seem to be shifted with respect to each other, possibly pointing to a false momentum scale applied in the simulation.

### 8.3.3.3 Splitting Data into Independent Subsets

An important cross-check of the internal consistency of a dataset is to split it into several independent subsamples of more or less similar size, and to compute

3) Very often just the reduced  $\chi^2$  ( $15.1/15 = 0.98$  and  $37.8/15 = 2.38$  in our example) is quoted and required to be close to 1. This is *not* sufficient, since the variance of the  $\chi^2$  distribution strongly depends on the number of degrees of freedom  $\text{ndf}$ ! Therefore one should always give both  $\chi^2$  and  $\text{ndf}$  and at best also the probability  $P(\chi^2, \text{ndf})$  to obtain a  $\chi^2$  value equal or larger than the observed  $\chi^2$  for a simulation with a perfect description of the data.

a separate result for each subsample. All these results should agree within the statistics (to be determined for example by a  $\chi^2$  test).

One possibility are subsamples that represent different experimental conditions, such as for instance different running periods. Typical examples are single LHC years, the Tevatron Runs I, IIa, and IIb, but also different trigger settings or magnet polarities. The result should of course not significantly depend on the run period or other experimental conditions. If for one subsample a significant deviation from the mean is observed, it is necessary to look for possible reasons for different conditions in this particular period or with this particular setting. (A look into the experiment log-book or a chat with other collaboration members may be useful!) However, be careful when excluding data samples without a very obvious reason. You may very easily bias yourself, as is described in Section 8.5.2. If there is no obvious reason and the significance of the deviation is not too large, it is better to keep the suspicious subset. Conversely, if the deviation is absolutely significant, the source should be found and the effect be cured before publishing a result: if a large unknown effect appears in one subset, it may also be present in other subsets, it may just not be as visible.

#### 8.3.3.4 Evaluating the Result in Intervals of an Analysis Parameter

A second possibility of dividing the data into independent subsets is the splitting into bins of an analysis parameter such as, for example, a track or jet momentum, missing energy, or other quantities crucial for the analysis. In general, the same considerations for possible deviations in single bins hold here as have been described in the previous paragraph. However, there are two additional possible features: first, the result may show a continuous trend as a function of the investigated analysis parameter. In this case, the source of this trend needs to be investigated, or at least a corresponding systematic uncertainty has to be evaluated (see for example the treatment of data-to-MC discrepancies in Section 8.4.2).

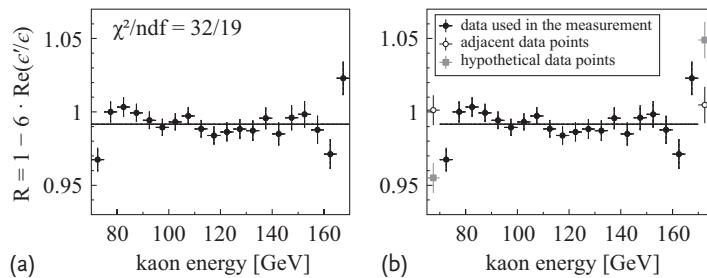
The second case is a deviating result at the very border of the selected region.

#### Example 8.2 CP violation in kaon decays

An example is the measurement of the parameter  $\text{Re}(\epsilon'/\epsilon)$  of direct CP violation in kaon decays by the NA48 experiment [2]. The measurement is done in independent bins of kaon momentum, as shown in Figure 8.3a. Clearly, the two outermost data points do not agree well with the other data, with about 2–3 standard deviations significance. Is this a reason to worry? The overall  $\chi^2/\text{ndf}$  is 32/19, which is not perfect but also not too worrisome (the probability of obtaining a  $\chi^2/\text{ndf}$  equal or worse than 32/19 is 3.1%). First of all, when observing such a behaviour, one should – as always when observing an unexpected feature – think of possible causes. Perhaps there are problems with too high and too low energies. This could be checked with independent investigations, at best with an independent dataset. In this example, many investigations were done, but the collaboration did not find any possible cause for this specific problem. Still, something could have been overlooked, and the problem of dealing with the edges of the energy region remains.

The solution is simple: if possible one should extend the energy region by one or two extra bins and look for a possible trend. The open circles in Figure 8.3b show the enlargement of the energy region by  $\pm 5$  GeV: the additional data points obviously do not follow the two outliers. It was therefore decided that the outliers are just statistical fluctuations, and the original result was published (of course *without* using the two additional bins! Everything else would introduce a bias towards data, which ‘look good’ (see also Section 8.5.2.2)).

A different situation would have occurred if the additional energy bins would have looked like the hypothetical data points indicated as small squares in Figure 8.3b: in this hypothetical case, there would have been an obvious problem in the analysis, which would have needed to be solved before a publication. In particular, it would not be sufficient to shorten the energy range without a real understanding of the cause.



**Figure 8.3** Measurement of  $R = 1 - 6 \cdot \text{Re}(\epsilon'/\epsilon)$  by the NA48 collaboration in bins of the kaon energy. (a) Original measurement. (b) Measurement with an extended energy region and hypothetical data points. (Adapted from [2].)

### 8.3.3.5 Analysis Software and Fit Routines

Most data analyses involve complex software and fit routines which may cause many problems, starting with a bias in the fit method itself and ranging to simple software errors. How can it be checked that a piece of software or a fit routine actually does what it should?

First of all, one should always perform a closure test by running the program on a Monte Carlo sample with known input parameters. The MC sample should be as large as possible, at best a factor ten or more bigger than the data sample. If this should not be possible with a full MC sample, it can often be easily achieved by using a toy Monte Carlo sample (see also Section 10.5.1). With this trivial check one should not only look for the correct output value, but also for correlations between fit parameters, correct error scaling with statistics, and so on.

Secondly, a fit routine itself may introduce a bias<sup>4)</sup> (see Section 2.2). Such a bias will become smaller with increasing statistics. Consequently, a test with large MC statistics is not sensitive to it. Instead, a large number ( $\sim 20$ –100) of different MC

4) Maximum-likelihood estimators are in general biased (see Section 2.2). However, the bias usually is very small.

samples each having about the data sample size should be evaluated: do the results follow a (Gaussian) distribution around the input value with a variance as expected from the statistical error?

Another useful cross-check is to re-run the analysis with a different histogram binning. In particular, problems with low statistics in some bins may show up this way. However, this should usually only be used as a cross-check (and as a possible hint to an underlying problem), as it is not straightforward to quantify the statistical fluctuations of the result caused by a different binning.

Finally, to find hidden software errors with no second, independent analysis being available, there is essentially only one advice: look at as many data and simulated distributions as possible. It is very important to check and understand every single feature of every distribution. There is no such thing as looking at too many plots! This work may be very tedious and therefore is often neglected, but it is practically the only way to find hidden errors. Only if really everything is consistent, one can be reasonably sure that the analysis does not contain serious errors. As generally stated already in Section 8.3.1, it is particularly helpful to communicate with other people about the analysis. In many cases potential problems are overlooked simply because they ‘have always been there’, did not seem to be important and eventually became forgotten. Colleagues with fresh eyes have a more unbiased view on the analysis and may be able to find many hidden troubles.

#### 8.4

##### **Estimation of Systematic Uncertainties**

There are, unfortunately, usually no cookbook recipes for the evaluation of systematics. Nevertheless, there do exist many ways to perform solid estimations of systematic uncertainties. This section, after discussing some rather trivial cases, aims at providing general methods and techniques for systematic error estimation using concrete examples. They should by no means been seen as the last word on systematic evaluation but rather as general ideas which may be adapted or at least give hints for data analyses in real life.

A common feature of the given examples and of cases that you may encounter in your own analysis is that you normally have to be creative. Very often there is no standard procedure to determine the systematics. Instead, you need to figure out how a specific source of systematic uncertainty may affect the result. You then try to enhance (or diminish) this explicit source and look at the effect on the result. Or you may have a look for larger effects in other channels or analyses, which are more sensitive to this systematic effect than yours. Finally, you may think of a selection criterion which is most sensitive to the suspected systematic effect and vary this criterion within a large range.

A simple example would be a possible error in the normalisation of the background that has to be subtracted from the data in order to obtain the signal yield. To access the systematic uncertainty of a result due to the background normalisation, one can artificially vary the normalisation within a ‘reasonable’ range and take

the difference to the original result as systematic uncertainty. The art, of course, is to define a ‘reasonable’ range. Sometimes it is enough to take relatively extreme values for the variation (see for example Section 8.4.2.1), but if the systematic effect turns out to be dominating, one also may have to perform detailed studies or even an additional dedicated analysis.

#### 8.4.1

##### Some Simple Cases

###### 8.4.1.1 External Input Parameters

A common case for an external uncertainty is that of an input parameter  $x$  with a known uncertainty  $\sigma_x$ . Examples of such input parameters are branching fractions, lifetimes, luminosities, detector energy scales, and so on. To estimate the resulting systematic uncertainty, the input parameter  $x$  is varied by  $\pm\sigma_x$ , and the deviation from the original result is taken as systematic uncertainty  $\pm\sigma_{\text{syst}}$ . If there is a known functional dependency of the result on the input parameter, the systematic can also be determined by simple error propagation.

###### 8.4.1.2 Tolerances

Sometimes the standard deviation  $\sigma_x$  of a parameter  $x$  is not known and instead a tolerance is given, that is the largest and smallest possible value of  $x$ . An example are hits in a scintillator strip: it is only known that the hits happen with equal probability anywhere inside the strip, but not outside. For a uniform probability distribution in the interval  $[x_{\text{low}}, x_{\text{high}}]$ , the standard deviation is given by  $\sigma_x = 1/\sqrt{12}(x_{\text{high}} - x_{\text{low}}) \approx 0.29 \cdot (x_{\text{high}} - x_{\text{low}})$ , that is a gain of about 40% with respect to the naive estimate of  $\sigma_x = \frac{1}{2}(x_{\text{high}} - x_{\text{low}})$ . Still, many people take the latter, ‘naive’ estimation in such cases, calling it ‘conservative’, even though they could do better without underestimating the corresponding systematic effect.

###### 8.4.1.3 Small Systematics

Another simple case is the estimation of a systematic uncertainty which is – even in the worst possible case – expected to be much smaller than other statistical or systematic uncertainties. In such a case, it is neither necessary nor useful to waste time to have a best possible estimate. Just think of the worst possible case, take it as a ‘conservative estimate’ and spend your time on the more important uncertainties.

#### 8.4.2

##### Educated Guesses

One of the most important methods to estimate systematic uncertainties is the so-called *educated guess*. It is used when no straightforward procedure of estimation is available. In such a case, one has to find a reasonable estimate of the uncertainty from more general considerations. Unfortunately, there is no general recipe on how to perform an educated guess. What is needed is a more general view on

the investigated systematic, and a mixture of knowledge, experience, reasoning, creativity, and common sense. We will therefore have a look at several examples in order to illustrate the idea and show a few possibilities. Nevertheless, for any given problem, you will usually have to find your own educated guess.

A very good introductory example is given in the text book of Roger Barlow [3], where the activity of a specific radioactive source is estimated from very general considerations on the age of the source. Other examples, somewhat closer to high energy physics analyses, are the following.

#### 8.4.2.1 Background Estimation

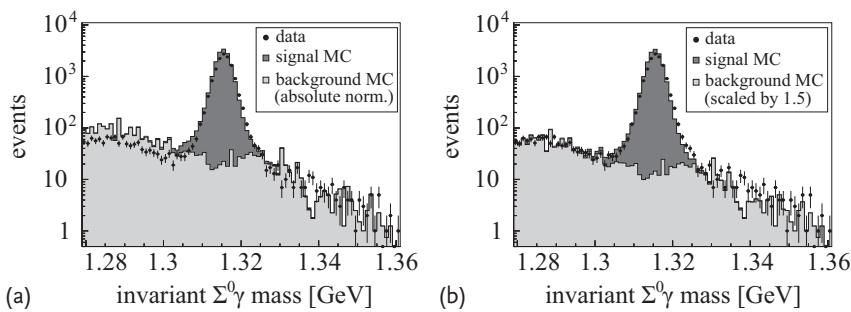
There are two general methods to estimate the background underneath a signal distribution: data can be extrapolated from a signal-free region into the signal region (in many cases by a simple *sideband subtraction*), or one can use a simulation. The first method has the advantage of not relying on potentially incorrect simulations, but it usually does not predict the shape of the background underneath a signal distribution. In the case of only combinatorial background from accidental coincidences this normally is a minor issue because the signal sidebands are fitted with relatively simple empiric functions (e.g. second-order or third-order polynomials) and the difference between fits with different functions is assigned as systematic uncertainty on the background estimation. However, for more complicated background or signal distributions this method is not easily applicable. Conversely, simulations do normally provide the shape of a background distribution but not the absolute normalisation. In practice, both methods often work together, for example, to obtain the background shape from simulation and the proper normalisation from the sidebands.

An example of the procedure discussed in the following is taken from an analysis of  $\Xi^0 \rightarrow \Sigma^0\gamma$  decays by the NA48/1 collaboration [4]. For the discussion here, the peculiarities of this decay are not important, but only the method of the background subtraction and the assignment of its systematic uncertainty.

#### Example 8.3 Measurement of $\Xi^0 \rightarrow \Sigma^0\gamma$ decays

Figure 8.4a shows the data points together with the fitted  $\Xi^0 \rightarrow \Sigma^0\gamma$  signal and the simulated background (normalised to the absolute  $\Xi^0$  flux). At low invariant masses, the background is obviously overestimated by about a factor of two (note the logarithmic scale). Such a mis-estimation may indeed happen, since a simulation does not necessarily model the background processes correctly nor does it always reproduce the tails of the detector resolution well. Using this method to determine the background yields an estimate of 3% background events in the signal region between 1.309 and 1.321 GeV.

The other option for normalising the background is using sidebands, where for our example a suitable choice is the region between 1.28 GeV and 1.30 GeV (Figure 8.4b). This results in a background estimation of 1.5% underneath the signal peak, a factor of two smaller than the first estimate. One may think that the latter method is much closer to the truth since the low-mass region is significantly better



**Figure 8.4** Example of systematic uncertainties due to background subtraction in the decay  $\Xi^0 \rightarrow \Sigma^0\gamma$ . (a) Simulated background normalised to the absolute flux. (b) Normalisation to the left sideband. (Adapted from [4].)

reproduced by the renormalised simulation. However, there may be many different aspects of the background behaviour playing together. Even if the resolution of the bulk of the background may be wrongly simulated, the tails of the resolution – only those contribute to the background – may still be correctly simulated. Also other backgrounds may play a role, as may be indicated by the high-mass region which is not well described by the sideband-subtracted background estimation. In this measurement, the collaboration decided to trust the relative normalisation from the sideband, but assigned a conservatively large systematic error of  $\pm 100\%$  on the estimation, which included the first method and resulted in a background estimation of  $(1.5 \pm 1.5)\%$ . As other systematics were dominating, this uncertainty did not significantly contribute to the total error. Therefore, no further work to reduce the error was performed. This is a good example of an ‘educated guess’: two extreme cases (no or double background) are taken and applied as limits of the systematic uncertainty.

#### 8.4.2.2 Detector Resolutions

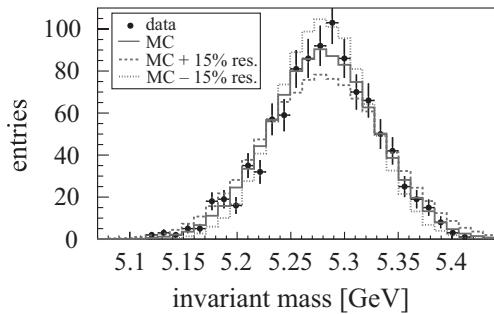
In general, simulations are not very good at reproducing detector resolutions. In particular, vertex positions, invariant masses, shower widths,  $\chi^2$  values of track fits, and so on are often not well reproduced by the detector simulation.

In such cases a simple procedure exists for estimating the corresponding uncertainties. At first, reasonable maximum and minimum mis-estimations of the resolution of a measured quantity  $x$  have to be found. They result in factors  $(1 + k_+)$  and  $(1 - k_-)$ , respectively, with which the resolution has to be multiplied.<sup>5)</sup> Then, since the true values  $x_{\text{true}}$  are known in the simulation, the reconstructed values  $x_{\text{reco}}$  can simply be modified to

$$x_{\text{reco}} \rightarrow x_{\text{reco}} \pm k_{\pm} \cdot (x_{\text{reco}} - x_{\text{true}}), \quad (8.1)$$

which means that the resolution is artificially increased or decreased by a factor of  $(1 + k_+)$  or  $(1 - k_-)$ , respectively.

5) Most often,  $k_+$  and  $k_-$  are not distinguished, that is  $k_+ = k_- \equiv k$ .



**Figure 8.5** Hypothetical invariant-mass distribution of a  $B$ -meson decay. Data points (with error bars) with original simulation (solid line) and mass resolution modified by  $\pm 15\%$  (dashed and dotted).

An example invariant-mass resolution of a hypothetical  $B$ -meson decay is shown in Figure 8.5: the dashed and dotted lines represent Monte Carlo simulations with  $+15\%$  and  $-15\%$  of the nominal mass resolution, respectively. Obviously, those lines do not fit the data as well as the nominal simulation (solid line), so they can be seen as worst cases of an under-estimation or over-estimation of the resolution. The analysis is then redone with the modified simulated events, and the difference to the nominal result is taken as (conservative) systematic uncertainty.<sup>6)</sup>

The advantages of this technique are the simple implementation and the fact that no additional Monte Carlo datasets need to be produced. In principle, one would of course try to modify detector resolutions at the source (e.g. the simulation of drift times in wire chambers). However, this is highly non-trivial and needs a thorough understanding of the detector. Instead it is much simpler to look at a distribution like a mass resolution, which can be directly compared to the data.

#### 8.4.2.3 Theory Uncertainties

Practically every data analysis needs some input from theory. Unfortunately, often more than one theoretical description is available. Moreover, all of them usually have some flaws in their description of the real world. Examples include parton density functions, fragmentation functions, expected decay distributions, and so on. Sometimes theorists give some estimate on the uncertainty of their calculations (e.g. from contributions of higher-order diagrams), in such cases one can just take the given uncertainty and follow the recipe in Section 8.4.1.1. In other cases, however, no uncertainties are given, but several competing theoretical predictions exist. Usually, one theoretical prediction is taken as 'default' (the most reliable or the most commonly used) and others are used for estimating the uncertainty. At first, let us consider only one additional theoretical prediction. If there are no other means for determining the uncertainty of the theory, the last resort is to take the difference to the second prediction as the estimate of the systematic error. The

6) The resolution issue usually comes into effect when fitting a signal peak on top of a background distribution, since the obtained yield will depend on the assumed resolution.

underlying idea is, that the results for all theoretical predictions will approximately follow a Gaussian distribution – therefore the difference  $\Delta$  between just two theoretical choices is a (very rough!) estimator of the width of this distribution. Of course, the quoted systematic uncertainty would be  $\pm\Delta$  and not  $\frac{+\Delta}{-0}$ .

If there is more than one theory to choose from, people tend to choose the largest discrepancy  $\Delta_{\max}$  to quote a systematic error of  $\pm\Delta_{\max}$ . This is somewhat conservative, as the correct uncertainty to quote would be the width of this distribution, that is the square root of the variance of all results obtained with the different theoretical inputs. The choice of taking either the maximum discrepancy or just the spread is somewhat arbitrary (therefore it should be well documented), but again an educated guess may help: are all theories trustworthy? Have they been checked in other measurements? Are they actually the same theory with just little differences or do they use very different approaches?

An example of theoretical uncertainties are *Quantum Electrodynamics (QED) radiative corrections*, arising from final state electromagnetic interactions. In the simulation, usually two possibilities exist: either the radiative corrections are switched on or they are switched off. However, also a simulation including radiative corrections will not be perfect, normally higher-order effects such as for example internal photon lines are neglected. How can the corresponding systematic uncertainty be estimated? The scenario without radiative corrections is surely wrong. The one with corrections is certainly better, but not fully correct either. It is therefore reasonable to take a fraction of the difference as systematic uncertainty. The art, however, is to decide the size of the fraction: for this again an educated guess is required, for example by looking at data–Monte Carlo comparisons with and without radiative corrections.

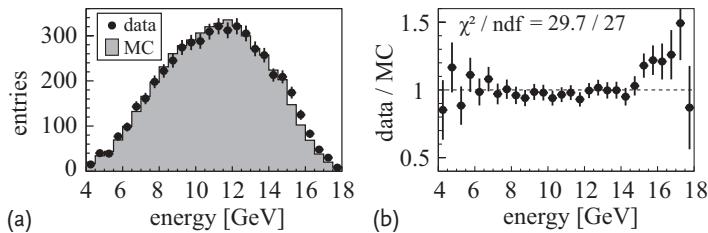
More on theory uncertainties can be found in the following Chapter 9, where the focus is on *strong interactions* which play a very important role in high energy physics.

#### 8.4.2.4 Discrepancies between Data and Simulation

In virtually every data analysis there is at least one data distribution which is not perfectly described by the Monte Carlo simulation. Imagine distributions as shown in Figure 8.6: in a large range, data and simulation agree well, which is also indicated by a practically perfect  $\chi^2/\text{ndf}$  of 29.7/27 between both distributions. However, at high energies there is a significant discrepancy between data and simulation,<sup>7)</sup> which should cause a systematic uncertainty. How can we estimate this uncertainty? First of all: as stated in many places in this chapter, one has to try to find the source of the discrepancy! It is always better to understand the origin and fight it directly than curing the symptoms.

Let us assume that – despite significant effort – the search for the origin was not successful. Still, a systematic error needs to be quoted. One possible source of the effect is incorrect modelling of the detector acceptance by the Monte Carlo

7) The reason for the good  $\chi^2$  value is that there is extremely good agreement at lower energies which arises from a statistical fluctuation. Simply looking at the distributions therefore still tells more than just the bare  $\chi^2$  value!



**Figure 8.6** Example of a non-perfect comparison between data and MC simulation. (a) Original distributions. (b) Data divided by simulation.

simulation. Assuming this to be the only source, then what is the size of the effect? About 10% of the data lie in the poorly described region. Inside this region the data lie about 20% above the MC simulation. So, the size of the effect is  $0.1 \cdot 0.2 = 2\%$ . Since we unfortunately have no other means to determine the uncertainty, we could quote  $\pm 2\%$  as the systematic error – and indeed this is usually done.

However, there is one great danger in this procedure: since the simulation usually needs to be normalised to the data statistics, it is unclear if the problem really lies in the upper part of the plotted distributions. Instead, one could renormalise the simulated data in order to have agreement in the upper part. Then there would be a similar disagreement of 20% between data and simulation, but now for the lower part of the distribution, which represents 90% of the data. The estimated systematic uncertainty would now be  $0.9 \cdot 0.2 = 18\%$  instead of 2%! Of course, it seems unlikely that only the small upper part would be simulated well and not the larger lower part of the data distribution. But as we do not know the cause of the disagreement, we actually may grossly underestimate the systematic uncertainty. Therefore again: it is always necessary to have at least a rough understanding of any disagreement between data and simulation.

#### 8.4.2.5 Analyses with Small Statistics

Analyses with limited amounts of statistics have the problem that comparisons between data and simulation are not easily possible. As statistical uncertainties are large, statistical fluctuations may mimic systematic problems or, conversely, systematics may be hidden underneath the large statistical errors. Therefore, if systematic effects are present, they can never be proven to be smaller than the statistical error.

In this case, two methods are possible: first, the data sample could be enlarged in a controlled way. This means to loosen some of the selection criteria which are not critical for the investigated systematic. Then, with the enlarged data statistics, meaningful systematic studies can be performed. A good example is the DELPHI analysis described in Section 8.3.3: the discrepancy of about 15% between data and the simulated  $\gamma\gamma$  background in the transverse-energy distribution (Figure 8.1c) was taken as the systematic uncertainty on the  $\gamma\gamma$  background normalisation.

Second, one can use similar, but more abundant control channels. This would for instance be a well-known decay with a similar final state or, as another example,

pion tracks instead of muon tracks, if you are interested in tracking systematics and not in particle identification.

#### 8.4.3

##### Cut Variations

With the variation of selection criteria ('cuts') the dataset of an analysis is enlarged or reduced and thus effects of specific sources of systematics may be enhanced or diminished. A cut variation is usually performed in several small steps. A typical total range of such variations is the change of the final sample size by, let's say, a factor of two, but in different situations it may be less. However, the range should always include possible larger changes in the analysis inputs such as background contributions, particle identification, trigger efficiencies, and so on.

In order to obtain a measurement, the data have to be corrected for detector acceptances which are usually determined from Monte Carlo simulations. Therefore, if a cut is varied, it has to be done simultaneously in the data and the simulation. A variation of the result with a cut variation often points to a problem with the acceptance as a function of the cut variable. (Another possibility would be a forgotten or badly understood background source.)

While cut variations are a necessary procedure for cross-checking the consistency of an analysis (see Section 8.3.2), they are also very commonly used for the determination of systematic uncertainties. However, this is not recommended here. There are several reasons not to use cut variations for estimating systematics effects quantitatively: first, the statistical significance is limited. If the size of the data sample changes by for example a factor of two, statistical fluctuations are of the order of the statistical error of the results. This means that it is impossible to find systematics smaller than the statistical error. Using uncorrelated errors with cut variations (see below) does help, but does not completely solve this problem. Second, cut variations do not help in understanding the underlying problem: if a result shifts with the variation of a cut, it is only a hint towards a flaw in the analysis. Investigating the underlying data and Monte Carlo distributions gives much more information, since a problem seen by a cut variation just reflects a discrepancy between data and Monte Carlo simulation. Finally, estimating systematic uncertainties from variations of several selection cuts, and adding them in quadrature while not taking into account potential correlations, may result in an incorrect estimate of the total systematic uncertainty.

Despite the strong discouragement to use cut variations for quantitative estimates of systematic errors, there may be situations where they are the only possibility, for example when a specific source for a discrepancy between data and Monte Carlo simulation could not be found, even after significant effort. Another situation exists when the source of the systematic uncertainty is known, but no other handle exists to access it quantitatively.

How can one find out whether or not the variation is significant at all and what the size of the systematic error attached to it is? The problem is the correlated datasets: if a cut is tightened, the remaining dataset is completely contained in the

sample with the default selection, which means that the statistics are highly correlated. Vice versa, loosening a cut just adds events to the default sample and again the datasets are correlated. The way out of this dilemma is to consider separate datasets, as sketched in Figure 8.7a. Let us assume that the data sample A with the default selection criteria yields a result  $x_A$  with a statistical uncertainty  $\sigma_A$ . Then a selection cut is tightened, so that a dataset B remains, which gives a result  $x_B \pm \sigma_B$ . Since B is completely contained in A, the relative statistical error  $\sigma_B/x_B$  on  $x_B$  is larger than  $\sigma_A/x_A$  (except for possible non-linear effects in, for example, fits), and  $\sigma_A$  and  $\sigma_B$  are correlated. However, one can construct a dataset C, which consists of all events of A except for those contained in B, as shown in Figure 8.7a.<sup>8)</sup> The sample C would give a result  $x_C \pm \sigma_C$ . Since B and C have no overlap, their two results are statistically uncorrelated. One can therefore calculate their weighted average (see (2.18) in Section 2.3.3), which of course is equal to the measured  $x_A \pm \sigma_A$ :

$$\bar{x} = \sum_i \frac{x_i}{\sigma_i^2} \Big/ \sum_i \frac{1}{\sigma_i^2} = \frac{x_B/\sigma_B^2 + x_C/\sigma_C^2}{1/\sigma_B^2 + 1/\sigma_C^2} \stackrel{!}{=} x_A , \quad (8.2)$$

$$\sigma_{\bar{x}}^2 = 1 \Big/ \sum_i \frac{1}{\sigma_i^2} = \frac{1}{1/\sigma_B^2 + 1/\sigma_C^2} \stackrel{!}{=} \sigma_A^2 . \quad (8.3)$$

Since  $x_A$ ,  $x_B$ ,  $\sigma_A$ , and  $\sigma_B$  are known,  $x_C$  and  $\sigma_C$  can be determined:

$$x_C = \frac{x_A/\sigma_A^2 - x_B/\sigma_B^2}{1/\sigma_A^2 - 1/\sigma_B^2} = \frac{\sigma_B^2 x_A - \sigma_A^2 x_B}{\sigma_B^2 - \sigma_A^2} \quad (8.4)$$

and

$$\frac{1}{\sigma_C^2} = \frac{1}{\sigma_A^2} - \frac{1}{\sigma_B^2} . \quad (8.5)$$

With this, the difference between the uncorrelated  $x_B$  and  $x_C$  is

$$x_C - x_B = \frac{\sigma_B^2 x_A - \sigma_A^2 x_B}{\sigma_B^2 - \sigma_A^2} - \frac{(\sigma_B^2 - \sigma_A^2) x_B}{\sigma_B^2 - \sigma_A^2} = \sigma_B^2 \frac{x_A - x_B}{\sigma_B^2 - \sigma_A^2} . \quad (8.6)$$

This difference has a statistical significance (in units of standard deviations) of

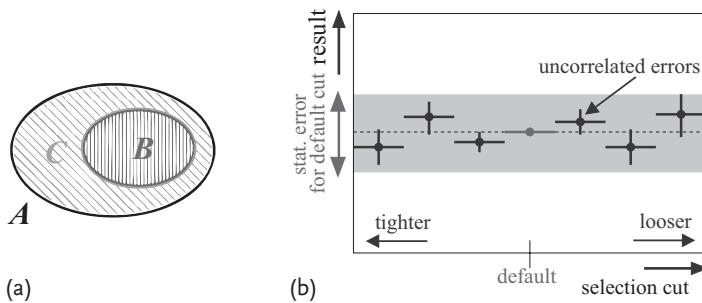
$$\frac{x_C - x_B}{\sqrt{\sigma_B^2 + \sigma_C^2}} = \sigma_B^2 \frac{x_A - x_B}{\sigma_B^2 - \sigma_A^2} \cdot \frac{1}{\sqrt{\sigma_B^2 + \frac{\sigma_A^2 \sigma_B^2}{\sigma_A^2 - \sigma_B^2}}} = \frac{x_A - x_B}{\sqrt{\sigma_B^2 - \sigma_A^2}} . \quad (8.7)$$

This, of course, is exactly the same statistical significance in units of standard deviations as the sought-after difference between  $x_A$  and  $x_B$ . Therefore, the uncorrelated part  $\sigma_{\text{uncorr}}$  of the uncertainties  $\sigma_A$  and  $\sigma_B$  is given by

$$\sigma_{\text{uncorr}}^2 \equiv |\sigma_B^2 - \sigma_A^2| . \quad (8.8)$$

The absolute value takes into account the case of loosening a cut, where  $\sigma_B < \sigma_A$ .

8) The usually rather small dataset C is only used for the mathematical derivation of the following formulae. Therefore, we neither need to select its events nor to analyse them.

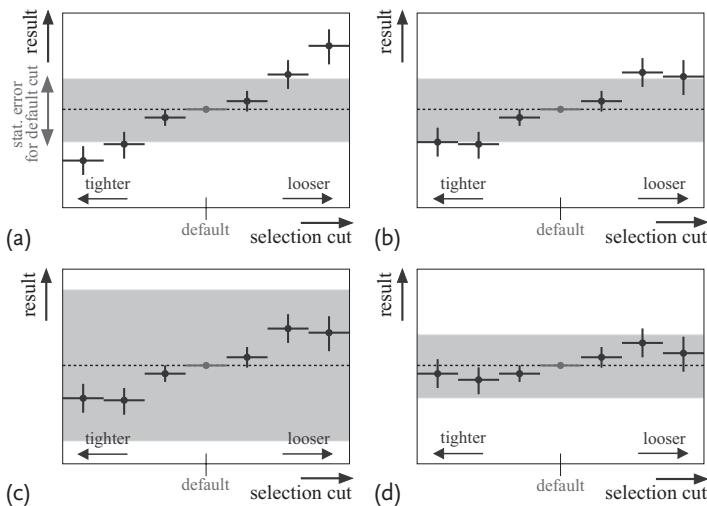


**Figure 8.7** (a) Separation of a data sample *A* into a dataset *B*, which fulfills a tighter cut criterion, and an independent dataset *C* = *A* − *B*. (b) Illustration of a cut variation.

Figure 8.7b is an illustration of a typical cut variation. It shows the result for the default set of cuts with no error bar (indicated as ‘default’), and the results with loosened or tightened selection criteria with uncorrelated errors according to (8.8) (the outer six data points). In the case of a well-understood analysis, the fluctuations of the results obtained with tightened or loosened selection criteria are just statistical and, at best, much smaller than the statistical uncertainty of the default result (indicated by the shaded band).

How and when can systematic uncertainties be estimated from cut variations? Figure 8.8 shows several typical examples of possible outcomes of a cut variation.

**Figure 8.8a:** The result shows a more or less linear behaviour with variation of the selection cut. It is not possible to assign a systematic uncertainty in this case, since the choice of the considered cut variation range is completely arbitrary. Instead, the source of the variation needs to be found and either removed or, at least, under-



**Figure 8.8** (a–d) Examples for different outcomes of cut variations.

stood. To investigate the problem, it is – among other things – useful to take a look at the data and Monte Carlo distributions in the cut variable. If there is an understanding of the problem, a systematic uncertainty can then possibly be assigned by using other methods than a cut variation.

**Figure 8.8b:** This example is similar to (a), except that the outermost variations show a somewhat better behaviour. This may have two different causes: either the second and the last but one cut variation give indeed the worst result, or the first and the last point just show downward fluctuations from an actually linear trend as in the previous example. From this plot it is impossible to tell which of the two cases may apply, and therefore the observed variation needs to be understood. As always, one needs to look at the underlying data and Monte Carlo distributions of the cut variable: do they show large discrepancies, also outside the range of the cut variation? If yes, then there is no way to attach a systematic uncertainty before the reason for the discrepancy is understood. If not, it is still strongly advised to find the reason for this particular behaviour in the region of the selection cut. If nothing is found, it is common practice to assign a systematic uncertainty of the size of the maximum variation. This is somewhat arbitrary, since the two most outlying data points may contain large statistical fluctuations. This method is therefore not advocated here. Instead, the source of the problem needs to be found.

It should in addition be noted that in this specific example the obtained systematic uncertainty would be about the size of the statistical error. This means that even a small under-estimation or over-estimation of the systematic uncertainty would directly affect the total error! This is another motivation to really understand the source of the variation.

**Figure 8.8c:** This example is the same as the previous one, but the statistical uncertainty of the result is twice as large as before and therefore dominant. This means that a slight mis-estimation of the systematic error, which would be determined from the observed variation, does not affect the final result much. Still, the reason for this variation – which is significant, as can be seen from the uncorrelated errors – should be better understood by for example looking at the underlying distributions. In particular one needs to exclude a linear trend as in example (a).

**Figure 8.8d:** Here we may have just a statistical fluctuation, as all the data points are compatible with no variation at all. Also the seeming trend of the five central points is not necessarily an indication of a problem if one keeps in mind that the single data points are uncorrelated to the default cut value, but are strongly correlated among each other. Conversely, a general trend cannot be excluded either. The advice is therefore the same as usual: first, compare the data and simulation distributions. Do they agree with each other? Second, investigate even tighter or looser selection cuts, if possible. Do they support a trend or not?<sup>9)</sup> Finally, as always, think about possible problems which could show up in the selected cut variable! Are there

9) If yes, this should already be seen by a disagreement of the underlying distributions!

reasons – like mis-estimated background, wrong energy scales, and so on – for a change of the result when varying the cut value? If yes, perform a direct evaluation of the systematics by, for example, changing the background or the energy scale, as explained in detail in Section 8.4.2 above. The result of the cut variation can then be neglected, as long as it is within the directly evaluated systematic uncertainty. If it is not, one unfortunately has to think of additional systematics, which could contribute to a trend in the cut variation results, and one has to repeat the procedure.

#### 8.4.3.1 Cut Variations in Multi-Dimensional Analyses

The method of cut variation can be expanded to the multi-dimensional case, where the result – for example from a multi-dimensional fit – is given by a set  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of  $n$  parameters with a covariance matrix  $\mathbf{V}$ . To determine the uncorrelated difference between two results  $\mathbf{x}_A$  and  $\mathbf{x}_B$  from two datasets  $A \subset B$  (or  $B \subset A$ ) the *uncorrelated covariance matrix*  $\mathbf{V}_{\text{uncorr}}$  (analogous to the uncorrelated error in (8.8)) is needed:

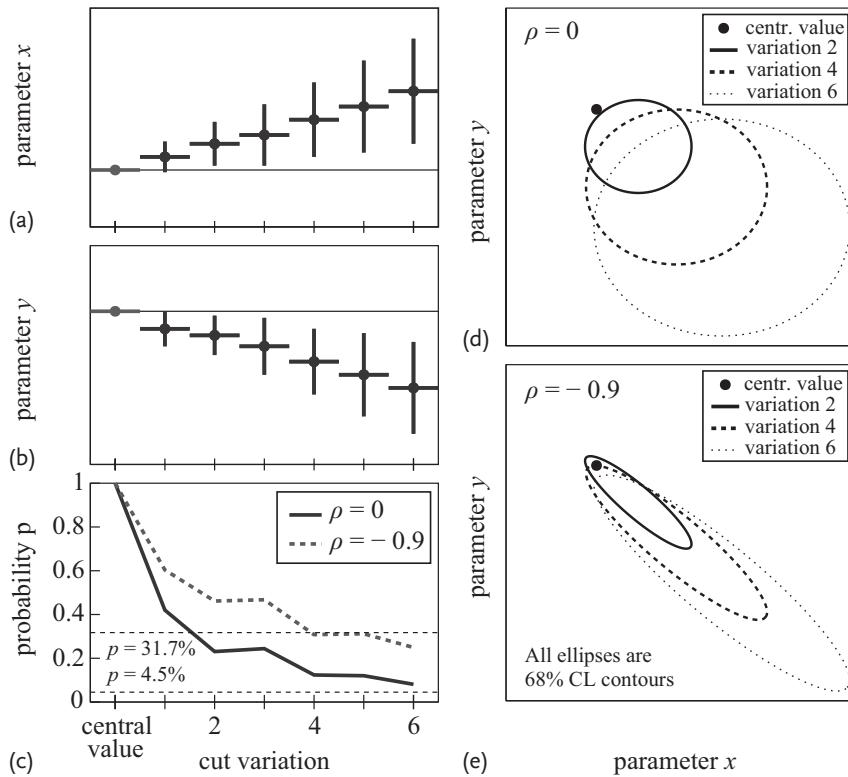
$$\mathbf{V}_{\text{uncorr}} \equiv \pm(\mathbf{V}_B - \mathbf{V}_A), \quad (8.9)$$

where the plus sign applies for  $B \subset A$  (as in Figure 8.7a), and the minus sign for  $A \subset B$ . The degree of agreement between the two results is then given by the uncorrelated  $\chi^2$  difference

$$\Delta\chi^2_{\text{uncorr}} \equiv (\mathbf{x}_A - \mathbf{x}_B)\mathbf{V}_{\text{uncorr}}^{-1}(\mathbf{x}_A - \mathbf{x}_B)^T, \quad (8.10)$$

which is the squared number of standard deviations between  $\mathbf{x}_A$  and  $\mathbf{x}_B$ . It is important to note that in the multi-dimensional case the standard deviations do *not* correspond to 68% confidence regions, as for a result with only one parameter, but to much smaller regions. For two parameters, the  $1\sigma$  contour (i.e. the error ellipse) encloses just a 39.4% confidence region, and for three parameters, the  $1\sigma$  ellipsoid corresponds to a 19.9% confidence region. The 68% confidence regions, which people are used to, are given by the  $1.52\sigma$  contour ( $\Delta\chi^2 = 2.30$ ) in the 2-dimensional and by the  $1.88\sigma$  contour ( $\Delta\chi^2 = 3.53$ ) in the 3-dimensional case.

An example for the 2-dimensional case is shown in Figure 8.9: an analysis measures two parameters  $x$  and  $y$  in a global fit. The outcome of a hypothetical cut variation is shown in Figure 8.9a–c. Both  $x$  and  $y$  show a clear trend away from the default result, with both about  $1.5\sigma$  off in the worst case, strongly indicating a potential problem of the analysis. However,  $x$  and  $y$  may be correlated with a correlation coefficient  $\rho$  which we have to take into account when deciding about the significance of the outcome of the cut variation. For this we need to compute the  $\chi^2$  difference using (8.10). The agreement is then given by the  $\chi^2$  probability  $P(\chi^2; \text{ndf})$ , where the number of degrees of freedom is the number of fit parameters. We first consider completely uncorrelated results  $x$  and  $y$  ( $\rho = 0$ ): in this case



**Figure 8.9** (a–e) Example of a cut variation of an analysis with either two uncorrelated ( $\rho = 0$ ) or two strongly correlated ( $\rho = -0.9$ ) variables  $x$  and  $y$ . See the text for a full explanation.

the  $\chi^2$  probability  $P = P(\Delta\chi^2, \text{ndf} = 2)$  drops quickly as the cut is varied,<sup>10)</sup> as is shown in Figure 8.9c.

If, in contrast, we look at an extreme case of a correlation coefficient  $\rho = -0.9$ , that is an almost full anticorrelation between  $x$  and  $y$ , we obtain a very different behaviour: because of the large anticorrelation, the probability always stays around the 31.7% line, corresponding to about one standard deviation in the 1-dimensional case, even though each single parameter shows a larger variation.<sup>11)</sup> Figure 8.9d,e show the error ellipses  $V_{\text{uncorr}}$  (8.9), scaled to contain 68% probability, for some of the cut variations. They also nicely illustrate the effect of anticorrelation between fit parameters. In practice, anticorrelations are quite common, for example when

10) Note that  $P$  is still larger than the product  $P(\Delta\chi_x^2; 1) \cdot P(\Delta\chi_y^2; 1)$  of the two single probabilities, because of the different meaning of a  $1\sigma$  contour in the 2-dimensional case.

11) This may be astonishing at first glance. One could argue that, if the second parameter did not show any variation at all (i.e. was

not affected by the applied cut), then one would automatically have the 1-dimensional case. However, because the second parameter shows an (anticorrelated) variation, the probability increases!

fitting the offset and slope of a straight line with data points far away from the coordinate origin.

#### 8.4.4

##### Combination of Systematic Uncertainties

In most cases, several different systematic uncertainties have to be combined to a common systematic error, which then is combined with the statistical error to obtain the total uncertainty of the measurement. For simple analyses, which measure only one single parameter, this combination is fairly simple: usually each source of systematic uncertainty should be independent from the other and is, of course, independent of the statistical error. Therefore, as for independent statistical errors, all the single uncertainties can be added in quadrature to obtain the total uncertainty:

$$\sigma_{\text{tot}}^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 = \sum_{i=1}^n \sigma_i^2. \quad (8.11)$$

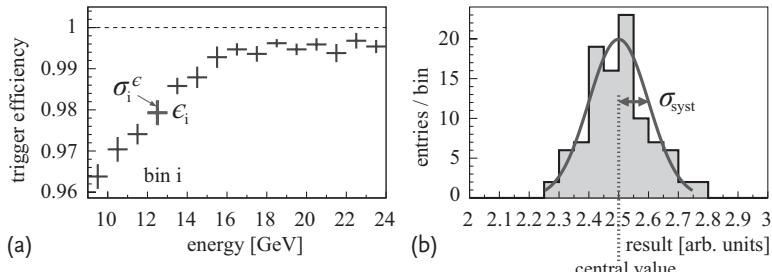
Somewhat more care has to be taken when systematics are potentially correlated, like for example the shape and the normalisation of a possible background. In this case, the correlation coefficient  $\rho_{ij}$  between each pair of sources  $i$  and  $j$  of systematic uncertainty needs to be determined, and the uncertainties are combined according to the law of error propagation as

$$\sigma_{\text{tot}}^2 = \sigma_1^2 + \sigma_2^2 + \cdots + 2\rho_{12}\sigma_1\sigma_2 + \cdots = \sum_{i=1}^n \sigma_i^2 + 2 \sum_{\substack{i,j=1 \\ i < j}}^n \rho_{ij} \sigma_i \sigma_j. \quad (8.12)$$

The situation is even more complicated when several parameters  $\mathbf{x} = (x_1, \dots, x_n)$  are measured together. A simple example would be a combined fit of the slope  $m$  and the offset  $b$  of a straight line to a data distribution. The fit yields a correlation coefficient  $\rho$  between the statistical uncertainties of the fit estimates  $m$  and  $b$ . But what is the correlation coefficient  $\rho^{\text{syst}}$  between the corresponding systematic uncertainties? In many analyses these are not determined – maybe not even thought about – and therefore neglected, resulting in a false total correlation. However, it is usually not difficult to determine the correlation of uncertainties due to one specific source of systematics. In general there are three cases for each set of two different systematics:

**Fully correlated uncertainties** This is the most frequent case. If in the above example of a straight-line fit an external parameter is varied by  $\pm\sigma_{\text{syst}}$ , slope and offset may vary by  $\pm\Delta m$  and  $\pm\Delta b$ , respectively. The variations  $|\Delta m|$  and  $|\Delta b|$  are then taken as systematic uncertainties  $\sigma_m$  and  $\sigma_b$  and, depending on whether  $\Delta m$  and  $\Delta b$  have the same or the opposite sign, the correlation coefficient is either +1 or -1.

Since parameter variation is the most commonly used tool for the estimation of systematics effects, most systematic uncertainties cause full correlation between the parameters of one measurement.



**Figure 8.10** Illustration of the use of a toy Monte Carlo simulation. (a) Hypothetical trigger efficiency as a function of energy. Each energy bin  $i$  has a separate trigger efficiency  $\epsilon_i \pm \sigma_i^e$ , which is uncorrelated to the other bins. (b) Possible results of a toy Monte Carlo simulation with  $N = 100$  events, which varied

all trigger efficiencies  $\epsilon_i$  simultaneously with their uncertainties. Each variation results in a slightly different analysis result; the total systematic error (and a possibly shifted mean) can be read off from the central value and the width  $\sigma_{\text{syst}}$  of the distribution.

**Uncorrelated uncertainties** Astonishingly, this is a rare case. Of course, a source of systematic uncertainty may only affect one of two parameters – but this is the trivial case of uncorrelated systematic errors, since one of them is zero. A source of systematics, which results in really independent systematic uncertainties on parameters of the same measurement, is rather untypical. Think, for example, of causes, which would independently affect slope and offset of a straight-line fit.

**Partially correlated uncertainties** Partial correlations are often a result of several sources of systematics, which themselves are correlated among each other. Each single source may cause fully correlated systematic uncertainties on the measurement results, but the correlation among the sources causes correlations  $\neq \pm 1$  among the measured parameters.

**Using a toy Monte Carlo simulation** To obtain the correct correlation coefficient, a toy Monte Carlo simulation may be useful: consider  $n$  non-independent systematic parameters  $\mathbf{a} = (a_1, \dots, a_n)$  in the above example of a straight-line fit with slope  $m$  and offset  $b$ , with correlations described by an  $(n \times n)$  covariance matrix  $\mathbf{V}$ . Then a large number of random-number ntuples  $\Delta \mathbf{a}$  can be generated according to the multi-dimensional Gaussian given by  $\mathbf{V}$ . For every single ntuple, all systematic sources  $i$  are varied by  $\Delta a_i$  and the slope  $m$  and the offset  $b$  of the fit are recomputed and filled into a 2-dimensional histogram. Finally, assuming everything to be sufficiently Gaussian (it usually is), the resulting 2-dimensional Gaussian distribution in  $m$  and  $b$  directly provides the systematic uncertainties  $\sigma_{\text{syst}}^m$  and  $\sigma_{\text{syst}}^b$ , as well as their correlation  $\rho_{\text{syst}}^{m,b}$ .

This method of using a toy Monte Carlo simulation is also very useful in the case of non-correlated systematic uncertainties. Consider, for example, a trigger efficiency  $\epsilon$  which is known in intervals  $i$  of energy:  $\epsilon_i \pm \sigma_i$  (see Figure 8.10.) The usual method to compute the overall systematic uncertainty would be to separately vary

the trigger efficiency  $\epsilon_i$  for each energy interval  $i$  by  $\pm\sigma_i$  and to add all resulting changes of the measurement in quadrature. However, with many energy bins this may become very tedious. A more efficient method is a small toy Monte Carlo simulation, which for each event generates efficiencies  $\epsilon_i$  for each bin  $i$  according to their single (usually Gaussian) uncertainties  $\pm\sigma_i$ . For each toy Monte Carlo event the measurement is performed with the newly generated efficiencies and the results are filled into one histogram. The mean and the width can then easily be read off, with the width giving the total systematic uncertainty.

#### 8.4.4.1 Combination of Covariance Matrices

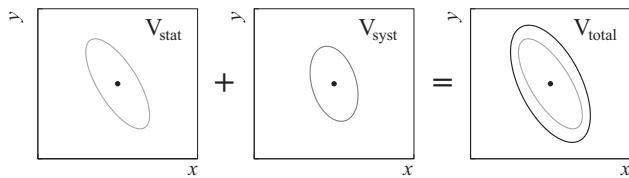
In the end, each source or each set of sources of systematic effects causes uncertainties on the measured parameters together with a correlation among them, both represented by a covariance matrix  $V_{\text{syst}}^i$  for the  $i$ th source. The combination of  $n$  systematics then is straightforward: all covariance matrices  $V_{\text{syst}}^i$  ( $i = 1, \dots, n$ ) are just added together to give the final systematic uncertainties together with their correlation:

$$V_{\text{syst}} = V_{\text{syst}}^{\text{full corr}} + V_{\text{syst}}^{\text{not corr}} + V_{\text{syst}}^{\text{part corr}} \quad (8.13)$$

$$\begin{aligned} &= \sum_i^{\text{full corr}} \begin{pmatrix} \sigma_{1,i}^2 & \cdots & \pm\sigma_{1,i}\sigma_{n,i} \\ \cdots & \cdots & \cdots \\ \pm\sigma_{1,i}\sigma_{n,i} & \cdots & \sigma_{n,i}^2 \end{pmatrix} + \sum_i^{\text{not corr}} \begin{pmatrix} \sigma_{1,i}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sigma_{n,i}^2 \end{pmatrix} \\ &\quad + \sum_i^{\text{part corr}} \begin{pmatrix} \sigma_{1,i}^2 & \cdots & \rho_{1n,i}\sigma_{1,i}\sigma_{n,i} \\ \cdots & \cdots & \cdots \\ \rho_{1n,i}\sigma_{1,i}\sigma_{n,i} & \cdots & \sigma_{n,i}^2 \end{pmatrix}, \end{aligned} \quad (8.14)$$

where the first sum runs over the fully correlated, the second over the uncorrelated, and the third over the partially correlated systematic errors.

Finally, the covariance matrices of the statistical and systematic uncertainties are added in the same way to obtain the covariance matrix of the complete result. In the case of just two parameters the covariance matrices can be represented as the usual error ellipses, as shown in Figure 8.11. In some cases, for instance when the correlations are very different, it may be useful to plot both the statistical and the total error ellipse.



**Figure 8.11** Addition of the covariance matrices of statistical and systematic uncertainties for two correlated measured parameters  $x$  and  $y$ .

## 8.5

### How to Avoid Systematic Uncertainties

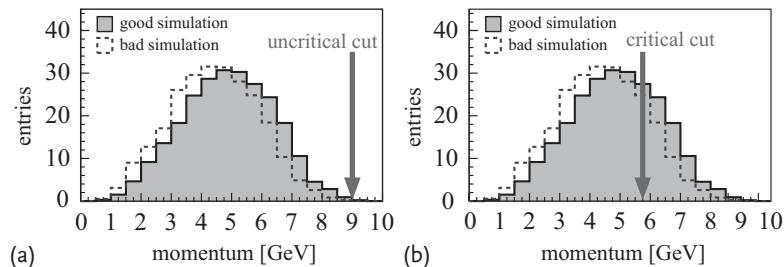
Most systematic uncertainties cannot easily be avoided when doing data analyses. Examples are systematics due to trigger efficiencies, calibrations, and backgrounds. However, some systematics may be suppressed or even avoided when an analysis is planned or performed.

#### 8.5.1

##### Choice of Selection Criteria

Possible systematic uncertainties should already be considered at the very beginning of an analysis, when selection criteria are determined. One should always keep in mind that every selection cut not only reduces the acceptance, but may also lead to systematic problems if the acceptance as a function of the cut variable is not well understood. Normally this does not apply much for geometrical cuts (except for inefficient detector regions), but rather for cuts on quantities which involve a detailed simulation. Typical examples of such quantities are timing parameters,  $\chi^2$  values of track or vertex fits, track impact parameters, particle identification, and so on. All these quantities are notoriously difficult to simulate, as the underlying detector resolutions need to be well known as a function of several, sometimes external, parameters.

In principle, every cut on a badly simulated quantity will lead to different acceptances in data and simulation and therefore to a systematic shift of the result. However, there is a simple way to avoid such systematics: just do not or only loosely cut on these quantities! This is illustrated in Figure 8.12: both plots show the simulated momentum distributions of Figure 8.2, with one of the simulations not agreeing with the data. If applying a momentum cut as shown in Figure 8.2a, this is not a big problem, since the acceptance for the momentum is always  $\approx 99\%$ , and therefore any possible error is  $< 1\%$ . In Figure 8.2b, however, a badly chosen momentum cut is shown. From the false Monte Carlo simulation, an acceptance of  $\approx 80\%$  is derived, while the real acceptance is just 70%. If a rate is measured, a



**Figure 8.12** Example of the choice of uncritical and critical selection criteria. (a) Uncritical cut that accepts all events with momentum  $< 9$  GeV. (b) Critical selection cut, resulting in a potentially miscalculated detector acceptance.

systematic error of  $0.8/0.7 - 1 = 14\%$  would therefore be made by using the wrong simulation. And the situation would become even worse should the cut be further lowered.

The bottom line of this discussion is that the necessity of every single selection criterion should be checked. Furthermore, all related systematic effects have to be kept under control.

Sometimes a cut on a badly described quantity cannot be avoided. An example is particle identification, where one normally has to apply hard selection criteria in order to get the background under control. In these cases, it is necessary to not rely on the simulation, but to use the data themselves to determine the acceptances, for example by using similar, but well-known channels.

In many cases also a trade-off is possible. By loosening a cut, you may for example enhance the background, but avoid systematics due to non-understood detector acceptances. The additional background will add an additional uncertainty, but it may be more straightforward to estimate than the acceptance uncertainty.

### 8.5.2

#### Avoiding Biases

There are many possible reasons for the introduction of a bias to a measured result. Some may be due to external factors such as faulty input parameters, others may be inherent in the analysis method, such as for example an unavoidable bias of a fit procedure. Such cases need to be investigated, but with the tools which have been described in this chapter one should hopefully be able to keep them under control and to estimate them accurately.

Another source of biases, however, cannot easily be found and estimated in the usual way: it is the analyst him/herself. Like anyone else, physicists have feelings, thoughts, emotions, and so on which may also affect an analysis, even though we may like to think of ourselves as unbiased scientists. This human factor often leads to biased results, which in many cases are not noticed, since they usually cannot be spotted with the customary techniques. The two most common cases – the expectation of a specific result and the use of signal events to define analysis parameters – are discussed in the following. Section 10.6.1 deals in greater detail with the problems arising from the experimenter's preconceptions – and with possible counter-measures.

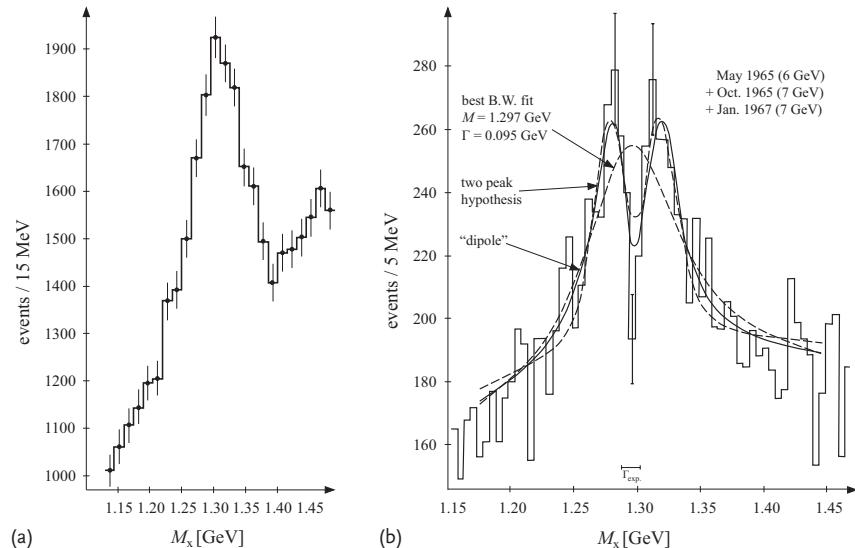
##### 8.5.2.1 Do not Expect a Certain Result

There are numerous examples of measurements where experimentalists were biased by expecting a specific measurement result and, miraculously, indeed observed this very outcome Sections 8.5.3 and 10.6.1.

##### Example 8.4 The ' $A_2$ mass splitting'

A prominent example is the 1967 measurement of the so-called ' $A_2$  mass splitting' by the CERN Missing-Mass Spectrometer group [5]: the experiment found that the

invariant-mass distribution of the  $A_2$  meson in pion–proton scattering exhibited a double-peak structure. Hints for such a structure had already been seen by the same group in two runs in 1965, but with much less significance and with a mass resolution similar to the observed width of the structure. In the 1967 data-taking, the recoil proton momentum was in addition determined by its range, which provided a better mass resolution. The full dataset, plotted with a bin width of about the mass resolution, did not show any unusual behaviour (Figure 8.13a). However, when about 60% of the 1967 data with the best mass resolution were combined with all the 1965 data (which still had a worse resolution than the excluded 1967 data), a clear double-peak structure was observed (Figure 8.13b). A fit with a single Breit–Wigner function yielded a  $\chi^2$  probability of 0.1%, but a tuned double-Breit–Wigner function obtained 70% probability.



**Figure 8.13** Measurement of the  $A_2$  meson by the CERN Missing-Mass Spectrometer. (a)  $\pi^- p \rightarrow p X$  data with  $p_\pi = 7 \text{ GeV}$  from the year 1967. The bin size is 15 MeV. (b) Selected 1967 data with detector mass resolution (FWHM)  $15 < \Gamma_{\text{exp}} < 25 \text{ MeV}$  plus data from two 1965 runs with a different detector configuration and  $\Gamma_{\text{exp}} \approx 30 \text{ MeV}$ . The bin size is 5 MeV. (Adapted from [5].)

Of course, many people – theorists and experimentalists – were excited about this very unexpected structure, as it did not fit into the meson octet structure. Following on from this, several experiments also reported the double-peak structure, but, strangely, always with a significance of about three standard deviations. Finally, with much more statistics available, a BNL experiment [6] excluded the double-peak structure, and the  $A_2$  returned to being a normal meson.

What had happened? Why did many independent experiments observe a fake, non-existing structure? Were all the physicists working on these analyses bad scientists, trying to fake results to become famous? That is hard to believe, in particular as very many people were involved. Instead, what obviously happened in the

understandable excitement of discovering a completely new phenomenon, was that the analysts – without being conscious of it – chose their selection criteria and fit methods in a way that enhanced a possible double peak. Of course, since there was no real mass-splitting, they were never able to observe significances of more than three standard deviations, but because of the biased analyses, they usually observed a signal just on the edge of an observation, thus boosting trust in the reality of the observed structure.

In the history of particle physics, there are many similar examples, from the discovery of a heavy neutrino with a mass of 17 keV [7] to the various sightings of pentaquarks [8] just a couple of years ago. You probably know Figure 2 of the Introduction to the Review of Particle Physics (PDG) [9], which shows the development of several experimental quantities with time. Every now and then, all those measured values show significant jumps, pointing either to a common systematic shift or to the effect of biased analyses.

If one thinks about it a little more, the underlying mechanism of these ‘bad’ measurements becomes quite clear. If a measurement on a quantity has already been published, every new data analysis may have two possible outcomes: either it agrees with the previous measurement or it does not. In the first case, the physicist who performs the new measurement will probably be content (usually he or she has achieved a smaller error), lean back, and finish the analysis without thinking more deeply about it. In the case of a not too large disagreement (about one to three standard deviations), however, the scenario becomes very different: the physicist would be somewhat worried and would have a closer look for potential problems. In particular, he or she would look for explanations of the discrepancy in his/her own analysis, thus trying to find corrections that shift the result towards the formerly measured value. Conversely, he/she would not particularly look for effects that could cause an even larger discrepancy. In this way, the new measurement becomes heavily biased towards yielding a result close to the original value. Only if the discrepancy were great ( $5\text{--}10\sigma$ ), would the physicist actually begin to think on both his/her and the original analysis.

What can we learn from these examples? The answer is quite simple: free yourself from any prejudice in regard to the expected result! Do not care about previous measurements and theory expectations. At best, you only compare your result to others once the analysis is completely finished. Of course, practically this is not feasible if a preliminary outcome is already present when still performing the analysis. Therefore, one should always try to hide the result until the end of the analysis is reached. This is called a *blind analysis* – a concept which is described in Sections 8.5.3 and 10.6.1.

#### 8.5.2.2 Do not Look at Signal Events

Another very common case of introducing a bias is the definition of selection criteria by using the same data sample as is actually used for the final measurement. There are literally thousands of cases where people chose the selection criteria by looking at the signal itself – in order to see more signal events, to make the sig-

nal look nicer, to reduce the background, and so on. This may seem stupid (and it is), but it actually happens easily, even to people who are aware of the problem: when for example introducing a new feature like a selection criterion to the analysis, one may easily be tempted to take a glance at a signal distribution which might or might not look better than before. Once such an effect has been observed on a signal distribution, it is almost impossible to forget about it; and often – since it is well known that cuts must not be tuned to the data – *a posteriori* reasons are invented to justify the newly introduced selection criterion. Of course this is still nothing else than a strong bias towards a larger or better signal.

It is therefore absolutely necessary to *never* look at the data when introducing or optimising cuts. Selection criteria can only be determined by simulation, from a dataset with completely disjunct events, even by arbitrary choice or personal taste, but never by looking at the signal. This may sound very simple, but it should be kept in mind that it is a very common mistake – even though people in principle know about it. You may for example think of all the selection criteria of your current or last analysis: were they really all introduced without looking at the data?

The best protection against this kind of bias is of course again performing a blind analysis.

### 8.5.3 Blind Analyses

A very good way to avoid personal bias, is to not know the result during the analysis. This method is called *blind analysis*. It strictly forbids looking at the result before finishing the analysis. It is most easily described for measurements with only a few expected signal events and for searches: first, a signal region is defined, for instance a certain range around the expected invariant-mass value in a search for a rare process or the whole region above a certain value of missing transverse momentum in a search for new particles. This signal region may also be multi-dimensional. The definition of the signal region normally relies on simulation since it is forbidden to look at the data beforehand. In a second step, all data events which fall into the signal region are removed from the data files on which the analysis is performed. Alternatively, one may have a corresponding initial statement in the event analysis code which processes the data.<sup>12)</sup> Now the analysis can be performed, working mostly with simulated data and real data from the signal sidebands or from similar channels. At the very end, when all selection criteria and the whole analysis chain have been fixed and all systematic uncertainties have been estimated (or at least the method of their estimation has been fixed), then the events of the signal region are added (*unblinding*), the final result is obtained, and the paper is written.

It is *forbidden* to change the analysis after unblinding. Therefore, great care has to be taken to foresee any possible effect in the signal region and in the estimation of systematic errors. Since nothing can be changed afterwards, any neglected

<sup>12)</sup> But be careful: if you can't resist your own curiosity it is much easier to remove a statement from the code than to add the missing events to the data files!

feature may ruin the analysis. This is less the case in searches, where one usually has a clear picture of background and other analysis issues even without looking at the signal region, but occurs to a greater extent in analyses that are more complicated, for example where decay parameters are measured from a large number of signal candidates and many subtle systematic effects may play a role. However, what seems like a drawback is actually an advantage since all details have to be understood *beforehand*, and not just when a problem is seen – which then could introduce a bias.

Sometimes even more than the signal region is blinded: in order to avoid possible tails from the signal or to be independent of background fluctuations near the signal, additional ranges of, let's say,  $\pm 3\sigma - \pm 5\sigma$  from the expected signal are blinded. When finishing the analysis, first the outer blinded region is opened and compared to expectations. Only if everything is fine is then the signal region opened. In this way the analyst still has a last chance of taking into account overlooked systematics before the final unblinding.

The method of blind analysis can also be extended to more complicated cases. An example is the measurement of the CP-violating parameter  $\sin 2\beta$  in the system of neutral  $B$  mesons [10, 11]. Here the signal is an asymmetry in the time dependence of  $B^0$  mixing with respect to  $\bar{B}^0$  mixing. In this measurement, the signal events cannot be excluded since an asymmetry of all signal events is measured. Still, the analyses were blinded by randomising the initial flavour of the  $B^0/\bar{B}^0$  mesons. This was achieved by having different groups working on the determination of the  $B$  flavour, so-called *tagging*, and on the actual asymmetry measurement itself. Only shortly before the release of the analysis would the two groups combine their measurements to obtain the final result.

Apart from the huge advantage of avoiding any data-driven bias – which cannot be overemphasised – the disadvantages of a blind analysis are the much greater effort and time required than for a normal analysis. Therefore, it does not always make sense to blind the analysis result. If a clear signal of many thousands or millions of events exists and the aim is to measure a cross section or a decay distribution, blinding may obscure the view for subtle effects. For such measurements there are often not even accurate predictions or precise previous results. In such a case it is difficult to work towards an ‘expected’ result, since no specific result is distinct from another. As always in life, the truth is neither black nor white: blind analyses should always be performed when necessary, in particular when searching for new phenomena; but in other cases a normal analysis may be as good and may save a lot of work.

## 8.6 Conclusion

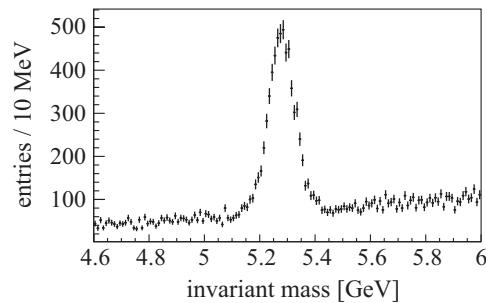
It is not easy to summarise a vast topic like the treatment of systematics, as almost every case has its own peculiarities and is somewhat different from others. Nevertheless, some general rules are always applicable, in particular for finding or

even avoiding potential systematic effects in a data analysis. Among other things, every physicist doing an analysis should therefore always consider the following questions:

- Have you thought of all potential systematic problems? How would they show up and what would the consequences be?
- Have you compared as many data distributions to the simulation as possible? Do they all look good? If not: what consequences would particular disagreements have on the result of the measurement?
- Have you performed as many cross-checks as possible? Are they all consistent with expectations?
- Have you shown your work on systematics to other people? Do they agree with your conclusions about the systematics?
- Is every selection criterion necessary? Can some be loosened or even be omitted in order to reduce potential systematic uncertainties?
- Are you sure that you have not been biased by looking at the results before finishing the analysis? Do you feel uneasy with the preliminary result because it is not what you expected? (The latter would be an indication of a biased analysis!)

With all these items checked, one should more or less be protected from having overlooked a serious problem.

Once found, systematic uncertainties have to be estimated. We have seen several examples of how to assess their size, which hopefully give practical ideas on how to evaluate systematic effects in concrete cases. Again, there are no general rules for the estimation of systematic uncertainties. What one needs are mostly some creativity, some experience, and, last but not least, a good portion of common sense. However: every experimental physicist should already have two out of these three talents, and experience will be gained with time.



**Figure 8.14** The invariant-mass distribution of the  $B$  meson decay discussed in the exercise.

## 8.7 Exercise

### Exercise 8.1 A hypothetical $B$ -meson decay

The rate of a hypothetical  $B$ -meson decay shall be measured with high precision. Figure 8.14 shows the distribution of the reconstructed invariant mass<sup>13)</sup> of signal candidate events around the nominal  $B$  mass of  $m_B = 5.28 \text{ GeV}$  together with background, which is assumed to be purely combinatorial (i.e. not containing other specific  $B$  decay channels). The branching fraction  $\mathcal{B}$  is then given by the number  $N$  of observed decay candidates in a given window around the nominal  $B$  mass, the number  $N_{\text{bkg}}$  of estimated background events in this window, and the signal acceptance  $\mathcal{A}$  by

$$\mathcal{B} = \frac{N - N_{\text{bkg}}}{\mathcal{A}} \cdot \frac{1}{\Phi_B},$$

where  $\Phi_B = 10^{10}$  shall be the total number (flux) of  $B$  decays in the detector. The acceptance  $\mathcal{A}$  is assumed to be 10%, without a possible selection cut on the invariant mass.

- a) The background can be estimated by sideband subtraction, using for example the regions from 4.68 to 4.98 GeV and from 5.58 to 5.88 GeV. What is the background estimate and its statistical error<sup>14)</sup> in a window of  $\pm 150 \text{ MeV}$  around the  $B$  mass? How could the systematic uncertainty on the background estimate be assessed?
- b) The detector resolution on the invariant mass was determined from Monte Carlo simulation to be  $\sigma^{\text{MC}} = 50 \text{ MeV}$ . What would be a good choice for the mass cut around  $m_B = 5.28 \text{ GeV}$ , when you either trust or do not trust this value for the resolution?
- c) From an additional study the detector resolution turned out to be known to only  $\pm 10\%$ . What is the corresponding branching ratio including the statistical uncertainty and the systematic uncertainties on both acceptance and background estimation for mass cuts of  $\pm 2\sigma^{\text{MC}}$  or  $\pm 3\sigma^{\text{MC}}$  around  $m_B$ , respectively? Which mass cut leads to the smaller total error?
- d) Rather than just counting events, in many analyses a fit of a Gaussian signal plus a polynomial for the background is performed. What is the result for the branching fraction when using a linear function to model the background?<sup>15)</sup> What would be possible sources of systematic uncertainties now and how could they be assessed? How could the difference between the fit result and the result obtained in Exercise 8.1c) for a  $3\sigma^{\text{MC}}$  mass cut be explained?

13) The corresponding data can be found in file `sysmeas.root`.

14) Note that this statistical uncertainty later turns into a systematic uncertainty on the branching ratio  $\mathcal{B}$ . The systematic uncertainty on the background estimate causes then another systematic uncertainty on  $\mathcal{B}$ .

15) Pay attention to the correct normalisation of the Gaussian!

### References

- 1 DELPHI Collab., Abdallah, J. *et al.* (2003) Searches for supersymmetric particles in  $e^+e^-$  collisions up to 208 GeV and interpretation of the results within the MSSM. *Eur. Phys. J. C*, **31**, 421.
- 2 NA48 Collab., Batley, J. *et al.* (2002) A precision measurement of direct CP violation in the decay of neutral kaons into two pions. *Phys. Lett. B*, **544** (1/2), 97.
- 3 Barlow, R.J. (1989) *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons.
- 4 NA48 Collab., Batley, J. *et al.* (2010) New precise measurements of the  $\Xi^0 \rightarrow \Lambda\gamma$  and  $\Xi^0 \rightarrow \Sigma^0\gamma$  decay asymmetries. *Phys. Lett. B*, **693**, 241.
- 5 Chikovani, G. *et al.* (1967) Evidence for a two-peak structure in the  $A_2$  meson. *Phys. Lett. B*, **25** (1), 44.
- 6 Bowen, D. *et al.* (1971) Measurements of the  $A_2^-$  and  $A_2^+$  mass spectra. *Phys. Rev. Lett.*, **26**, 1663.
- 7 Wietfeldt, F. and Norman, E. (1996) The 17 keV neutrino. *Phys. Rep.*, **273**, 149.
- 8 Danilov, M. and Mizuk, R. (2008) Experimental review on pentaquarks. *Phys. At. Nucl.*, **71**, 605.
- 9 Nakamura, K. *et al.* (2010) Review of particle physics. *J. Phys. G*, **37**, 075021.
- 10 BELLE Collab., Abashian, A. *et al.* (2001) Measurement of the  $CP$  violation parameter  $\sin 2\phi_1$  in  $B_d^0$  meson decays. *Phys. Rev. Lett.*, **86** (12), 2509.
- 11 BABAR Collab., Aubert, B. *et al.* (2001) Measurement of  $CP$ -violating asymmetries in  $B^0$  decays to  $CP$  eigenstates. *Phys. Rev. Lett.*, **86** (12), 2515.

**9****Theory Uncertainties***Markus Diehl***9.1  
Overview**

The computation of observables (like for example cross sections or decay widths) in high energy physics involves uncertainties that originate from rather different sources. Some of these are of statistical nature, others are not. Sources of theory uncertainties can be broadly divided into three categories:

- In general, the calculation of an observable relies on approximations. Many results are based on an expansion in a small parameter, such as a coupling or the inverse of a large momentum scale. To estimate a theoretical uncertainty in this case means to guess the size of uncalculated higher-order terms in the expansion. In other cases, results are based on model assumptions; the associated uncertainty can then sometimes be estimated by using an alternative model.
- Theory results typically depend on parameters of the Standard Model. Some parameters, like the fine-structure constant  $\alpha$  or the mass  $M_Z$  of the Z-boson, are known to very high precision. Others, however, can be an important source of uncertainty, such as the strong coupling constant  $\alpha_s$ , the quark masses  $m_b$  and  $m_t$ , or certain elements of the CKM matrix. These parameters are extracted from a suitable set of experiments; their uncertainties are in part due to the statistical and systematic uncertainties of the measurements and in part to the uncertainties in the theoretical formulæ used to extract the parameters.
- In addition, many theoretical expressions contain non-perturbative parameters or functions from QCD. The most prominent example at hadron colliders are parton distribution functions (PDFs), which themselves depend on  $\alpha_s$  via their scale evolution. Other examples are hadronic form factors or wave functions necessary to calculate particle decays like  $B \rightarrow D\ell\nu$  or  $B \rightarrow \pi K$ . Many of these quantities need to be extracted from measurements and thus again have experimental and theoretical uncertainties.

Certain non-perturbative quantities can be computed in lattice QCD, such as hadron masses, decay constants, but also the strong coupling constant  $\alpha_s$ . Mo-

ments of parton densities (see Section 9.5) can also be calculated, but not the parton densities themselves. Lattice results come with uncertainty estimates, both for the systematic uncertainties of the method and for the statistical errors due to the use of Monte Carlo integration. For a general overview of recent lattice results we refer to the proceedings in [1].

In this chapter we discuss a number of these issues in detail. We focus on QCD, which typically accounts for the largest uncertainties, both because it has the largest coupling constant among all gauge interactions and because it includes a non-perturbative sector where our ability to calculate is still rather limited. Electroweak interactions have largely become the domain of precision physics, in particular thanks to the measurements made at LEP and to the progress in perturbative computations. Our current knowledge of electroweak parameters in the Standard Model is summarised in Section 10 of [2].

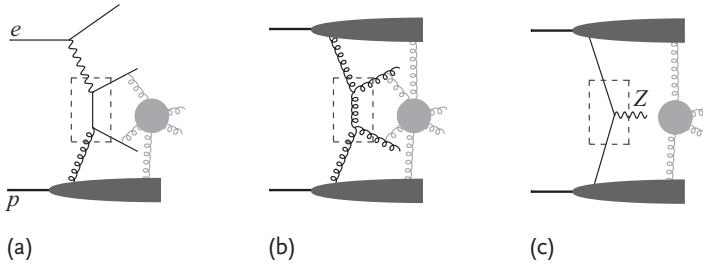
Understanding uncertainty estimates for theoretical calculations requires some understanding of the methods used in those calculations. Section 9.2 is devoted to the concept of factorisation and to the uncertainties of perturbative calculations in QCD. In Section 9.3 we briefly discuss the accuracy of the factorisation approach itself, and Section 9.4 deals with a number of aspects that are beyond the reach of this approach. Uncertainty estimates of parton densities have reached a stage of high sophistication and are described in detail in Section 9.5. We note that a detailed discussion of theoretical uncertainties, together with recommendations for estimating them, has recently been given in [3] for various Higgs-boson production channels at the LHC.

## 9.2

### Factorisation: A Cornerstone of Calculations in QCD

A characteristic feature of quantum field theory is that couplings such as  $\alpha$  or  $\alpha_s$  depend on a renormalisation scale  $\mu_R$ , which roughly speaking reflects the momentum scale at which particles couple to each other. It is the hallmark of QCD that  $\alpha_s$  increases as  $\mu_R$  becomes small, changing from  $\alpha_s(M_Z) \approx 0.118$  to  $\alpha_s(m_\tau) \approx 0.33$  between the masses of the  $Z$ -boson and the  $\tau$  lepton. At yet lower scales the perturbative expansion in  $\alpha_s$  eventually breaks down. Quantitative calculations in QCD largely rely on the concept of factorisation, which allows us to express suitable observables as a product involving a limited number of non-perturbative factors and a factor that is dominated by a high momentum scale  $Q$  and can be computed in perturbation theory. In the following we concentrate on the case where the non-perturbative factors are parton densities, which is most important for high-energy collisions and discussed in detail in [4, 5]. Related applications of the factorisation concept are for instance used to describe  $B$  meson decays and involve form factors and meson wave functions, see for example [6, 7].

Examples for processes that can be calculated using factorisation are jet production in electron–proton or proton–proton collisions,  $e p \rightarrow \text{jet} + X$  or  $p p \rightarrow$



**Figure 9.1** Example graphs for the amplitude of jet production in  $ep$  and  $pp$  collisions (a,b) and for  $pp \rightarrow Z + X$  (c). The half oval blobs at the top and the bottom of the graphs correspond to parton densities, and the dashed

boxes indicate the hard parton-level subprocesses. Gluons attached to the circular blobs are soft, and their effects cancel in inclusive cross sections described by a factorisation formula.

jet +  $X$ , or the production of a  $Z$ -boson,  $pp \rightarrow Z + X$ . Here  $X$  denotes all other particles produced in the collision. Feynman graphs corresponding to these processes are shown in Figure 9.1. For  $ep$  collisions, the corresponding cross-section formula has the form

$$\frac{d\sigma}{dx d\Phi} = \frac{1}{Q^d} f(\mu_F) \otimes_x C \left[ \Phi, \frac{\mu_F}{Q}, \frac{\mu_R}{Q}, \alpha_s(\mu_R) \right] + \frac{1}{Q^d} \mathcal{O} \left( \frac{A}{Q} \text{ or } \frac{A^2}{Q^2} \right), \quad (9.1)$$

where  $\mu_F$  denotes the factorisation scale and  $f(x, \mu_F)$  is a PDF of the proton. In  $pp$  or  $p\bar{p}$  collisions one has instead a product of the PDFs of the two colliding hadrons. A sum over the different parton types (quarks, antiquarks, gluons) has been omitted in (9.1) for brevity. The hard-scattering kernel  $C$  describes a parton-level process and is calculated in perturbation theory. In addition to the strong coupling constant  $\alpha_s$  it may depend on further parameters such as  $m_b$ ,  $M_Z$ ,  $\sin \theta_W$ , and so on. By ‘ $\otimes$ ’ we denote the convolution

$$f \otimes_x C = \int_x^1 \frac{dz}{z} f(z) C \left( \frac{x}{z} \right), \quad (9.2)$$

where  $x$  is a scaling variable formed from measurable kinematic quantities (e.g. the Bjorken variable in inclusive deep inelastic  $ep$  scattering, DIS). In addition, the cross section can depend on kinematic variables (energies, momenta, angles) of particles produced in the parton-level subprocess, which we collectively denote by  $\Phi$ . Both  $f$  and  $C$  are dimensionless by definition, so that the integer  $d$  in (9.1) is determined by the mass dimension of  $d\sigma/(dx d\Phi)$ .

An essential condition for the validity of factorisation is the presence of a large scale  $Q$ , which ensures that the parton-level subprocess is dominated by large virtualities of intermediate particles and thus involves a sufficiently small value of  $\alpha_s$ . Typically  $Q$  is a large mass, such as  $M_Z$  or  $m_t$  in the production of a  $Z$ -boson or a  $t\bar{t}$  pair, or a large transverse momentum as in jet production. The case of several hard

scales is discussed in Section 9.2.1.4. At a practical level, the hard scale  $Q$  plays two distinct roles in the factorisation formula (9.1). Firstly, it specifies the accuracy of the formula, which has corrections suppressed by some power of  $\Lambda/Q$  as indicated on the right-hand side of (9.1). Here  $\Lambda$  is a scale of order 1 GeV and represents the non-perturbative sector of QCD. These *power corrections* will be discussed further in Section 9.3. Secondly,  $Q$  determines the typical size of the renormalisation and factorisation scales,  $\mu_R$  and  $\mu_F$ , as explained in the next section.

Let us take a closer look at the scope and limitations of factorisation formulæ like (9.1):

1. One can calculate the production cross section for *inclusive* final states of the type  $p_1 + p_2 + \dots + p_n + X$ , where  $p_1$  to  $p_n$  are specified particles produced in the parton-level subprocess, whereas  $X$  denotes a set of unobserved particles including in particular the remnants of the initial hadrons. The variables in  $\Phi$  must only refer to  $p_1$  to  $p_n$ , and the formula does not give any details about the set of particles  $X$ .
2. The parton densities  $f(x, \mu_F)$  specify the longitudinal momentum fraction  $x$  of the partons entering the parton-level subprocess, but not their ‘intrinsic’ transverse momentum, the typical size of which is of order  $\Lambda$ . As a consequence, the kinematic information on transverse momenta in  $\Phi$  is accurate only up to effects of order  $\Lambda$ . For large transverse momenta, for instance of jets, this is a small correction. If one is interested in small transverse momenta, then one must use a more complicated theoretical formalism that involves transverse-momentum-dependent parton densities [5]. At present this formalism is applicable only to a limited number of processes, such as the production of a  $Z$  or a Higgs boson,  $p p \rightarrow Z + X$  or  $p p \rightarrow H + X$ .
3. If the particles  $p_1$  to  $p_n$  produced in the parton-level subprocess are colour-neutral (and decay into colour-neutral particles if they are unstable, such as  $Z \rightarrow \mu^+ \mu^-$  or  $H \rightarrow \gamma \gamma$ ), then one can directly measure them in the final state. This is, however, not the case for quarks and gluons, which do not exist as free particles and the conversion of which into observable hadrons is beyond the control of perturbation theory. Observables that are calculable in the factorisation approach must be such that they make sense both at the parton level and at the hadron level. A prime example are suitably defined hadronic jets, as will be further discussed in Section 9.4.2.

These restrictions are the price to be paid for the great simplicity and predictive power of the factorisation formula (9.1) and its counterparts for  $p p$  and  $p \bar{p}$  collisions. Partons in the initial and the final state (including the spectator partons in the incident hadrons) are subject to long-range, non-perturbative interactions which in the language of Feynman graphs are represented by the exchange and production of soft gluons as shown in Figure 9.1. An analysis in perturbation theory shows that such soft-gluon effects are present at the level of individual Feynman graphs but cancel in the factorisation formula (with a possible remainder at the level of corrections suppressed by a power of  $\Lambda/Q$ ). The sum over unobserved par-

ticles X in point 1) and the integral over intrinsic parton momenta in point 2) are essential to obtain this cancellation [5]. To make predictions for observables where soft-gluon effects do not cancel remains a challenge and typically requires the use of models for strong interactions in the non-perturbative regime.

Each of the three points above constitutes an important difference between factorisation formulæ and Monte Carlo event generators. At the expense of an increased model dependence, the latter give results for completely *exclusive* final states at hadron level, with every particle in the final state being fully specified. We will discuss this in more detail in Section 9.4.3.

### 9.2.1

#### The Perturbative Expansion and Uncertainties from Higher Orders

The hard-scattering kernel  $C$ , as well the derivatives  $d\alpha_s/d\mu_R$  and  $df(x, \mu_F)/d\mu_F$  have expansions in powers of  $\alpha_s$  with coefficients known up to a certain order. An important part of the theoretical uncertainty is the size of uncalculated higher orders in the perturbative series. In the two following subsections we will see to which extent these terms can be estimated from the dependence of  $C$  on the factorisation and renormalisation scales,  $\mu_F$  and  $\mu_R$ . These scales are respectively associated with collinear and with ultraviolet divergences in higher-order contributions to the hard scattering (see Figure 9.2), and one may choose them independently in (9.1).

##### 9.2.1.1 The Renormalisation Scale

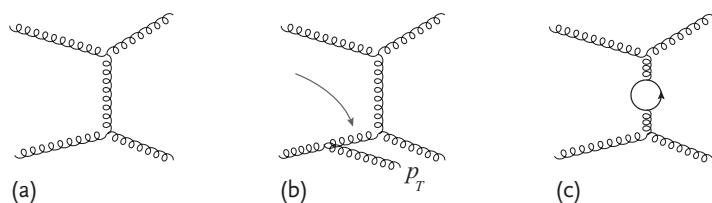
The running of the strong coupling is governed by the renormalisation group equation (see e.g. [8]),

$$\frac{d}{d \ln \mu_R^2} \alpha_s(\mu_R) = \beta(\alpha_s(\mu_R)), \quad (9.3)$$

where the renormalisation group function  $\beta$  has a perturbative expansion

$$\beta(\alpha_s) = -(b_0 \alpha_s^2 + b_1 \alpha_s^3 + b_2 \alpha_s^4 + b_3 \alpha_s^5 + \dots) \quad (9.4)$$

with coefficients  $b_i$  that depend on the number of active quark flavours,  $n_F$ . The coefficients  $b_0$  to  $b_3$  have been calculated [9]. In practice, one truncates the series (9.4)



**Figure 9.2** Parton-level graphs for jet production at leading (a) and next-to-leading order (b,c) in  $\alpha_s$ . The gluon line indicated by an arrow gives a collinear divergence when one integrates over  $p_T$ . The quark loop in graph (c) is ultraviolet divergent. The divergence in graph (b) is related with the factorisation scale  $\mu_F$  and the one in graph (c) with the renormalisation scale  $\mu_R$ .

at a given order and solves (9.3) analytically by expanding  $\alpha_s$  in inverse powers of  $\ln \mu_R$ , see for example [10]. Alternatively one can solve (9.3) numerically. In the rest of this subsection we write  $\mu$  instead of  $\mu_R$  for brevity. Depending on the order at which the series (9.4) is truncated one speaks of the running coupling at leading order (LO), next-to-leading order (NLO), next-to-next-to-leading order (NNLO), and so forth.

The coefficients in the  $\alpha_s$  expansion of a hard-scattering kernel  $C$  depend on  $\mu$ . If the expansion starts at order  $\alpha_s^k$ , we have

$$C\left(\frac{\mu}{Q}, \alpha_s(\mu)\right) = C_0 \alpha_s^k(\mu) + C_1 \left(\frac{\mu}{Q}\right) \alpha_s^{k+1}(\mu) + C_2 \left(\frac{\mu}{Q}\right) \alpha_s^{k+2}(\mu) + \dots \quad (9.5)$$

The kernel  $C$  is defined to be dimensionless, so that  $C_i$  depends on  $Q$  and  $\mu$  only via their ratio, and we suppress the dependence of  $C$  and  $C_i$  on all other variables. The  $\mu$  dependence of the coefficients  $C_i$  in the expansion (9.5) is such that to a given order in  $\alpha_s$  one has  $dC/d\mu = 0$ . This ensures that the cross section (9.1) does not depend on the unphysical parameter  $\mu$  to the accuracy at which it is calculated.

To see how this works let us first relate  $\alpha_s$  at the two scales  $\mu$  and  $Q$ ,

$$\alpha_s(Q) = \alpha_s(\mu) + a_1 \left(\frac{\mu}{Q}\right) \alpha_s^2(\mu) + a_2 \left(\frac{\mu}{Q}\right) \alpha_s^3(\mu) + \mathcal{O}(\alpha_s^4), \quad (9.6)$$

and then take the derivative with respect to  $\ln Q^2$  on both sides. On the right-hand side we get derivatives  $da_i/d\ln Q^2$ , whereas the left-hand side becomes

$$\begin{aligned} \beta(\alpha_s(Q)) &= -b_0 \alpha_s^2(Q) - b_1 \alpha_s^3(Q) + \mathcal{O}(\alpha_s^4) \\ &= -b_0 \alpha_s^2(\mu) - 2a_1 \left(\frac{\mu}{Q}\right) b_0 \alpha_s^3(\mu) - b_1 \alpha_s^3(\mu) + \mathcal{O}(\alpha_s^4) \end{aligned} \quad (9.7)$$

according to (9.4) and (9.6). Matching the coefficients of the expansion in  $\alpha_s(\mu)$ , we obtain a set of differential equations that are easily solved. We find

$$\begin{aligned} \frac{da_1}{d\ln Q^2} &= -b_0 & \Rightarrow \quad a_1 \left(\frac{\mu}{Q}\right) &= -b_0 L \left(\frac{\mu}{Q}\right), \\ \frac{da_2}{d\ln Q^2} &= -2a_1 b_0 - b_1 & \Rightarrow \quad a_2 \left(\frac{\mu}{Q}\right) &= -b_1 L \left(\frac{\mu}{Q}\right) + b_0^2 L^2 \left(\frac{\mu}{Q}\right), \end{aligned} \quad (9.8)$$

with  $L(\mu/Q) \equiv \ln(Q^2/\mu^2)$ . Increasing powers of  $\alpha_s(\mu)$  in the expansion (9.6) come with increasing powers of  $L$ ; this structure is characteristic of renormalisation group equations. The fixed-order relation (9.6) is hence reliable as long as  $L$  is not too large.

If we set  $\mu = Q$  in expression (9.5) and substitute  $\alpha_s(Q)$  using (9.6) and (9.8), we readily obtain, after a few lines of algebra,

$$\begin{aligned} C_1\left(\frac{\mu}{Q}\right) &= C_1(1) - k b_0 C_0 L\left(\frac{\mu}{Q}\right), \\ C_2\left(\frac{\mu}{Q}\right) &= C_2(1) - [(k+1)b_0 C_1(1) + k b_1 C_0]L\left(\frac{\mu}{Q}\right) \\ &\quad + \frac{k(k+1)}{2} b_0^2 C_0 L^2\left(\frac{\mu}{Q}\right), \end{aligned} \quad (9.9)$$

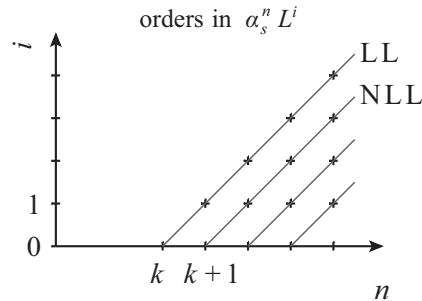
which the reader is encouraged to check. Notice that the coefficients of  $L$  and  $L^2$  grow with the order  $k$  at which the expansion of  $C$  begins, reflecting that the scale dependence is stronger when  $\alpha_s$  is raised to a higher power.

Our result (9.9) shows the interplay between the  $\mu$  dependence and different orders in the  $\alpha_s$  expansion. If one calculates  $C$  at LO, that is the term  $C_0 \alpha_s^k(\mu)$  in (9.5), then one also knows the terms with  $\alpha_s^k(\mu)[\alpha_s(\mu)L]^i$  at higher orders, which are called the leading logarithmic (LL) terms. Likewise, knowledge of the NLO coefficient  $C_1(1)$  gives us the next-to-leading logarithmic (NLL) terms  $\alpha_s^{k+1}(\mu)[\alpha_s(\mu)L]^i$ . This is illustrated in Figure 9.3. We note that NLO calculations are available for a large number of processes. The next order (NNLO) is known for a rather restricted set of reactions, and one order more ( $N^3LO$ ) for a few cases only.

Taking a different perspective, we see that if we vary  $\mu$  in the expression of  $C$  at order  $\alpha_s^n$ , the resulting variation of  $C$  corresponds to the higher-order terms

$$\alpha_s^{n+1}(\mu) \sum_{i=1}^{n+1-k} (\text{known coefficient}) \cdot L^i + \mathcal{O}(\alpha_s^{n+2}), \quad (9.10)$$

but contains no information about the term  $\alpha_s^{n+1}(\mu)C_{n+1}(1)$  which is not accompanied by a power of  $L$ . This makes it clear that a variation of the renormalisation scale reflects the size of higher-order terms in the perturbative series, but *not of all* such terms. It is customary to assess the uncertainty from missing higher orders



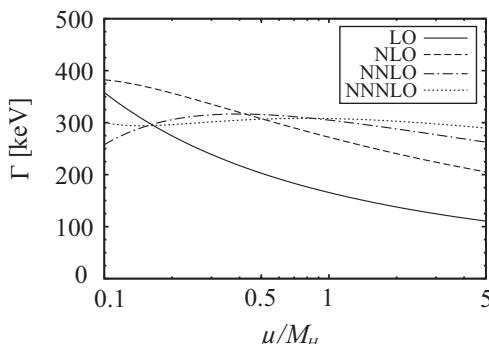
**Figure 9.3** Orders in  $\alpha_s^n(\mu)L^i$  appearing in the expansion of a hard-scattering kernel. The ratio of the expansion coefficients for any two points on a diagonal line is fixed by the renormalisation group equation, as explained in the text.

by varying  $\mu$  between 1/2 and 2 times a central value (these factors are not set in stone and other choices are sometimes made). Whether this gives a good estimate of the contributions from higher orders depends on whether the uncalculated coefficients  $C_{n+1}(1)$ ,  $C_{n+2}(1)$ , and so on are of similar size as the calculated ones,  $C_0$  to  $C_n(1)$ . When  $C_{n+1}$  is eventually calculated, one can check the reliability of the previous uncertainty estimate. There are cases where this estimate works well and where the higher-order result lies within the band resulting from varying the scale in the lower-order expression, and other cases where it does not.

As an illustration let us take the partial width  $\Gamma$  for the Standard Model Higgs boson decaying into hadrons via a top-quark loop (i.e. excluding graphs where the Higgs boson couples to  $b\bar{b}$  and lighter quarks). This quantity does not involve any parton densities or similar non-perturbative quantities and is given by a dimensionless hard-scattering kernel  $C$  times a factor depending on the Higgs mass,  $M_H$ , and Fermi's constant,  $G_F$ . It starts at order  $\alpha_s^2$  and has been calculated up to order  $\alpha_s^5$  (i.e. up to N<sup>3</sup>LO) in [11]. In Figure 9.4 we show  $\Gamma$  as a function of  $\mu$  for the successive perturbative approximations and see the substantial  $\mu$  dependence of the LO approximation, as well as its gradual decrease at higher orders. Both are quite typical of observables in QCD.

Before an uncertainty can be estimated by *varying* the scale, one has to choose a *central value* of  $\mu$ , which should be the best guess for this scale in the sense that higher-order corrections are small. As we have seen, higher-order terms in the perturbative expansion come with powers of  $\alpha_s(\mu) \ln(Q^2/\mu^2)$ . To keep this product small, one should take  $\mu$  of order  $Q$ . In practice this general rule leaves some freedom of choice, and one can, for instance, not decide a priori whether  $\mu = m_t$  or  $\mu = 2m_t$  is more appropriate for the production of a  $t\bar{t}$  pair.

There are more definite prescriptions in the literature for setting the renormalisation scale, going by the names of *fastest apparent convergence* (FAC) [12], *principle of minimal sensitivity* (PMS) [13] and *Brodsky–Lepage–Mackenzie* (BLM) prescrip-



**Figure 9.4** The partial width  $\Gamma$  for a Higgs-boson decay into hadrons via a top-quark loop, calculated at successive orders in  $\alpha_s$  and plotted as a function of the renormalisation scale  $\mu$  in units of  $M_H$ . Parameters used

are the pole masses  $M_H = 120$  GeV and  $m_t = 175$  GeV. The running coupling is taken at the same perturbative order as the observable, with  $\alpha_s(M_Z) = 0.1184$  in all cases.

tion [14]. Applying these prescriptions to lower-order expressions of observables for which higher-order terms are known, one finds cases where they work well and others where they do not. Note also that the uncertainty estimated by varying the scale around a central value depends on the choice of the central value. In particular, the principle of minimal sensitivity selects  $\mu$  such that  $dC/d\mu = 0$  for the perturbative series truncated at a given order, which tends to minimise the uncertainty estimated by scale variation.

The reader should be aware that there is no consensus in the community about an ‘optimal’ choice of the scale. Also, some authors argue that there is only one adequate scale choice and that hence the variation of the scale around a certain value does not give a useful estimate for higher-order contributions at all. Whichever point of view one takes on this issue, it should be clear from the material in this section that both setting and varying the renormalisation scale only relates to higher-order effects coming with renormalisation group logarithms  $L$ . One should not forget that the choice of scale is only *one* aspect of higher-order corrections. Other effects can be more important at the quantitative level and need to be investigated on a case-by-case basis. From a practical point of view, however, scale variation often provides the only readily available estimate of higher-order corrections (short of a full higher-order calculation) and should be better than no estimate at all, as long as one keeps in mind that the method is not fail-safe.

### 9.2.1.2 The Factorisation Scale

The role of the factorisation scale  $\mu_F$  in the cross section (9.1) is quite similar to the one of  $\mu_R$ . The scale dependence of PDFs is governed by the DGLAP evolution equation [15, 16],

$$\frac{d}{d \ln \mu_F^2} f(x, \mu_F) = f(\mu_F) \otimes_x P(\alpha_s(\mu_F)), \quad (9.11)$$

with the convolution defined in (9.2). The splitting functions  $P$  describe the splitting of one parton into several partons and have a perturbative expansion

$$P(z, \alpha_s(\mu_F)) = P_0(z) \alpha_s(\mu_F) + P_1(z) \alpha_s^2(\mu_F) + P_2(z) \alpha_s^3(\mu_F) + \dots \quad (9.12)$$

with coefficients known up to  $P_2$  [17, 18]. For simplicity we omit again explicit labels for the different parton types, as we did in formula (9.1). A set of PDFs  $f(x, \mu_F)$  is referred to as LO, NLO or NNLO depending on the order in  $\alpha_s$  used for its evolution.

The cross section (9.1) must not depend on the artificial scale  $\mu_F$  to any order in the perturbative expansion, that is  $d(f \otimes C)/d\mu_F = 0$ . This requirement implies

$$\begin{aligned} \frac{d}{d \ln \mu_F^2} C\left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}, \alpha_s(\mu_R)\right) \\ = -P(\alpha_s(\mu_F)) \otimes_x C\left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}, \alpha_s(\mu_R)\right). \end{aligned} \quad (9.13)$$

To solve this equation, we recall the perturbative expansion (9.5) of  $C$ ,

$$\begin{aligned} C\left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}, \alpha_s(\mu_R)\right) &= C_0 \alpha_s^k(\mu_R) \\ &+ C_1 \left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}\right) \alpha_s^{k+1}(\mu_R) + C_2 \left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}\right) \alpha_s^{k+2}(\mu_R) + \dots \end{aligned} \quad (9.14)$$

Here we have restored the  $\mu_F$  dependence, which was not displayed in the previous subsection. Using (9.6) with  $Q$  replaced by  $\mu_F$ , we can rewrite (9.12) as an expansion in  $\alpha_s(\mu_R)$ . Inserting this and the expansion (9.14) into (9.13), we can match the coefficients of  $\alpha_s^i(\mu_R)$ . This yields a set of differential equations for  $dC_i/d\ln\mu_F^2$ , which can be solved iteratively. For the first coefficient one obtains

$$C_1\left(\frac{\mu_F}{Q}, \frac{\mu_R}{Q}\right) = C_1\left(1, \frac{\mu_R}{Q}\right) + P_0 \otimes C_0 \ln \frac{Q^2}{\mu_F^2}, \quad (9.15)$$

and in the higher-order coefficients one finds higher powers of  $\ln(Q^2/\mu_F^2)$ , with a pattern very similar to the one we encountered in the previous subsection.

As in the case of  $\mu_R$ , the  $\mu_F$  dependence of the cross section calculated at a given order in  $\alpha_s$  is due to missing higher orders. The variation of  $\mu_F$  may be used to estimate higher-order terms that come with factorisation-scale logarithms (instead of the renormalisation-scale logarithms discussed in the previous subsection). Concerning the choice of a ‘best value’ for  $\mu_F$ , definite prescriptions similar to those we mentioned for  $\mu_R$  do not exist, so that one is left with the general guideline to take  $\mu_F$  of order of the hard scale  $Q$ . One often sets  $\mu_F = \mu_R$ , but one may also allow the two scales to be different since they are associated with logarithms that correspond to different kinematic regions in Feynman graphs (see Figure 9.2). An uncertainty estimate obtained by varying  $\mu_F = \mu_R$  by a factor between 1/2 and 2 has the virtue of simplicity, but independent variation of the two scales may be regarded as a better estimate of higher-order contributions.

To conclude this section let us give some examples for the uncertainty obtained by scale variation between 1/2 and 2 times a central value.

### Example 9.1 Cross-section prediction for $p p \rightarrow t\bar{t}H + X$

As an example of a process the cross section of which starts at order  $\alpha_s^2$ , we take  $p p \rightarrow t\bar{t}H + X$ . With  $\sqrt{s} = 7$  TeV,  $M_H = 120$  GeV and independent variation of  $\mu_F$  and  $\mu_R$ , a scale uncertainty between -26.2% and +39.8% is quoted in [3] for the cross section calculated at LO. At NLO this uncertainty decreases to the range between -9.4% and +3.5%. The rapidity distributions for inclusive  $Z$  and Higgs-boson production at the LHC have been calculated in [19] and [20], with uncertainty bands obtained by joint scale variation of  $\mu_R = \mu_F$ . In both cases one finds that the uncertainty band for the LO cross section does not contain the NLO result at central rapidities, whereas the bands of the NLO and the NNLO cross sections do overlap.

### 9.2.1.3 Combining Different Orders

As a rule, one calculates cross sections with the scale dependence of the parton densities and of  $\alpha_s$  truncated at the same perturbative order as the hard-scattering kernel  $C$ . In some situations one may however wish to take PDFs or the running coupling at a higher order than is available for the hard-scattering subprocess. This holds especially for final states with high multiplicity, say with six jets, where  $C$  is only known at LO. Such a combination is not necessarily inconsistent, as long as one is aware that the overall accuracy is determined by the lowest order used for the different ingredients in the factorisation formula. Certain combinations should be avoided, for instance taking parton densities that have been fitted to inclusive ep structure functions evaluated at a certain perturbative order and using them to calculate the same structure functions with  $C$  truncated at a different order. In less clear-cut cases, different practitioners may take different points of view as to whether a combination of different orders is sensible or not.

### 9.2.1.4 Multi-Scale Problems and Resummation Methods

In the previous two subsections we saw that the perturbative expansion of parton-level cross sections contains powers of  $\alpha_s \ln(Q/\mu_R)$  and  $\alpha_s \ln(Q/\mu_F)$ , where  $Q$  ‘represents the hard scale’ of the scattering process. However, many reactions involve *several* hard scales, for example the photon virtuality and  $m_b$  for bottom production in deep inelastic scattering, or the mass and the transverse momentum of the W-boson in  $p p \rightarrow W + X$ . For such multi-scale problems it is even less obvious which choice of scales  $\mu_R$  and  $\mu_F$  keeps logarithms at higher orders reasonably small. Often one can try and identify a ‘typical virtuality’ in higher-order Feynman graphs, but there is neither a general nor a fail-safe method for this.

If a process involves a small ratio  $r$  of momentum or mass scales (which is often reflected in scaling variables  $x$  being close to 0 or 1), it can happen that higher-order corrections come with one or even two powers of  $\ln r$  for each power of  $\alpha_s$ . In such cases *no* choice of  $\mu_R$  and  $\mu_F$  can eliminate all large logarithms at higher orders. Perturbative results truncated at a fixed order in  $\alpha_s$  can then be unreliable and must be complemented by ‘resumming’ the large logarithms to all orders. Corresponding resummation procedures have been worked out for several types of logarithms.

*Sudakov logarithms* come with two powers of  $\ln r$  for each power of  $\alpha_s$ . They appear in different contexts when gluon radiation is inhibited by the selection of kinematics. The small ratio  $r$  can be of the form  $p_T/Q$ , where  $Q$  is the mass or virtuality of a particle and  $p_T$  is its transverse momentum. Resummation of these *recoil* or *transverse-momentum logarithms* is, for instance, important for the production of a Z or a Higgs boson in  $p p$  collisions. *Threshold logarithms* appear when the squared centre-of-mass energy  $\hat{s}$  of the parton-level collision is close to the squared invariant mass  $Q^2$  of the produced particle; the small ratio is then  $r = (\hat{s} - Q^2)/\hat{s}$ . It is also possible to perform a joint resummation of recoil and threshold logarithms. More information and references can be found in [21].

*High-energy logarithms* appear when the overall squared collision energy  $s$  is much larger than any other scale in the process. They come in powers of  $\alpha_s \ln x$ , where the scaling variable  $x$  is inversely proportional to  $s$ . The resummation of these logarithms is performed with the help of the BFKL equation, which describes the change of scattering amplitudes with  $s$ . Physically, the logarithms are connected to the growth of gluon radiation with energy and to the rise of the gluon density in the proton at small momentum fractions. We return to this topic in Section 9.3.2.

The practical implementation of resummation and its matching with fixed-order perturbative results typically requires some theoretical choices, which may be regarded as sources of theoretical uncertainties. As an example, some formulations include integrals over the running coupling of the form  $\int d\mu \alpha_s(\mu) F(\mu)$ , which require regularisation in the region where  $\mu$  becomes small and the perturbative expression for  $\alpha_s(\mu)$  is divergent.

### 9.3

#### Power Corrections

The computation of perturbative corrections in  $\alpha_s$  (and in electroweak coupling) has achieved a high level of sophistication, with higher-order results being available for many processes and observables. In contrast, the evaluation of corrections that are suppressed by an inverse power of a large mass or momentum scale is only possible in a quite restricted number of cases. Such an evaluation may depend on several poorly known non-perturbative parameters. It is then not very precise, but still offers the possibility to estimate the theoretical uncertainty due to the power corrections that affect any factorisation formula like (9.1). We do not attempt a systematic overview of this topic here, but rather mention a few important examples.

##### 9.3.1

#### Operator Product Expansion

A systematic framework for computing power-suppressed terms is the *operator production expansion* (OPE). This method is very powerful, but it can only be applied to sufficiently inclusive observables (and not to cross sections or decay rates that are differential in some variable or involve cuts on the final state). In short, the method works by writing an observable as a matrix element of two operators separated by a small distance (which is the Fourier conjugate of a large mass or momentum). Expanding the operator product gives a representation in terms of local operators, the matrix elements of which need to be determined as a non-perturbative input. These operators are classified by their ‘twist’, with operators of the lowest twist giving the leading contribution to an observable and operators of higher twist giving power corrections. Power corrections are therefore often referred to as *higher-twist corrections*.

An important application of this method is the description of inclusive hadronic  $\tau$  decays, where one has (see Section 9.4 of [22])

$$R_\tau = \frac{\Gamma(\tau \rightarrow \nu_\tau + \text{hadrons})}{\Gamma(\tau \rightarrow \nu_\tau + e\bar{\nu}_e)} = R_0 \left[ 1 + \frac{\alpha_s}{\pi} + 5.2 \frac{\alpha_s^2}{\pi^2} + 26.4 \frac{\alpha_s^3}{\pi^3} + c_2 \frac{m^2}{m_\tau^2} + c_4 \frac{\langle m\bar{\psi}\psi \rangle}{m_\tau^4} + c_6 \frac{\langle \bar{\psi}\psi\bar{\psi}\psi \rangle}{m_\tau^6} \right]. \quad (9.16)$$

Here  $m^2$  is a combination of squared light-quark masses, and  $\langle m\bar{\psi}\psi \rangle$  and  $\langle \bar{\psi}\psi\bar{\psi}\psi \rangle$  are vacuum expectation values of local quark operators. With phenomenological estimates for these quantities and for the constants  $c_2$ ,  $c_4$ ,  $c_6$ , the measured value of  $R_\tau$  has been used to extract the strong coupling at  $\mu_R = m_\tau$ .

The OPE also gives predictions for power-suppressed effects in inclusive decays  $B \rightarrow \ell\nu + X$ , where the large scale is  $m_b$ . This is in particular used for the extraction of the CKM matrix element  $|V_{cb}|$ , see p. 1017 in [2].

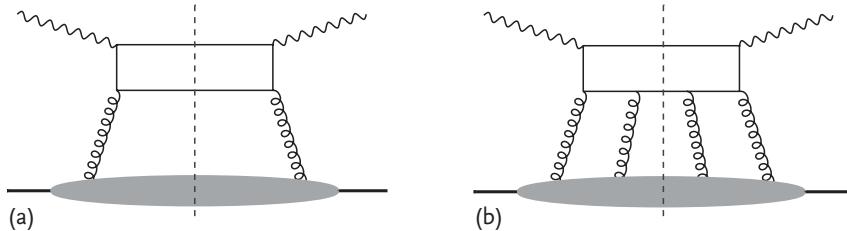
A classic field of application for the OPE are sum rules in deep inelastic scattering, that is integrals of structure functions over the Bjorken variable. For many of these sum rules, contributions suppressed by  $1/Q^2$  have been computed [23], where  $-Q^2$  is the squared momentum transfer to the lepton.

### 9.3.2

#### Power Corrections in Cross Sections

Factorisation formulæ for cross sections generically have power-suppressed corrections, as indicated in (9.1). These corrections have only been computed for a restricted set of processes, the most prominent one being inclusive deep inelastic scattering [24, 25]. Power-suppressed terms are then expressed in terms of higher-twist distributions that generalise the parton densities in (9.1). An example graph for deep inelastic scattering is given in Figure 9.5b, where four instead of two gluons from the proton enter the hard-scattering subprocess. Higher-twist distributions depend on several parton momentum fractions, and a large number of them contributes to a given process. They have barely been used in phenomenology so far.

A particular case in this context is the small- $x$  region already mentioned in Section 9.2.1.4. The graph in Figure 9.5b is suppressed by  $1/Q^2$ , but it grows faster with decreasing  $x$  than the one in Figure 9.5a. Certain power corrections are thus enhanced in small- $x$  kinematics. In the BFKL or high-energy factorisation approach, the small variable used for justifying approximations is in fact  $x$ . This is different from the hard-scattering factorisation approach described in Section 9.2, where the small expansion parameter is  $1/Q$ . In the BFKL approach, a sufficiently large scale  $Q$  is still required to justify the use of QCD perturbation theory, but calculations are performed at leading or next-to-leading order in  $\ln x$ . An additional expansion in powers of  $1/Q$  is not necessary, although it is made in certain cases.



**Figure 9.5** Graphs for the cross section of inclusive deep inelastic scattering. Graph (a) involves the usual gluon density, whereas the power-suppressed contribution in graph (b) comes with a higher-twist distribution. The

dashed vertical line denotes the final-state cut, that is the part of the graph to its left represents the scattering amplitude and the part to its right represents the complex conjugate amplitude.

An introduction to the BFKL approach can be found in [26], and an overview of theoretical developments in small- $x$  physics in [27, 28].

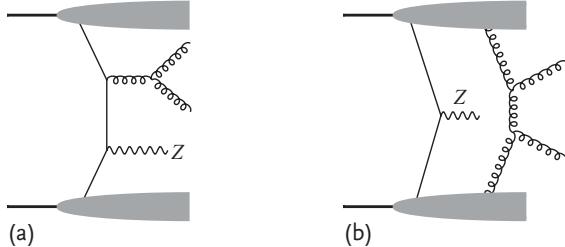
Whether a theory approach based on the large- $Q$  or on the small- $x$  limit is more appropriate in a given physical situation cannot be easily determined and is sometimes controversial. Overall, it is fair to say that at present the small- $x$  approach can barely compete with the large- $Q$  expansion as far as the availability of higher-order results in  $\alpha_s$  is concerned, whereas in many cases the large- $Q$  approach is not very predictive when power corrections become important. To estimate theoretical uncertainties in this context is not easy. For a number of processes, predictions are available from event generators based on the small- $x$  approach (e.g. from the CASCADE generator [29]) and can be compared with the results of general-purpose programs based on hard-scattering factorisation (see Section 9.4.3). The difference between the results can give an indication of the uncertainties related to the two approaches, but care needs to be taken because they may also originate from more mundane sources, for instance from the degree to which free parameters in the generators have been tuned to data.

## 9.4 The Final State

For most experimental analyses, a rather detailed understanding of the final state produced in a collision is necessary, both for the signal and background processes. Some aspects of this broad field are discussed in the following.

### 9.4.1 Underlying Event and Multi-Parton Interactions

Standard factorisation formulæ such as (9.1) describe the production of a specified final state by a single collision at parton level, along with unobserved particles. As already noted in Section 9.2, the detailed dynamics of a collision are more com-



**Figure 9.6** Graphs for  $p p \rightarrow Z + 2 \text{ jets} + X$  by a single (a) or by two (b) hard parton-level interactions. The half oval blobs in graph (b) represent two-parton distribution functions.

The gluons emitted by the protons in graph (b) produce jets and therefore carry large energy, unlike the soft gluons in Figure 9.1c.

plicated because partons are subject to soft interactions among themselves. In  $p p$  collisions this implies in particular that the hard interaction between two partons (which is described by the hard-scattering kernel) is accompanied by further soft collisions between the spectator partons in the two incoming protons. The particles produced in these extra collisions, together with the remnants of the beam hadrons, are often referred to as the *underlying event* and can have an important impact on the characteristics of the final state. To describe the underlying event, one has to resort to models which are typically implemented in event generators.

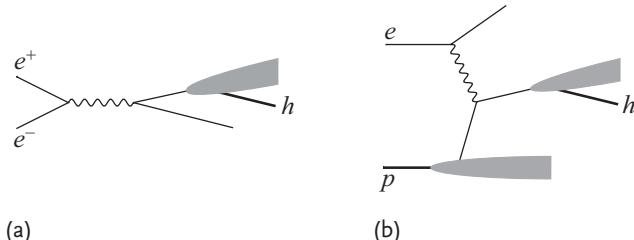
At sufficiently high energies, in particular at the LHC, one can also have several *hard* parton-level processes in one and the same  $p p$  collision. Such *multiple partonic* or *multi-parton interactions* are relevant for final states with high parton multiplicity. An example is the production of a Z-boson associated with two jets, as illustrated in Figure 9.6. Double hard scattering is only possible in the kinematic region where the Z has a small transverse momentum, but in that region the single and double hard-scattering contributions can have comparable size [30]. The theoretical description of multiple hard scattering is a challenging and active area of research [31, 32]. It involves multi-parton distributions in the proton, which are largely unknown. Simple models, as implemented in event generators, assume the absence of correlations between different partons in a proton and essentially replace multi-parton distributions by a product of single-parton densities.

For the effects of the underlying event and of multiple hard interactions, one can estimate a theoretical uncertainty by comparing the results obtained with different event generators. A more cautious procedure would be to compare results obtained with these effects being switched on or off, even though their complete absence is barely plausible and in contrast with several experimental findings [33].

#### 9.4.2

##### From Partons to Hadrons

A fundamental problem in describing particle production in high-energy collisions is that factorisation approaches (both in the large- $Q$  and in the small- $x$  limit) cal-



**Figure 9.7** Graphs for  $e^+e^- \rightarrow h + X$  at large centre-of-mass energy (a) and for  $ep \rightarrow e + h + X$  at large momentum transfer to the electron (b). The blobs with an emerging hadron  $h$  represent fragmentation

functions. Additional subgraphs for soft-gluon exchange as in Figure 9.1 are not shown for simplicity; their effects average out in the one-particle inclusive cross section with integrated transverse momenta.

culate the production of partons and not of hadrons. Only for special observables, typically for one-particle or two-particle inclusive distributions, can one use factorisation formulæ that involve fragmentation functions, which describe the transition  $\text{parton} \rightarrow \text{hadron} + \text{anything}$  and are thus close analogues of parton densities, which describe the transition  $\text{hadron} \rightarrow \text{parton} + \text{anything}$ . Examples of processes where fragmentation functions appear are shown in Figure 9.7, and an overview about our current knowledge of these functions is given in Section 17 of [2]. In more general cases one has to resort to event generators, which describe the transition from partons to the final-state hadrons by hadronisation models (see Section 9.4.3).

Although partons produced in high-energy processes do not appear as particles in the final state, they leave an imprint in the form of hadronic jets, which – roughly speaking – are sets of collimated hadrons with total momentum close to the momentum of the parton they originate from. Jets need to be defined by a precise algorithm, and it generally depends on the situation at hand which type of algorithm is best suited for a given purpose. For an overview and discussion of jet-finding algorithms for LHC physics we refer the reader to [34]. An important criterion for any jet algorithm is that it should be infrared and collinear safe, which means that it should be insensitive to the emission of soft particles and to the splitting of a fast particle into two fast particles with a small opening angle. This ensures that the jet definition is insensitive to the part of the phase space where perturbative calculations at parton level break down, and thus minimises the effects of the non-perturbative dynamics of hadronisation. As a first approximation, one can then directly compare a perturbative calculation (with the jet algorithm applied to the produced partons) and experimental data (where the same algorithm is applied to the observed hadrons)<sup>1)</sup>. In a second step one may apply *hadronisation corrections* to the perturbative result using the hadronisation models implemented in event generators.

- 1) Jet observables defined by algorithms that are not infrared and collinear safe cannot even be calculated in perturbation theory beyond tree level, since soft and collinear gluon radiation would give infinite results.

The reverse procedure of applying hadronisation corrections to data in order to obtain parton-level results is highly problematic since it can impart a strong model dependence on the measurement. Also, ‘parton-level results’ obtained with one generator cannot easily be compared with predictions from another generator that uses a different approach to describe perturbative radiation and non-perturbative hadronisation. In the words of [34], ‘experiment and theory should meet at the level of an observable.’

A particular class of observables are *hadronic event shapes* such as *thrust*, which are functions of the four-momenta of the hadrons produced in an event. At sufficiently high momentum scales they can be calculated in perturbation theory, provided that they are infrared and collinear safe. Resummation of large logarithms as discussed in Section 9.2.1.4 is often required. The effects of hadronisation give power corrections to the perturbative result, and there is a number of analytic approaches for estimating them [35, 36]. A recent application of this framework is the determination of  $\alpha_s$  from event shapes in  $e^+ e^-$  annihilation [37].

#### 9.4.3

##### Monte Carlo Event Generators

Monte Carlo event generators have become an indispensable tool for the analysis of high-energy collisions. There are several general-purpose generators that give results for a wide range of final states. A general introduction to event generators can be found in [38], and a detailed description of the generators **HERWIG**, **PYTHIA** and **SHERPA** in [39]. To establish a connection with the material of the preceding sections, let us briefly sketch their underlying physics model here.

The dynamical picture implemented in the event generators just mentioned is closely connected with the concept of hard-scattering factorisation described in Section 9.2, with the important difference that an event generator gives predictions for all details of the final state, specifying all particles and their momenta. Schematically, events are generated with a probability determined by the following ingredients:

1. a hard parton-level process, with incoming partons distributed according to parton densities that are provided as an external input (except of course in  $e^+ e^-$  annihilation). This part of the event generation closely follows the factorisation formula (9.1) and its analogues for  $p p$  and  $p \bar{p}$  collisions;
2. parton showers, which describe the radiation of additional partons in the kinematical region where perturbation theory is applicable. Depending on whether the radiating parton enters the hard-scattering process or is produced by it, one speaks of initial-state radiation or final-state radiation. Initial-state parton showers are closely related to the physics described by the DGLAP evolution equation of parton densities;
3. an infrared cutoff that has to be imposed on the parton showers, in order to remain in the region where perturbation theory is reliable. For the dynamics below this infrared cutoff, a hadronisation model is used to describe the transi-

tion from the partons generated in the previous two steps to the hadrons in the final state.

For hadron–hadron collisions, these ingredients are supplemented with models for the underlying event and for multi-parton interactions, plus possibly a model for purely soft interactions.

The matching of steps (1) and (2) requires care in order to prevent double-counting of partons produced by the hard process and by the parton shower. This is relatively easy for hard  $2 \rightarrow 2$  scattering processes calculated at leading order in  $\alpha_s$ , but it becomes more involved when one considers a hard scattering that produces more than two partons, and when calculating the hard process at NLO.

An essential input for step (3) is the detailed colour configuration of the partons before hadronisation. A prescription for their colour assignment needs to be given and is sometimes referred to as *colour reconnection*. We note that the string-fragmentation model in PYTHIA uses ‘fragmentation functions’ as an input, for which several options are available. These fragmentation functions should not be confused with their namesakes discussed in Section 9.4.2, since despite their apparent resemblance they are conceptually different quantities. The former describe the fragmentation of a colour string and are part of a hadronisation model, whereas the latter are universal functions describing the fragmentation of quarks, antiquarks or gluons, with a definition in QCD and a scale dependence given by DGLAP equations akin to those for parton densities.

Naturally, the model dependence in step (3) is larger compared with steps (1) and (2), which are constructed based on perturbation theory and kinematic approximations. Which aspects are responsible for the uncertainty in the predictions of a generator depends, however, on the process and observable in question. To estimate this uncertainty one may compare different generators or different model options and parameter settings (also called *tunes*) of the same generator. Uncertainties due to parton distributions can be evaluated separately by using different sets of PDFs. One should take care to use only generator tunes and parton densities that have been validated against data and are not deprecated – otherwise the comparison will reflect the progress of the field rather than the uncertainty in the best available theory predictions.

## 9.5

### From Hadrons to Partons

Parton densities are a crucial input for calculating reaction cross sections in lepton–hadron and hadron–hadron collisions. Their moments, that is integrals of the form  $\int dx x^n f(x, \mu_F)$ , can be calculated in lattice QCD for low  $n$  (typically from  $n = 0$  to  $n = 2$ ). To compute cross sections with a factorisation formula of the type (9.1), one however needs the PDFs as functions of  $x$  and in practice has to extract them from data. In Table 9.1 we list a number of recent PDF determinations, along with some of their characteristics to be discussed in the following. A more detailed overview of

**Table 9.1** Recent PDF sets.  $\mu_0$  is the starting scale of evolution, and  $Q_{\min}^2$  is the minimal value of  $Q$  required for DIS data in a fit (this value may be higher for individual subsets

of the data). The values of  $m_c$  and  $m_b$  are understood to be the pole masses. Further explanations of the entries are given in the text.

PDF set	Order	Fitted PDF parameters	$\mu_0^2$ (GeV $^2$ )	$Q_{\min}^2$ (GeV $^2$ )	Value	$\alpha_s(M_Z)$	Comment
JR09 [44]	NNLO	20	0.55	2	0.1124(20)		Fitted
ABKM09 [45]	NNLO	21	9	2.5	0.1135(14)		Fitted
MSTW08 [46, 47]	LO	28	1	2	0.139		Fixed
	NLO				0.120		
	NNLO				0.117		
HERAPDF1.0 [48]	NLO	10	1.9	3.5	0.1176		Fixed
CT10 [49]	NLO	26	1.69	4	0.118		Fixed
NNPDF2.1 [50, 51]	LO	259	2	3	0.119, 0.130		Fixed
	NLO				0.119		
	NNLO				0.119		

PDF set	$m_c$ (GeV)	$m_b$ (GeV)	Heavy flavour scheme	Tolerance $T$	
				68% CL	90% CL
JR09	1.3	4.2	FFNS ( $n_F = 3$ )	4.54	
ABKM09	1.5	4.5	FFNS ( $n_F = 3$ )	1	
MSTW08	1.4	4.75	GM-VFNS	$\approx 1$ to 6.5	$\approx 2.5$ to 11
HERAPDF1.0	1.4	4.75	GM-VFNS	1	
CT10	1.3	4.75	GM-VFNS		10
NNPDF2.1	1.414	4.75	GM-VFNS	—	—

modern PDFs and their uncertainties can for example be found in [40, 41]. Useful web resources for obtaining numerical values of PDFs are [42, 43].

The principle of PDF fits is to make an ansatz for the parton densities  $f(x, \mu_0)$  at a reference scale  $\mu_0$ , also called the ‘starting scale’. The DGLAP evolution equation then gives  $f(x, \mu_F)$  at other scales  $\mu_F$ , and the resulting PDFs are used to compute cross sections that are compared with data. For conventional PDF fits, a functional form for the distributions  $f(x, \mu_0)$  is chosen, and the parameters of these functions are determined in a  $\chi^2$  fit to the data. An alternative procedure, pursued by the NNPDF collaboration and explained in Section 9.5.1.3, uses functions called *neural networks* to represent  $f(x, \mu_0)$  and determines their parameters using a Monte Carlo method. Neural networks are described in more detail in Section 5.3.4.

An important feature of the DGLAP equation is that fixing a PDF for  $x \geq x_0$  at one scale  $\mu_0$  fixes the PDF for  $x \geq x_0$  at any other scale, see (9.11) and (9.2). Likewise, the convolution in the factorisation formula (9.1) shows that a cross section measured at scaling variable  $x_0$  is sensitive to parton distributions with  $x \geq x_0$ . This implies that data in a given kinematic region are only sensitive to (and can only constrain) PDFs with  $x$  above a certain value, but not below.

PDF extractions differ in the choice of data included in the fit. The backbone of all extractions are inclusive cross-section data for deep inelastic scattering, measured in  $e p$  collisions at HERA and in fixed target experiments with electron, muon or neutrino beams. Further relevant processes are charm and jet production in DIS, Drell–Yan lepton-pair production in fixed-target hadron–hadron collisions, and  $W$ ,  $Z$  and jet production in  $p \bar{p}$  collisions at the Tevatron. In the future, data from  $p p$  collisions at the LHC will play a prominent role. The MSTW08, CT10 and NNPDF2.1 sets in Table 9.1 are based on fits using the largest number of datasets, JR09 and ABKM09 have more restricted data selections, and HERAPDF1.0 uses only HERA data. Note that including more data in a fit does not automatically lead to more reliable PDFs: if a particular measurement has errors that are not reflected by its quoted uncertainties, or if the theory for a particular process is less precise, the resulting PDFs may turn out to be less rather than more accurate.

Parton densities extracted from data are affected by two types of uncertainty. First, the data to which one fits the PDFs have statistical and systematic uncertainties, which propagate to uncertainties on the parameters describing the PDFs. Modern PDF sets include such parametric uncertainties, which are discussed in Section 9.5.1. Second, there are various uncertainties in the theory used to connect data and PDFs:

- The perturbative order used in the evolution of the PDFs and in the cross-section formulæ. For most modern PDF sets this is NLO or NNLO, whereas PDF sets with LO precision are less frequently provided. For some observables, resummed calculations as discussed in Section 9.2.1.4 may be used.

Concerning the use of different orders in the PDFs, in  $\alpha_s$  and in the factorisation formulae we refer to Section 9.2.1.3. We note in particular that, while full NNLO calculations are available for the evolution of PDFs and for the inclusive cross sections of DIS and the Drell–Yan process, the same is not true for jet production. The NNLO fits of MSTW08 and NNPDF2.1 include jet-production data described at NLO with partial NNLO corrections for the threshold region [46, 51];

- The value of the strong coupling constant  $\alpha_s$  (which is conventionally quoted at the reference scale  $M_Z$ ). In some PDF fits, such as JR09 and ABKM09 in Table 9.1,  $\alpha_s$  is among the parameters fitted to the data, that is it is treated just like a parameter of the input PDFs. The corresponding fit uncertainty is included in the PDF uncertainties. The other PDF sets in the table are provided for several values of  $\alpha_s$ , with particular choices being preferred as indicated. In the MSTW08 set these preferred values are determined by fitting to data but kept fixed when determining the parametric errors of the PDFs. Further discussion of the  $\alpha_s$  values extracted from PDF fits and their comparison with other results can be found in [52, 53].

Note that there is a strong correlation between the value of  $\alpha_s$  used in a PDF fit and the resulting gluon distribution  $g(x)$ . This is because gluons are only subject to strong interactions, so that  $g(x)$  enters cross sections multiplied with at least one factor of  $\alpha_s$ . In contrast, quarks and antiquarks can directly couple

to a photon,  $W$  or  $Z$ , so that corresponding cross sections can be independent of  $\alpha_s$  at lowest order;

- The values of the heavy quark masses  $m_c$  and  $m_b$  (the value of  $m_t$  is of lesser importance since it is much larger than the typical scale of most of the data that dominate PDF fits). For all PDF sets given in Table 9.1, the fits are performed for fixed values of the quark masses. The ABKM09 and HERAPDF1.0 sets include a certain variation of these values when determining the PDF uncertainties, as respectively explained in [45] and [48]. Both MSTW08 and NNPDF2.1 provide PDF sets obtained with a number of alternative values of  $m_c$  and  $m_b$ .  
As noted in [54] the values of  $m_c$  used in current PDF fits tend to be systematically lower than what is obtained from dedicated determinations of the charm quark mass. With future theory improvements (notably for charm production in DIS) this tension will hopefully disappear;
- Different schemes used to treat heavy quarks. For definiteness let us consider charm; an analogous discussion can be given for bottom or top quarks.
  - In the *fixed flavour number scheme* (FFNS) only gluons and the light quarks  $u, d, s$  are treated as partons and assigned PDFs. The production of charm quarks is calculated in the hard-scattering process via their coupling to gluons (and to strange quarks via the weak current). This scheme is most suitable when the hard scale  $Q$  of the process is of order  $m_c$ . It fails for  $Q \gg m_c$  because the fixed-order hard-scattering kernels miss large logarithms  $[\alpha_s \ln(Q/m_c)]^i$  from higher orders.
  - In the *zero-mass variable flavour number scheme* (ZM-VFNS), charm is treated as a massless parton, with a PDF that is set to zero at a certain scale (typically  $\mu_F = m_c$ ) and obtained by DGLAP evolution at higher scales. The logarithms just mentioned are resummed by the evolution. This scheme is adequate for processes with  $Q \gg m_c$  such as the production of charm at high transverse momentum. Since the charm mass is neglected, the scheme is inadequate for  $Q \sim m_c$ .
  - *General-mass variable flavour number schemes* (GM-VFNS) aim to interpolate smoothly between the two extremes just described. Such schemes involve a number of technical choices, like the treatment of  $m_c$  in kinematic variables and the transition between the descriptions with three and with four parton flavours. Different GM-VFNS schemes are in use, specified by acronyms such as ACOT, TR, FONLL, BMSN, and so on. A detailed discussion of these issues can be found in [45, 47, 50, 55] and a numerical comparison of different schemes in Section 22 of [56].
- Power corrections. Simple approaches for power-suppressed terms are sometimes included in the cross-section formulæ, with parameters adjusted to data. However, many extractions attempt to minimise the importance of power corrections by restricting the kinematic range of the fitted data. This holds in particular for DIS, where experimental measurements often include rather small values of the squared momentum transfer  $-Q^2$  to the lepton (which should provide the hard scale in the factorisation formula). Note that the value  $Q_{\min}$  above

- which data are included in a fit does in general not coincide with the starting scale  $\mu_0$  of evolution (see Table 9.1);
- The use of nuclear targets in certain measurements, which requires corrections for nuclear effects to relate PDFs in the nucleus with those in the nucleon.

A possible bias in PDF fits can come from the functional form chosen for the input distributions  $f(x, \mu_0)$ , as well as from relations assumed between the input distributions for different quarks or antiquarks. The latter point in particular concerns strangeness. Among the sets in Table 9.1 only MSTW08 and NNPDF2.1 allow  $s(x, \mu_0)$  and  $\bar{s}(x, \mu_0)$  to differ, whereas all others impose  $s(x, \mu_0) = \bar{s}(x, \mu_0)$ . The HERAPDF1.0 set additionally constrains the shape of the strangeness distribution to be  $\bar{s}(x, \mu_0) = c_s \bar{d}(x, \mu_0)$  with a predefined value of  $c_s$ . This set provides explicit ‘model and parameterisation uncertainties’, which are obtained by variation of the parameters  $\mu_0$ ,  $Q_{\min}$ ,  $c_s$  and of the quark masses  $m_c$ ,  $m_b$ , as well as by fits with 11 instead of 10 free parameters describing the PDFs. One of the main aims of the NNPDF approach is to avoid a parameterisation bias by taking very flexible forms of the input distributions, namely neural networks. This leads to a large number of free parameters (see Table 9.1) which are not determined by a conventional  $\chi^2$  fit but by a Monte Carlo method.

### 9.5.1

#### Parametric PDF Uncertainties

Let us now discuss in more detail the uncertainties on the parameters determined in PDF fits. These are often just called ‘PDF uncertainties’ or ‘PDF errors’ and in all modern determinations are provided together with the fitted parton densities. The HERAPDF1.0 set refers to these uncertainties as ‘experimental’ since their origin are the uncertainties in the fitted data, and additionally provides ‘model’ and ‘parameterisation’ uncertainties. Adding those in quadrature gives the total PDF uncertainty of this set.

As noted in the previous subsection, the JR09 and ABKM09 fits determine  $\alpha_s$  together with the PDFs. In these cases the parametric errors on the PDFs include the uncertainty of the strong coupling. The other fits in Table 9.1 provide parton distributions for a range of  $\alpha_s$  values, which can be used to compute uncertainties due to varying  $\alpha_s$  in a given range. Detailed recommendations for how to do this for the different sets and how to obtain combined ‘PDF and  $\alpha_s$  uncertainties’ can be found in [41]. When PDFs are provided for different quark mass values (MSTW08 and NNPDF2.1) one could similarly compute quark-mass uncertainties.

The PDF uncertainties just discussed can be propagated to uncertainties of physical observables such as cross sections. The values of  $\alpha_s$  and of the quark masses used in a PDF set should of course be consistent with the ones used for calculating the observable.

In the following two subsections we discuss some aspects of parametric uncertainties in conventional PDF fits, and then we briefly explain the Monte Carlo method used by the NNPDF collaboration.

### 9.5.1.1 The Hessian Matrix

Put in a simplified manner, conventional PDF determinations minimise the quantity

$$\chi^2(\mathbf{p}) = \sum_i^N \frac{[D_i - T_i(\mathbf{p})]^2}{\sigma_{i,\text{stat}}^2 + \sigma_{i,\text{syst}}^2} \quad (9.17)$$

in a fit. The sum runs over all  $N$  data points considered in the fit, which usually pertain to a number of different measurements, observables and experiments. The data point  $i$  has a measured value  $D_i$  and statistical and systematic uncertainties  $\sigma_{i,\text{stat}}$  and  $\sigma_{i,\text{syst}}$ . The theoretical prediction  $T_i$  for  $D_i$  depends on a set of  $n$  parameters  $\mathbf{p} = (p_1, p_2, \dots, p_n)$  that are determined by the fit. As discussed above, this set may include quantities like  $\alpha_s$  in addition to the parameters that determine the PDFs  $f(x, \mu_0)$  at the starting scale of evolution.

The form (9.17) is appropriate for uncorrelated systematic uncertainties of the  $N$  data points. The treatment of correlated systematic uncertainties is more involved, and different modifications of (9.17) are being used (see for example the discussions in [46, 48, 57]). A particular case is the overall normalisation uncertainty on the cross section, which is common to all data of a given measurement. Different ways of including this uncertainty in the definition of  $\chi^2$  are discussed in [46, 57–59]. We will not delve into these complications here and continue our discussion using the simple form (9.17).

The values of  $\mathbf{p}$  obtained by minimising  $\chi^2$  provide a set of estimators for the parameters. Let us assume that the measurements  $D_i$  follow a multi-dimensional Gaussian distribution around their true physical values, and that the number  $N$  of data points is sufficiently large for the central limit theorem to hold. One then has

$$\Delta\chi^2(\mathbf{p}) = \chi^2(\mathbf{p}) - \chi^2_{\min} = \sum_{ij}^n (p - p_{\min})_i H_{ij} (p - p_{\min})_j , \quad (9.18)$$

where  $\mathbf{p}_{\min}$  is the set of parameters for which  $\chi^2$  is minimised and  $\mathbf{H}$  is the Hessian matrix. The estimators  $\mathbf{p}_{\min}$  follow a multi-dimensional Gaussian distribution with covariance matrix  $\mathbf{V} = \mathbf{H}^{-1}$ . Consider now a function  $F(\mathbf{p})$  of the fit parameters, which may be a PDF or a physical observable such as a cross section. Then  $F(\mathbf{p}_{\min})$  follows again a Gaussian distribution, and linear error propagation,

$$\Delta F = T \sqrt{\sum_{ij}^n \frac{\partial F}{\partial p_i} (\mathbf{H}^{-1})_{ij} \frac{\partial F}{\partial p_j}} , \quad (9.19)$$

gives the statistical uncertainty of  $F$ . Here the tolerance  $T$  corresponds to a value  $\Delta\chi^2 = T^2$ , with  $T = 1$  at 68% CL,  $T = 2.71$  at 90% CL, and so on. The evaluation of the expression (9.19) is rather involved if  $F$  is a cross section since one needs the derivatives  $\partial F / \partial p_i$ , which are typically not available in closed analytic form. To avoid evaluating the uncertainty (9.19), one can use the so-called *Hessian method* [46, 60]. For this, one has to diagonalise the Hessian matrix  $\mathbf{H}$  and to rescale the associated

eigenvectors in parameter space, such as to get a set of linear combinations  $z_i$  of  $(p - p_{\min})_i$  with

$$\Delta\chi^2(z) = \sum_i^n z_i^2. \quad (9.20)$$

The uncertainty on  $F$  can then be computed as

$$\Delta F = T \sqrt{\sum_i^n \left( \frac{\partial F}{\partial z_i} \right)^2} = \sqrt{\sum_i^n \left[ \frac{F(S_i^+) - F(S_i^-)}{2} \right]^2}. \quad (9.21)$$

In the second step of (9.21) we have linearised  $F$  around  $z = \mathbf{0}$  and introduced  $2n$  ‘eigenvector PDFs’  $S_i^\pm$  for each parton species, which are the PDFs evaluated with parameters  $z_i = \pm T$  and  $z_j = 0$  for  $j \neq i$ . One can thus compute  $\Delta F$  by evaluating  $F$  for a set of given PDFs, just as one computes the central value of  $F$  from the best fit PDFs (i.e. with all  $z_i = 0$ ).

If a PDF fit is performed with more free parameters than the data can reliably determine, then the Hessian matrix typically has some eigenvalues that are very small and correspond to ‘flat directions’ in the space of PDF parameters. This can cause severe numerical instabilities in the diagonalisation of the Hessian matrix and the computation of eigenvector PDFs. A way to avoid this is to perform the best fit with a larger number of parameters (so as to have a more flexible form of the PDFs), whilst fixing some of those parameters when calculating the PDF eigenvector sets. This has for instance been done in [47].

It is straightforward to generalise the Hessian method to the case of several functions  $F_1, F_2, \dots$ , expressing their uncertainties and their correlations in terms of eigenvector PDFs, see for example [61].

If the measurements  $D_i$  follow a multi-dimensional Gaussian distribution, then one expects that an individual experiment  $m$  with  $N_m$  data points contributes  $\chi^2_{m,\min} \sim N_m$  to the overall  $\chi^2_{\min}$  of the fit [62]. Some (but not all) fitting groups find, however, that for particular experiments  $\chi^2_{m,\min}$  is significantly smaller or larger than  $N_m$ , and conclude that under these circumstances an uncertainty estimate based on expression (9.21) with the canonical values of  $T$  for 68% or 90% CL is not reliable. This may be due to problems with the central values or uncertainty estimates of individual experiments, but it may also originate from shortcomings of the theoretical description or a too rigid PDF parameterisation. As a remedy, different collaborations are using modified values for the tolerance  $T$  in formula (9.20). Associated with a given CL (conventionally 68% or 90%), these values are chosen such that the individual experiments in the fit contribute to the overall  $\chi^2_{\min}$  in a way one would expect for an ideal situation without tensions in the data or the theory. The tolerance values used in current fits are quoted in Table 9.1 and are further discussed in [46, 60, 63]. The MSTW08 collaboration uses different values of  $T$  for each eigenvector PDF, hence the ranges given in the table.

The Hessian method is adequate in kinematic regions where the PDFs are well constrained by data, so that their uncertainties are small. However, the method

becomes unreliable when uncertainties become large, so that corrections to the quadratic dependence (9.18) of  $\Delta\chi^2$  on  $\mathbf{p}$  or to the linear error propagation and the replacement of derivatives by differences in the uncertainties (9.19) and (9.21) become important. The methods described in the following two sections provide two quite different remedies for such a situation.

### 9.5.1.2 Lagrange Multipliers

In the method of *Lagrange multipliers*, which was elaborated for PDF determinations in [57], one performs fits that minimise  $\chi^2$  under the constraint that  $F(\mathbf{p})$  takes a particular value  $v$ . As in the previous section,  $F(\mathbf{p})$  can be a cross section or any other function of the PDFs (including the PDFs themselves). The constrained minimisation can be performed using Lagrange multipliers as explained in Section 7.3, hence the name of the method.

One can then determine the value  $v$  for which the minimum  $\chi^2$  in the constrained fit satisfies

$$\chi_{\min}^2|_{g=v} = \chi_{\min}^2 + T^2, \quad (9.22)$$

where  $\chi_{\min}^2$  is obtained in the unconstrained fit and  $T$  is the chosen value of the tolerance. The largest and the smallest of those values  $v$  gives an upper and a lower uncertainty estimate for  $F$ , respectively. It is not difficult to show that these estimates are identical to those of the Hessian method if  $F(\mathbf{p})$  is strictly linear and  $\Delta\chi^2(\mathbf{p})$  strictly quadratic in  $\mathbf{p}$ . The constrained minimum  $\chi_{\min}^2|_{g=v}$  is then quadratic in  $v$ . A plot of this function therefore provides a good indication when the uncertainty estimate of the Hessian method becomes unreliable.

The method of Lagrange multipliers is not used very frequently because it requires dedicated PDF fits for each function  $F$  one wants to compute. Example studies can be found in [57, 63, 64] and also in the fit [65] of polarised parton densities.

### 9.5.1.3 The NNPDF Approach

We now briefly describe the method of determining PDFs that has been pioneered by the NNPDF collaboration, referring readers to [66] for a detailed account. As already mentioned, the PDFs at the starting scale are given by neural networks with about ten times more free parameters than in conventional fits, with the aim of avoiding a parameterisation bias on the PDFs. The PDFs are then determined by the following procedure:

1. One generates a Monte Carlo ensemble of replicas of the original dataset to be fitted. The data points  $D_i$  of the replicas are distributed according to the central values and uncertainties specified by the measurements. Depending on the situation, sets of  $N_{\text{rep}} = 100$  or 1000 replicas are usually used.
2. A best-fit PDF is constructed for each replica. This is not done by simple  $\chi^2$  minimisation, since a corresponding fit with several hundred parameters would be hopelessly under-determined. Instead, the data of each replica set are randomly divided into a *training set* and a *validation set*. The fitting algorithm then aims to minimise  $\chi^2$  calculated from the training set but stops when the

$\chi^2$  of the validation set starts to increase. This procedure is used to avoid ‘fitting the noise’ in the data, that is to obtain reasonably smooth PDFs that provide a good interpolating description of the data rather than going through almost every data point. A more detailed discussion can be found in Chapter 5.

3. One thus obtains a statistical ensemble of PDFs. The average and the variance of a function  $F(\mathbf{p})$  can then be calculated in the standard way as

$$\bar{F} = \frac{1}{N_{\text{rep}}} \sum_r^{N_{\text{rep}}} F_r, \quad (\Delta F)^2 = \frac{1}{N_{\text{rep}} - 1} \sum_r^{N_{\text{rep}}} (F_r - \bar{F})^2, \quad (9.23)$$

where  $F_r$  is evaluated from the best-fit PDF of the replica  $r$ . If the  $F_r$  follow a Gaussian distribution, then  $\bar{F} - \Delta F < F < \bar{F} + \Delta F$  is a 68% CL interval. Alternatively, one can determine this interval from the condition that the complementary intervals with  $F < \bar{F} - \Delta F$  and  $F > \bar{F} + \Delta F$  each correspond to 16% of all replicas [67, 68]. This gives a proper confidence interval even if the probability distribution of  $F_r$  is not Gaussian.

### 9.5.2

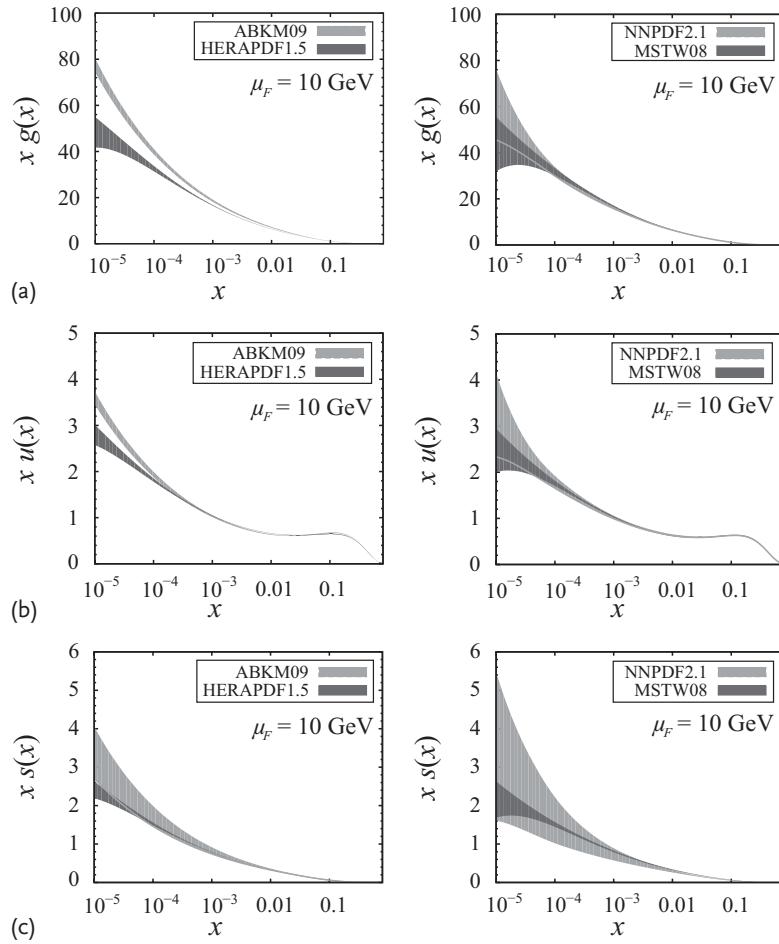
#### A Comparison of Recent PDF Sets

In Figure 9.8 we compare the densities of gluons, up quarks and strange quarks obtained in recent fits at NNLO. Here HERAPDF1.5 [69] is a successor of the HERAPDF1.0 set that is listed in Table 9.1 and documented in [48]. We see that at the chosen factorisation scale the PDFs are rather well determined at larger  $x$  values, whereas the uncertainties significantly increase towards lower  $x$ . Among all distributions (also for the quarks and antiquarks not shown in the figure), the ones for  $s$  and  $\bar{s}$  are least well known. We note that due to evolution effects the size of the uncertainty bands typically decreases with the increasing scale  $\mu_F$ .

We also observe that at higher values of  $x$  the different sets agree rather well within their uncertainty bands, whereas this is not the case as  $x$  decreases. It is important to realise that this disagreement at small  $x$  does *not* imply that the uncertainty estimates of the PDFs are unreliable. As we have discussed in detail, the bands reflect how the uncertainties in the fitted data propagate into uncertainties in the PDF parameters. A disagreement between the bands of different PDFs can be due to any of the other uncertainties discussed in Section 9.5.

When comparing different PDFs one should also bear in mind that they are not directly observable quantities. They depend on different theoretical choices, often called ‘schemes’, each of which comes with a particular prescription of how to calculate cross sections. All PDFs shown here are defined in the  $\overline{\text{MS}}$  renormalisation scheme, but as discussed above they correspond to different schemes for treating heavy flavours.

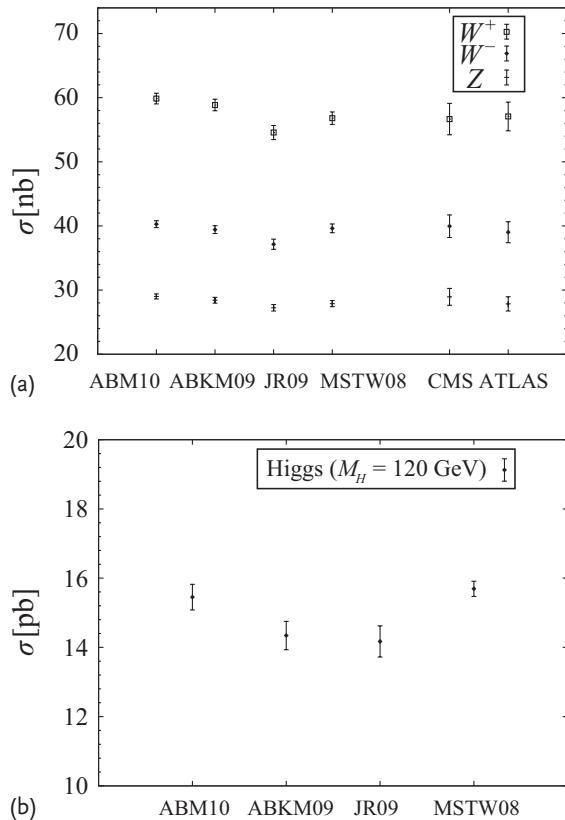
In Figure 9.9 we compare the cross sections for  $W^\pm$ -,  $Z$ - and Higgs-boson production at the LHC, computed with different PDFs at NNLO in [71]. The ABM10 set [74] is an update of ABKM09 [45], using the combined DIS data from H1 and



**Figure 9.8** Recent NNLO determinations of the density of gluons (a), up quarks (b) and strange quarks (c) at  $\mu_F = 10$  GeV. The bands correspond to 68% CL and in the case of NNPDF2.1 have been obtained us-

ing (9.23). For HERAPDF1.5 they reflect only the ‘experimental’ PDF errors as specified at the beginning of Section 9.5.1. The ABKM09 distributions are for  $n_F = 5$  quark flavours. All numerical values have been taken from [70].

ZEUS [48] as well as a more flexible PDF parameterisation and improved higher-order corrections for heavy-flavour production in DIS. We clearly see that the different predictions do not overlap within the error bars that reflect the parametric PDF uncertainties. This reinforces our above warning that these uncertainties alone need not give a reliable estimate for the overall uncertainty of a prediction due to our imperfect knowledge of PDFs. A more appropriate procedure is therefore to compare predictions from different PDF sets that are judged to be sufficiently up-to-date regarding theory and the description of experimental data. De-



**Figure 9.9** Cross sections for inclusive  $W^\pm$  and  $Z$  production (a) and for inclusive production of a Higgs boson with mass  $M_H = 120 \text{ GeV}$  (b) in  $p p$  collisions at  $\sqrt{s} = 7 \text{ TeV}$ . The theoretical predictions have been computed with different PDF sets at NNLO accu-

racy in [71]; their error bars represent the parametric PDF uncertainties at 68% CL. In (a) we include for comparison recent results from CMS [72] and ATLAS [73], where we have divided the experimental measurements by the leptonic branching ratios given in [2].

tailed comparisons of this type of benchmark processes can for instance be found in [3, 41, 71, 75, 76].

## 9.6

### Exercises

#### Exercise 9.1 The running QCD coupling

Extend (9.6) and (9.8) to one order higher in  $\alpha_s$ , that is calculate  $\alpha_3(\mu_R/Q)$ . Use the result to numerically compute  $\alpha_s(\mu_R)$  at  $\mu_R = m_t, M_Z/2$  and  $m_b$ , taking the value of  $\alpha_s(M_Z) = 0.1184$  as input. The  $\beta$ -function coefficients  $b_i$  can be found

in [10]. Compare the result with what is obtained from the expansion of  $\alpha_s(\mu_R)$  in  $1/\ln(\mu_R/\Lambda_{\text{QCD}})$  at the appropriate order, which is also given in [10].

### Exercise 9.2 Perturbative expansion I

Verify that the expansion given by (9.5) and (9.9) is independent of  $\mu_R$  up to terms of order  $\alpha_s^{k+3}$  by calculating the derivative

$$\frac{dC}{d \ln \mu_R^2} = \frac{\partial C}{\partial \ln \mu_R^2} + \beta(\alpha_s) \frac{\partial C}{\partial \alpha_s}, \quad (9.24)$$

where  $\alpha_s$  is always taken at scale  $\mu_R$ .

### Exercise 9.3 Perturbative expansion II

Verify the relation (9.15) and calculate the coefficient  $C_2$  in the expression (9.14). Combine your result with (9.9) and express

$$C_2 \left( \frac{\mu_F}{Q}, \frac{\mu_R}{Q} \right) \quad (9.25)$$

in terms of  $C_2(1, 1)$  and  $C_1(1, 1)$ .

### Exercise 9.4 Evolution equations

Show that the convolution defined in (9.2) is associative, that is that  $(f \otimes g) \otimes h = f \otimes (g \otimes h)$ . This is needed to obtain the evolution equation (9.13) for the hard-scattering kernel from the one for PDFs.

### Exercise 9.5 Renormalisation scale variation

Use the different curves in Figure 9.4 to read off the value of  $\Gamma$  for  $\mu = M_H$  at a given order in  $\alpha_s$ , as well as its uncertainty estimate obtained by varying  $\mu$  by a factor between 1/2 and 2. Repeat the exercise for the central value  $\mu = M_H/2$ .

### Exercise 9.6 Lagrange multipliers and Hessian method

Show that under the conditions spelled out below (9.22) both the Lagrange multiplier method and the Hessian method give the uncertainty estimate in expression (9.19).

## References

- 1 Rossi, G. (2010) *Lattice Field Theory*. Proc., 28th Int. Symp., Lattice 2010, Vilasimius, Italy, 14–19 June 2010. *PoS*, Lattice 2010.
- 2 Nakamura, K. et al. (2010) Review of particle physics. *J. Phys. G*, **37**, 075021.
- 3 Dittmaier, S. et al. (2011) Inclusive observables, in *Handbook of LHC Higgs Cross Sections*, arXiv:1101.0593.
- 4 Collins, J.C., Soper, D.E., and Sterman, G.F. (1989) Factorization of hard processes in QCD, in *Perturbative Quantum Chromodynamics* (ed. A. Mueller), World Scientific, arXiv:hep-ph/0409313.
- 5 Collins, J. (2011) *Foundations of Perturbative QCD*. Cambridge University Press.
- 6 Beneke, M. et al. (2000) QCD factorization for exclusive, nonleptonic  $B$  meson decays: General arguments and the case of heavy light final states. *Nucl. Phys. B*, **591**, 313.
- 7 Beneke, M. et al. (2001) QCD factorization in  $B \rightarrow \pi K$ ,  $\pi\pi$  decays and extraction of Wolfenstein parameters. *Nucl. Phys. B*, **606**, 245.
- 8 Peskin, M.E. and Schroeder, D.V. (1995) *An Introduction to Quantum Field Theory*, Perseus Books.
- 9 van Ritbergen, T., Vermaseren, J., and Larin, S. (1997) The four loop  $\beta$  function in quantum chromodynamics. *Phys. Lett. B*, **400**, 379.
- 10 Bethke, S. (2009) The 2009 world average of  $\alpha_s$ . *Eur. Phys. J. C*, **64**, 689.
- 11 Baikov, P. and Chetyrkin, K. (2006) Higgs decay into hadrons to order  $\alpha^5(s)$ . *Phys. Rev. Lett.*, **97**, 061803.
- 12 Grunberg, G. (1980) Renormalization group improved perturbative QCD. *Phys. Lett. B*, **95**, 70.
- 13 Stevenson, P.M. (1981) Optimized perturbation theory. *Phys. Rev. D*, **23**, 2916.
- 14 Brodsky, S.J., Lepage, G., and Mackenzie, P.B. (1983) On the elimination of scale ambiguities in perturbative quantum chromodynamics. *Phys. Rev. D*, **28**, 228.
- 15 Ellis, R., Stirling, W., and Webber, B. (1996) *QCD and Collider Physics*, Cambridge University Press.
- 16 Brock, R. et al. (1995) Handbook of perturbative QCD. *Rev. Mod. Phys.*, **67**, 157.
- 17 Moch, S., Vermaseren, J., and Vogt, A. (2004) The three loop splitting functions in QCD: The nonsinglet case. *Nucl. Phys. B*, **688**, 101.
- 18 Vogt, A., Moch, S., and Vermaseren, J. (2004) The three loop splitting functions in QCD: The singlet case. *Nucl. Phys. B*, **691**, 129.
- 19 Anastasiou, C. et al. (2004) High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO. *Phys. Rev. D*, **69**, 094008.
- 20 Anastasiou, C., Melnikov, K., and Petriello, F. (2005) Fully differential Higgs boson production and the di-photon signal through next-to-next-to-leading order. *Nucl. Phys. B*, **724**, 197.
- 21 Laenen, E. (2004) Resummation for observables at TeV colliders. *Pramana*, **63**, 1225.
- 22 Amsler, C. et al. (2008) Review of particle physics. *Phys. Lett. B*, **667**, 1.
- 23 Maul, M. et al. (1997) OPE analysis of the nucleon scattering tensor including weak interaction and finite mass effects. *Z. Phys. A*, **356**, 443.
- 24 Politzer, H. (1980) Power corrections at short distances. *Nucl. Phys. B*, **172**, 349.
- 25 Ellis, R., Furmanski, W., and Petronzio, R. (1983) Unraveling higher twists. *Nucl. Phys. B*, **212**, 29.
- 26 Forshaw, J.R. and Ross, D. (1997) *Quantum Chromodynamics and the Pomeron*, Cambridge University Press.
- 27 Andersson, B. et al. (2002) Small  $x$  phenomenology: Summary and status. *Eur. Phys. J. C*, **25**, 77.
- 28 Andersen, J.R. et al. (2006) Small  $x$  phenomenology: Summary of the 3rd Lund small  $x$  workshop in 2004. *Eur. Phys. J. C*, **48**, 53.
- 29 Jung, H. et al. (2010) The CCFM Monte Carlo generator CASCADE version 2.2.03. *Eur. Phys. J. C*, **70**, 1237.
- 30 Diehl, M. and Schäfer, A. (2011) Theoretical considerations on multiparton interactions in QCD. *Phys. Lett. B*, **698**, 389.

- 31** Bartalini, P. *et al.* (2010) Multiple partonic interactions at the LHC. Proc., 1st Int. Workshop, MPI'08, Perugia, Italy, 27–31 October 2008, arXiv:1003.4220.
- 32** Bartalini, P. *et al.* (2011) Multi-parton interactions at the LHC. arXiv:1111.0469.
- 33** Sjöstrand, T. and Skands, P.Z. (2005) Transverse-momentum-ordered showers and interleaved multiple interactions. *Eur. Phys. J. C*, **39**, 129.
- 34** Buttar, C. *et al.* (2008) Standard Model handles and candles working group: Tools and jets summary report. arXiv:0803.0678.
- 35** Dasgupta, M. and Salam, G.P. (2004) Event shapes in  $e^+e^-$  annihilation and deep inelastic scattering. *J. Phys. G*, **30**, R143.
- 36** Banfi, A., Salam, G.P., and Zanderighi, G. (2010) Phenomenology of event shapes at hadron colliders. *JHEP*, **1006**, 038.
- 37** Gehrmann, T., Jaquier, M., and Luisoni, G. (2010) Hadronization effects in event shape moments. *Eur. Phys. J. C*, **67**, 57.
- 38** Gieseke, S. and Nagy, Z. (2011) Monte Carlo generators and fixed-order calculations: Predicting the (un)predicted, in *Physics at the Teascale* (eds I. Brock and T. Schoerner-Sadenius), John Wiley & Sons.
- 39** Buckley, A. *et al.* (2011) General-purpose event generators for LHC physics. *Phys. Rep.*, **504**, 145.
- 40** Forte, S. (2010) Parton distributions at the dawn of the LHC. *Acta Phys. Pol. B*, **41**, 2859.
- 41** Alekhin, S. *et al.* (2011) The PDF4LHC Working Group Interim Report. arXiv:1101.0536.
- 42** Durham PDF server, <http://hepdata.cedar.ac.uk/pdfs> (last accessed 21 February 2013).
- 43** Les Houches Accord PDF Interface, <http://projects.hepforge.org/lhapdf> (last accessed 21 February 2013).
- 44** Jimenez-Delgado, P. and Reya, E. (2009) Dynamical NNLO parton distributions. *Phys. Rev. D*, **79**, 074023.
- 45** Alekhin, S. *et al.* (2010) The 3, 4, and 5-flavor NNLO parton from deep-inelastic-scattering data and at hadron colliders. *Phys. Rev. D*, **81**, 014032.
- 46** Martin, A. *et al.* (2009) Parton distributions for the LHC. *Eur. Phys. J. C*, **63**, 189.
- 47** Martin, A. *et al.* (2010) Heavy-quark mass dependence in global PDF analyses and 3- and 4-flavour parton distributions. *Eur. Phys. J. C*, **70**, 51.
- 48** Aaron, F. *et al.* (2010) Combined measurement and QCD analysis of the inclusive  $e^\pm p$  scattering cross sections at HERA. *JHEP*, **1001**, 109.
- 49** Lai, H.L. *et al.* (2010) New parton distributions for collider physics. *Phys. Rev. D*, **82**, 074024.
- 50** Ball, R.D. *et al.* (2011) Impact of heavy quark masses on parton distributions and LHC phenomenology. *Nucl. Phys. B*, **849**, 296.
- 51** Ball, R.D. *et al.* (2012) Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO. *Nucl. Phys. B*, **855**, 153.
- 52** Blümlein, J. (2010) The QCD coupling and parton distributions at high precision. *Mod. Phys. Lett. A*, **25**, 2621.
- 53** Alekhin, S. *et al.* (2011)  $\alpha_s(M_Z^2)$  in NNLO analyses of deep-inelastic world data. arXiv:1104.0469.
- 54** Alekhin, S. and Moch, S. (2011) Heavy-quark deep-inelastic scattering with a running mass. *Phys. Lett. B*, **699**, 345.
- 55** Tung, W.K. *et al.* (2007) Heavy quark mass effects in deep inelastic scattering and global QCD analysis. *JHEP*, **02**, 053.
- 56** Andersen, J. *et al.* (2010) The SM and NLO multileg working group: Summary report. arXiv:1003.1241.
- 57** Stump, D. *et al.* (2001) Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method. *Phys. Rev. D*, **65**, 014012.
- 58** Pumplin, J. *et al.* (2002) New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, **0207**, 012.
- 59** Ball, R.D. *et al.* (2010) Fitting parton distribution data with multiplicative normalization uncertainties. *JHEP*, **1005**, 075.
- 60** Pumplin, J. *et al.* (2001) Uncertainties of predictions from parton distribution functions. 2. The Hessian method. *Phys. Rev. D*, **65**, 014013.

- 61** Nadolsky, P.M. *et al.* (2008) Implications of CTEQ global analysis for collider observables. *Phys. Rev. B*, **78**, 013004.
- 62** Collins, J.C. and Pumplin, J. (2001) Tests of goodness of fit to multiple datasets. arXiv:hep-ph/0105207.
- 63** Martin, A. *et al.* (2003) Uncertainties of predictions from parton distributions. 1: Experimental errors. *Eur. Phys. J. C*, **28**, 455.
- 64** Pumplin, J. *et al.* (2009) Collider inclusive jet data and the gluon distribution. *Phys. Rev. D*, **80**, 014019.
- 65** de Florian, D. *et al.* (2009) Extraction of spin-dependent parton densities and their uncertainties. *Phys. Rev. D*, **80**, 034030.
- 66** Ball, R.D. *et al.* (2009) A determination of parton distributions with faithful uncertainty estimation. *Nucl. Phys. B*, **809**, 1.
- 67** Ball, R.D. *et al.* (2009) Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering. *Nucl. Phys. B*, **823**, 195.
- 68** Ball, R.D. *et al.* (2010) A first unbiased global NLO determination of parton distributions and their uncertainties. *Nucl. Phys. B*, **838**, 136.
- 69** HERAPDF table, [www.desy.de/h1zeus/combined\\_results/herapdftable](http://www.desy.de/h1zeus/combined_results/herapdftable) (last accessed 21 February 2013).
- 70** Durham PDF server, <http://hepdata.cedar.ac.uk/pdf/pdf3.html> (last accessed 21 February 2013).
- 71** Alekhin, S. *et al.* (2011) NNLO benchmarks for Gauge and Higgs boson production at TeV hadron colliders. *Phys. Lett. B*, **697**, 127.
- 72** CMS Collab. (2011) Measurement of the inclusive  $W$  and  $Z$  production cross sections in  $p p$  collisions at  $\sqrt{s} = 7$  TeV with the CMS experiment. *JHEP*, **1110**, 132.
- 73** ATLAS Collab. (2012) Measurement of the inclusive  $W^\pm$  and  $Z/\gamma$  cross sections in the electron and muon decay channels in  $p p$  collisions at  $\sqrt{s} = 7$  TeV with the ATLAS detector. *Phys. Rev. D*, **85**, 072004.
- 74** Alekhin, S., Blümlein, J., and Moch, S.O. (2010) Update of the NNLO PDFs in the 3-, 4-, and 5-flavour scheme. *PoS, DIS2010*, 021.
- 75** Demartin, F. *et al.* (2010) The impact of PDF and  $\alpha_s$  uncertainties on Higgs production in gluon fusion at hadron colliders. *Phys. Rev. D*, **82**, 014002.
- 76** Watt, G. (2011) Parton distribution function dependence of benchmark Standard Model total cross sections at the 7 TeV LHC. *JHEP*, **1109**, 069.

## 10

### Statistical Methods Commonly Used in High Energy Physics

*Carsten Hensel and Kevin Kröninger*

#### 10.1

##### Introduction

The previous chapters in this book have introduced a variety of statistical methods – some of them in an abstract way, some of them applied to very concrete use cases. A person working on the analysis of data using these methods might encounter problems when confronted with the details of the implementation. An example are the ensemble tests mentioned in earlier chapters: how are ensembles built? Does a correlation among the ensembles have an impact on the final result? And how many ensembles are necessary for the study in question?

This chapter discusses some details which one sometimes struggles with when analysing data. It also shows some concrete applications of frequently used methods which were presented previously. The discussions include the use of ensemble tests and some of their applications. A classical fitting procedure – the template fit – is given as an example of Bayesian inference, and two examples for the estimation of quantities (efficiencies, background contributions) using data-driven methods are explained. The chapter ends with a discussion of the experimenter's role in data analysis and presents strategies for performing blind analyses which help to avoid subtle biases stemming from the analyser's preconceptions.

This chapter is the first of three focusing on the applications of statistical methods rather than on concepts. Chapter 11 combines the developed strategies and presents a walk-through of two typical analyses – a search and a measurement. Chapter 12 summarises applications of statistical methods in astronomy and gives an introduction to the field of astrostatistics.

#### 10.2

##### Estimating Efficiencies

The estimation of efficiencies is a common data-analysis task in high energy physics. Typical examples include the identification efficiency of objects, such as

electrons or muons, the efficiencies of triggers and the efficiencies to select certain types of processes. In this section, trigger efficiencies will serve as an example. Three different approaches to estimating trigger efficiencies will be described, followed by a discussion on how to estimate the corresponding uncertainties.

### 10.2.1

#### Motivation

Collider experiments are typically unable to record, store and process all available data because of finite CPU resources and storage capacities and because of limitations of the rate at which data can be written to storage media. These constraints necessitate fast and efficient selection mechanisms, so-called *triggers*, to distinguish interesting from less interesting collision events. The selection criteria are based on characteristic event signatures like the number and energies of jets, photons, charged leptons and the missing transverse energy. For most data analyses it is crucial to have a precise estimate of the trigger efficiency, that is of the fraction of relevant events that are really selected by the trigger. This is particularly important when measuring cross sections or when searching for new particles. Moreover, systematic uncertainties related to trigger inefficiencies need to be treated correctly over the course of an analysis.

The usefulness of *trigger emulations* is limited since they might be based on inaccurate detector descriptions. Also, non-simulated machine backgrounds which are difficult to quantify might introduce additional uncertainties or biases. It is thus desirable to extract trigger efficiencies directly from data. The advantages and limitations of such data-driven methods are discussed in the following.

### 10.2.2

#### Trigger Efficiencies and Their Estimates

The trigger efficiency  $\epsilon(T)$  is defined as the probability for an object (or a combination of objects) to fulfil the requirements of a trigger  $T$ . It can be estimated as

$$\hat{\epsilon}(T) = \frac{n_T}{n_R} . \quad (10.1)$$

Here,  $n_T$  is the number of triggered objects and  $n_R$  is the total number of reconstructed objects, that is of objects that were obtained by an offline reconstruction algorithm and which were probed by the trigger mechanism.<sup>1)</sup>

Several requirements to pass a trigger decision can be combined, also on different objects. An example of a combined trigger is a dijet trigger requiring two jets with transverse momenta  $p_T$  above a certain threshold. In general, there is no limit

1) Care needs to be taken when choosing the offline reconstruction algorithm. Its definition should be close to the one used at the trigger level, that is the definition of an ‘offline electron’ should be close to that of a ‘trigger electron’. A too large difference between these two objects could lead to a biased estimate of the trigger efficiency.

on the number of requirements or objects. The efficiency of the event to fulfil all trigger requirements is then a combination of the individual trigger efficiencies. The following discussion is based on single-object triggers for clarity and simplicity, but can easily be extended to more complex trigger requirements.

It is obvious that a data-driven estimate of trigger efficiencies can only be based on events which were already triggered themselves. The challenge of data-driven estimates of trigger efficiencies therefore is to select a subsample of reconstructed offline objects which is independent (or unbiased) with respect to the trigger under study.

#### 10.2.3

##### **The Counting Method**

The *counting method* starts from an unbiased sample of reconstructed objects by analysing events which were selected by a trigger that is uncorrelated to the one under study. The trigger efficiency can be estimated by simply counting the numbers of reconstructed and of triggered objects.

The counting method offers simple access to efficiencies for a variety of triggers. However, the actual choice of uncorrelated triggers is only seemingly straight forward. A clearly unbiased sample can be selected by choosing a random trigger. Such a trigger selects events based on no other criteria than that the data acquisition system records events randomly. However, such a sample will suffer from the fact that it typically only contains a small number of interesting events, that is ‘hard’ events containing objects with large transverse momenta.

In order to enhance the number of interesting events in the selected sample, the random trigger is usually replaced by a *minimum-bias trigger* which requires only a minimal detector activity. Larger subsamples can be collected by using a genuine trigger with selection criteria that are independent of those of the trigger in question. Such triggers are called *orthogonal*. For example, muon-trigger efficiencies can be estimated from samples that were selected with electron triggers.

It should be stressed that the improved statistics in samples selected using minimum-bias or orthogonal triggers comes at a cost because the selected objects might not be completely independent of the trigger under study. The correlation can come from processes in which one object is often accompanied by another one. One example are events  $W \rightarrow \nu e$  where the transverse momentum of the charged lepton is correlated to the missing transverse energy of the event, potentially leading to a bias on the trigger efficiency estimate.

#### 10.2.4

##### **The Tag-and-Probe Method**

The *tag-and-probe method* offers an elegant way to estimate trigger efficiencies with only small statistical uncertainties and without a bias. Events are selected which are compatible with a well-known physics process that features two objects of the same type. The paradigmatic example is the determination of muon-trigger or electron-

trigger efficiencies using events in which a  $Z$ -boson is produced and subsequently decays into two electrons or two muons.

One of these two reconstructed leptons – the so-called *tag lepton* – is required to have fired the trigger. The second lepton is called the *probe lepton*. The trigger efficiency is estimated by the ratio of the number of triggered probe leptons and the total number of probe leptons.

In principle the method is only limited by the number of  $Z$ -boson events and thus improves with increasing integrated luminosity. However, tag and probe leptons are kinematically correlated. Furthermore, the estimated trigger efficiencies are limited to the properties (kinematics etc.) of the probe object. This might lead to a limited coverage of the phase space, for example to high  $p_T$  values for  $Z$ -bosons. The coverage can be extended to lower momenta by including events containing  $J/\psi$  mesons decaying in the same way as the  $Z$ -bosons.

Care must be taken if the amount of background in the selected data sample is not negligible because the correlation between the tag lepton and the probe lepton can be different for different processes. Several approaches exist to suppress background contributions, for example by applying tighter requirements in the selection or by fitting shapes of the data sample with signal and background templates [1].

Although the tag-and-probe method can be used to estimate trigger efficiencies in a truly unbiased way, the method obviously can not be applied to triggers based on global event properties like the missing transverse energy. An application of the tag-and-probe-method can be found in [2].

### 10.2.5

#### The Bootstrap Method

The *bootstrap method* is a simple approach for estimating trigger efficiencies.<sup>2)</sup> It starts with the assumption that the trigger efficiency for a trigger  $T_1$  with a low threshold, such as the transverse momentum of a jet or the missing transverse energy, is already known. This particular trigger is then used to select a sample of events. The relative trigger efficiency for a trigger  $T_2$  with a higher trigger threshold, given that  $T_1$  had already fired, is denoted  $\epsilon(T_2|T_1)$  and can be determined by counting the number of events that also pass  $T_2$ . Let us assume that the conditional probability to fulfil the trigger requirements of the first trigger  $T_1$ , given that the one with a higher threshold  $T_2$  had fired already, is one. Then the trigger efficiency for  $T_2$  is given as

$$\epsilon(T_2) = \epsilon(T_2|T_1) \cdot \epsilon(T_1). \quad (10.2)$$

The number of use cases for the bootstrap method is limited. However, the method is particularly useful when dealing with small data samples: the early LHC data samples did not contain enough events with  $Z$ -bosons to use the tag-and-probe

2) Note that the bootstrap method discussed here is not to be confused with the resampling method described in Section 10.5.

method. However, by using a large sample of  $J/\psi$  events it was possible to sample the turn-on region of the muon trigger with a low transverse-momentum threshold. Starting from there, the bootstrap method was used to extrapolate to higher  $p_T$  values and to extract the efficiencies for triggers with higher thresholds. An application of the bootstrap method can be found in [3].

#### 10.2.6

##### Calculating Uncertainties on Trigger Efficiencies

Each of the described methods has systematic uncertainties that are inherent to the method and strongly depend on the experimental conditions. Some of these uncertainties can be estimated by comparing the outcome of different methods. However, this is not always possible.

Uncertainties due to statistical limitations of the data sample at hand can be treated in a formal way. In most cases, the triggering process can be interpreted as a Bernoulli experiment.<sup>3)</sup> The efficiency is then the probability for an experiment to give a positive outcome (in this case for a positive trigger decision). Let  $n_T$  and  $n_R$  be the numbers of triggered and reconstructed objects, respectively. The probability for  $n_T$  positive outcomes is given by the binomial formula:

$$p(n_T; \epsilon, n_R) = \frac{n_R!}{n_T!(n_R - n_T)!} \epsilon^{n_T} (1 - \epsilon)^{n_R - n_T}. \quad (10.3)$$

The maximum-likelihood estimator is given by  $\hat{\epsilon} = n_T/n_R$  (see (10.1)). The uncertainty on  $\hat{\epsilon}$  can be estimated from the variance of the binomial distribution as

$$\hat{\sigma}[\hat{\epsilon}] = \sqrt{V[\hat{\epsilon}]} = \sqrt{\frac{\hat{\epsilon}(1 - \hat{\epsilon})}{n_R}}. \quad (10.4)$$

Substituting  $\hat{\epsilon}$  using for example (10.1) results in

$$\hat{\sigma}[\hat{\epsilon}] = \sqrt{\frac{n_T(n_R - n_T)}{n_R^3}}. \quad (10.5)$$

For a small number of events, or for efficiencies close to 0 or 1, this approximation is not valid. In particular, the uncertainty vanishes for the extreme cases of  $n_T = 0$  and  $n_T = n_R$ . An interesting discussion on how to use Bayes' theorem in estimating trigger efficiencies in such cases and how to choose an appropriate prior is given in [4]. A detailed discussion of binomial uncertainties is given in Sections 4.4.1 and 4.3.3.3.

3) A *Bernoulli experiment* or *Bernoulli process* is a sequence of decisions with only two outcomes. A typical example of such a process is the tossing of a coin.

### 10.3

#### Estimating the Contributions of Processes to a Dataset: The Matrix Method

The *matrix method* is a widely used approach in high energy physics to estimate the contributions of different processes to a dataset. It is a purely data-driven method and does not rely on Monte Carlo simulations. A typical use case is the estimation of the background contamination in a selected dataset, for example the fraction of QCD multi-jet events in a sample selected for top-quark studies. The former are difficult to simulate, and a data-driven approach is thus preferred.

##### 10.3.1

###### Estimating the Background Contributions to a Data Sample

In the following, a typical cut-based cross-section analysis is considered: a set of cuts,  $S$ , optimised to enhance the number of signal events is applied to a dataset, resulting in a number of  $N(S)$  selected events. In order to estimate the total production cross section, the number of produced events,  $\nu_S$ , needs to be known.<sup>4)</sup> This number can be estimated using the selection efficiency  $\epsilon_S$ , that is

$$\nu(S) = \epsilon_S \nu_S \quad (10.6)$$

is the number of expected events after the selection, and it is estimated by

$$\hat{\nu}(S) = N(S). \quad (10.7)$$

Equation 10.6 assumes that the sample is purely composed of signal events – which is obviously not a very realistic scenario. Let us instead consider the case in which (before the selection)  $n$  different sources of background contribute to the expected number of events and let these contributions be denoted  $\nu_{B_1}, \dots, \nu_{B_n}$ . Usually, the selection criteria  $S$  do not reject all background contributions. Thus, (10.6) has to be modified to

$$\nu(S) = \epsilon_S \nu_S + \epsilon_{S,B_1} \nu_{B_1} + \dots + \epsilon_{S,B_n} \nu_{B_n}, \quad (10.8)$$

where  $\epsilon_{S,B_i}$  is the selection efficiency for the background source  $i$ . The selection criteria  $S$  were chosen to select the signal process with a reasonable selection efficiency. If it is not possible to further tighten the selection requirements – for example because the resulting loss in statistics would not be tolerable – the remaining background contributions  $\epsilon_{S,B_i} \nu_{B_i}$  need to be estimated. This is possible by introducing further sets of selection criteria,  $B_i$ , designed to prepare samples enriched in events from background process  $i$ . Equation 10.8 can then be extended

4) It is understood that  $\nu_S$  is the number of events expected to be produced in the fiducial region of the detector according to Poisson statistics. In addition, the detector acceptance is ignored in this example, as are further sources of inefficiency.

to a system of equations:

$$\begin{aligned}\nu(S) &= \epsilon_S \nu_S + \epsilon_{S,B_1} \nu_{B_1} + \cdots + \epsilon_{S,B_n} \nu_{B_n}, \\ \nu(B_1) &= \epsilon_{B_1,S} \nu_S + \epsilon_{B_1} \nu_{B_1} + \cdots + \epsilon_{B_1,B_n} \nu_{B_n}, \\ &\cdots = \cdots + \cdots + \cdots + \cdots, \\ \nu(B_n) &= \epsilon_{B_n,S} \nu_S + \epsilon_{B_n,B_1} \nu_{B_1} + \cdots + \epsilon_{B_n} \nu_{B_n}.\end{aligned}\quad (10.9)$$

Alternatively, this can be written in matrix notation (which gave the method its name),

$$\begin{pmatrix} \nu(S) \\ \nu(B_1) \\ \cdots \\ \nu(B_n) \end{pmatrix} = \begin{pmatrix} \epsilon_S & \epsilon_{S,B_1} & \cdots & \epsilon_{S,B_n} \\ \epsilon_{B_1,S} & \epsilon_{B_1} & \cdots & \epsilon_{B_1,B_n} \\ \cdots & \cdots & \cdots & \cdots \\ \epsilon_{B_n,S} & \epsilon_{B_n,B_1} & \cdots & \epsilon_{B_n} \end{pmatrix} \begin{pmatrix} \nu_S \\ \nu_{B_1} \\ \cdots \\ \nu_{B_n} \end{pmatrix}, \quad (10.10)$$

or even shorter as

$$\boldsymbol{\nu}_{\text{sel}} = \boldsymbol{\epsilon} \boldsymbol{\nu}. \quad (10.11)$$

Here,  $\boldsymbol{\nu}_{\text{sel}}$  is the vector of events selected using the different criteria  $S$  and  $B_i$ ,  $\boldsymbol{\nu}$  is the vector of the expected numbers of events for the different processes, and  $\boldsymbol{\epsilon}$  is the efficiency matrix. For perfectly efficient selection criteria, the matrix  $\boldsymbol{\epsilon}$  is diagonal with  $\epsilon_{ii} = 1$ . In more realistic cases,  $\boldsymbol{\epsilon}$  is not diagonal and not necessarily symmetric.

We can now estimate the original signal and background sample contributions as

$$\boldsymbol{\nu} = \boldsymbol{\epsilon}^{-1} \boldsymbol{\nu}_{\text{sel}} \quad (10.12)$$

and thus

$$\hat{\boldsymbol{\nu}} = \boldsymbol{\epsilon}^{-1} \boldsymbol{N}_{\text{sel}}, \quad (10.13)$$

where  $\boldsymbol{N}_{\text{sel}}$  is the vector containing the numbers of observed events for the different sets of selection criteria.

### Example 10.1 QCD multi-jet background in studies of $t\bar{t}$ pair production

The most significant sources of background in the single-lepton decay mode of  $t\bar{t}$  pair production are QCD multi-jet production and  $W + \text{jets}$  production. The first is often estimated using the matrix method. In the single-electron channel, the signature of top-quark pairs is defined by a single high- $p_T$  electron, at least four jets and large missing transverse momentum. QCD multi-jet events can mimic this signature if one of the jets is misidentified as an electron (*fake electron*). These processes are difficult to simulate, and thus the matrix method is used to estimate this contribution from data. Events containing a real electron can either stem from  $t\bar{t}$  pair production or from  $W + \text{jets}$  production – the separation of the latter two processes being an analysis task which is not further discussed here.

A typical selection of top-quark events includes stringent ('tight') requirements on the electron identification; the number of selected events in this 'tight' sample is called  $N_{\text{tight}}$ . Furthermore, a second sample is selected with less stringent or 'loose' requirements, containing  $N_{\text{loose}}$  events. Both samples are considered to contain events with both *real* and *fake* leptons, leading to expected numbers of events

$$\nu_{\text{loose}} = \nu_{\text{lep}} + \nu_{\text{fake}}, \quad (10.14)$$

$$\nu_{\text{tight}} = \epsilon_{\text{lep}} \nu_{\text{lep}} + \epsilon_{\text{fake}} \nu_{\text{fake}}. \quad (10.15)$$

Here,  $\nu_{\text{lep}}$  and  $\nu_{\text{fake}}$  are the expected numbers of real and fake leptons in the loose selection, respectively;  $\epsilon_{\text{lep}}$  and  $\epsilon_{\text{fake}}$  are the probabilities for real and fake leptons to fulfil the tight selection criteria given that the loose selection criteria are already fulfilled and assuming that the selection efficiency for loose events is one. These efficiencies are typically estimated in auxiliary measurements using, for example, the tag-and-probe method (Section 10.2.4). The system of equations can be solved for  $\nu_{\text{fake}}$ . The number of multi-jet events in the tight sample is then given by

$$\epsilon_{\text{fake}} \nu_{\text{fake}} = \frac{\epsilon_{\text{fake}}}{\epsilon_{\text{fake}} - \epsilon_{\text{lep}}} (\nu_{\text{tight}} - \epsilon_{\text{lep}} \nu_{\text{loose}}). \quad (10.16)$$

The number of multi-jet events is thus estimated by substituting  $\nu_{\text{loose}}$  and  $\nu_{\text{tight}}$  with  $N_{\text{loose}}$  and  $N_{\text{tight}}$ . The statistical uncertainties on the estimate of  $\nu_{\text{fake}}$  can be calculated using error propagation. More details on this application can be found in [5].

### 10.3.2

#### Extension to Distributions

The matrix method results in an estimate of the contribution of one or more processes to a dataset, that is the normalisation factors. However, the method as presented above does not give an estimate of the distribution of these processes, for example that of the transverse energy of a fake electron. The distributions can be estimated using an extended matrix method by calculating a weight for each event in the dataset, and using the weighted distribution as an estimate of the distribution of the quantity of interest.

Using the above example, these weights  $w$  can be calculated using (10.16) for each event separately. If the event is contained in the loose sample, but not in the tight sample, then the weight becomes

$$w = \epsilon_{\text{fake}} \nu_{\text{fake}} = \frac{\epsilon_{\text{fake}} \epsilon_{\text{lep}}}{\epsilon_{\text{lep}} - \epsilon_{\text{fake}}}. \quad (10.17)$$

If the event is contained in both samples, its weights will be

$$w = \frac{(\epsilon_{\text{lep}} - 1) \epsilon_{\text{fake}}}{\epsilon_{\text{lep}} - \epsilon_{\text{fake}}}. \quad (10.18)$$

Note that the efficiencies  $\epsilon_{\text{fake}}$  and  $\epsilon_{\text{lep}}$  can depend on kinematic quantities, such as the  $p_T$ ,  $\eta$  or  $\phi$  of the object under study. The weights are nothing but the expected number of events (with fake leptons) given that only one event with these specific kinematic properties was observed. They can thus be interpreted as probabilities. These weights can be used to estimate the distribution of any event or object quantity.

### 10.3.3

#### Limitations of the Matrix Method

Despite its seeming simplicity, there are limitations to the matrix method.

First, the method relies on the different processes to be sufficiently well distinguishable from each other. Otherwise, the matrix  $\epsilon$  could become singular.

Second, depending on the problem, the entries of the matrix  $\epsilon$  can be obtained from auxiliary measurements in control regions of the data or from Monte Carlo simulations. The efficiencies themselves come with additional uncertainties, and these can be correlated with the sources of systematic uncertainty considered in the main analysis, making the estimate of the total uncertainty more complicated.

One potential source of uncertainty related to the efficiencies determined in auxiliary measurements is that of the extrapolation. The extrapolation from a control region to the signal region is not necessarily straightforward or without uncertainty. This is particularly important if both regions have different kinematic properties or use different detector components. In such cases it is advisable to define a second control region or to perform a cross-check using an altogether different method.

## 10.4

### Estimating Parameters by Comparing Shapes of Distributions: The Template Method

A powerful alternative method to estimate the contributions from different processes to an observed dataset is the so-called *template method*. The method exploits shape differences in the distribution of certain observables for the set of processes contributing to the data. These distributions are often given as histograms and are referred to as *templates*. Templates are typically obtained from Monte Carlo simulations or from control regions of the data. The normalisation constants of each template, corresponding to the contributions of the processes, are interpreted as model parameters and fitted to the data histogram by optimising the agreement between the data and the sum of templates. The template method is explained in the following and motivated using Bayes' theorem. Its performance can be evaluated using ensemble tests as discussed in Section 10.5. The template method is implemented in the `BAT` program [6] which was used for the examples in this section.

It is assumed that a set of  $n$  events has been measured in an experiment. The events are known to come from a number of different processes,  $N^P$ . A single process  $j$  contributes with  $n_j^P$  events to the observed dataset such that  $\sum_{j=1}^{NP} n_j^P = n$ . In the model considered here,  $n_j^P$  is a random number drawn from a Poisson distribution with an expectation value  $\nu_j^P$ . Neither  $n_j^P$  nor  $\nu_j^P$  are known.<sup>5)</sup> The aim of the analysis is to estimate the numbers of expected events of all processes involved,  $\nu_j^P$ , which are the parameters of the statistical model used.

The template method works as follows: a variable  $x$  is calculated for all measured events. The distribution of  $x$  is filled into a histogram which is divided into  $N_{\text{bins}}$  bins. It is assumed that the number of events in bin  $i$ ,  $n_i$ , fluctuates around an expectation value  $\nu_i$  according to a Poisson distribution and that fluctuations in the individual bins are independent. Furthermore, the probability density  $f_j(x)$  of the variable  $x$ , that is the template,<sup>6)</sup> is assumed to be known for all processes  $j$ , and it is typically taken from Monte Carlo predictions or obtained through data-driven methods. The number of events expected in bin  $i$  can then be expressed as the sum of the expected numbers of events from each process weighted with the probability to occur in that bin,

$$\nu_i = \sum_{j=1}^{NP} \nu_j^P \cdot \int_{\Delta x_i} f_j(x) dx \approx \sum_{j=1}^{NP} \nu_j^P \cdot f_j(x_i) \cdot \Delta x_i , \quad (10.19)$$

where  $x_i$  is the center of bin  $i$  and  $\Delta x_i$  is the width of the bin.

The problem can be formulated in a Bayesian framework. The *posterior probability density*, or *posterior*, for the parameters given the data,  $p(\boldsymbol{\nu}^P; \mathbf{n})$ , can be calculated using Bayes' theorem:

$$p(\boldsymbol{\nu}^P; \mathbf{n}) = \frac{p(\mathbf{n}; \boldsymbol{\nu}^P) \cdot \pi_0(\boldsymbol{\nu}^P)}{\int p(\mathbf{n}; \boldsymbol{\nu}^P) \cdot \pi_0(\boldsymbol{\nu}^P) d\boldsymbol{\nu}^P} . \quad (10.20)$$

Here,  $p(\mathbf{n}; \boldsymbol{\nu}^P)$  is the probability density (likelihood) for observing  $\mathbf{n} = (n_1, \dots, n_i, \dots, n_{N_{\text{bins}}})$  events, given  $\boldsymbol{\nu}^P = (\nu_1^P, \dots, \nu_j^P, \dots, \nu_{N_P}^P)$ . The expression  $\pi_0(\boldsymbol{\nu}^P)$  is the *prior probability density*, or *prior*; it summarises the knowledge about the parameters  $\boldsymbol{\nu}^P$  before the analysis of the current data is performed. In the present case, the likelihood is defined as a product of Poisson terms:

$$p(\mathbf{n}; \boldsymbol{\nu}^P) = \prod_{i=1}^{N_{\text{bins}}} \frac{\nu_i^{n_i}}{n_i!} \cdot e^{-\nu_i} . \quad (10.21)$$

It is worth mentioning again (see Section 2.5.3) that the product of Poisson distributions can be derived from a multinomial distribution if the total number of

- 5) The case that prior knowledge (for example from auxiliary measurements) is used will be discussed separately in Section 10.4.2.
- 6) The term *template* is used for the probability density  $f_j(x)$  as well as for the distribution of probabilities at the bin centres,  $f_j(x_i)$ .

events,  $n$ , is allowed to fluctuate according to a Poisson distribution with expectation value  $\nu = \sum_{j=1}^{N_p} \nu_j^p$ :

$$p(\mathbf{n}; \boldsymbol{\nu}^p) = \left( \frac{n!}{\prod_{i=1}^{N_p} n_i!} \prod_{i=1}^{N_p} \left( \frac{\nu_i}{\nu} \right)^{n_i} \right) \cdot \frac{\nu^N}{n!} e^{-\nu} = \prod_{i=1}^{N_p} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}. \quad (10.22)$$

An estimator for the model parameters – the expectation values  $\boldsymbol{\nu}^p$  – is the *global mode*, that is the set of parameters which maximises the posterior. A probability distribution for a single model parameter,  $p(\nu_k^p)$ , can be obtained by calculating the *marginal distribution* of the posterior:

$$p(\nu_k^p; \mathbf{n}) = \int p(\boldsymbol{\nu}^p; \mathbf{n}) \prod_{j \neq k}^{N_p} d\nu_j^p. \quad (10.23)$$

Typical estimators for  $\nu_k^p$  are the mode, mean and median of the marginal distribution.<sup>7)</sup> Uncertainties on  $\nu_k^p$  can be obtained by calculating, for example, the 16–84% quantiles of the distribution, the standard deviation or the smallest interval containing 68% probability. Similarly, limits on a single parameter can be obtained by solving the equation (here: for the 90% probability upper limit)

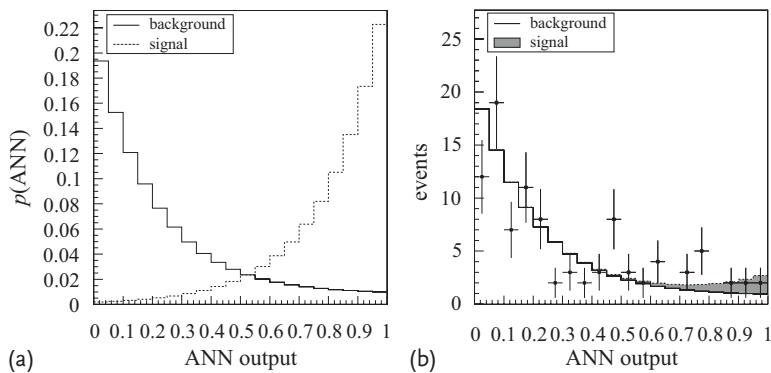
$$0.9 = \int_{\nu_{k,\min}^p}^{\nu_{k,90\%}^p} p(\nu_k^p; \mathbf{n}) d\nu_k^p \quad (10.24)$$

for  $\nu_{k,90\%}^p$ , where  $\nu_{k,\min}^p$  is the lower bound of the parameter range.

### Example 10.2 Searches using neural network outputs

Artificial neural networks (ANN, see Chapter 5) can be used to distinguish signal from background events. Consider the search for a hypothetical particle. A neural network was trained, and the output distributions (normalised to unity) for the signal and the background are shown in Figure 10.1a. Figure 10.1b shows the output distribution observed in data. A typical strategy would be to cut at a certain value of the output in order to obtain a pure signal sample. Instead, however, one can also fit the entire output spectrum using the output distributions of the neural network for signal and background events as templates. This increases the sensitivity to a potential signal. Figure 10.1b also shows the sum of the two templates where each template is normalised to the best-fit result. A small contribution from the signal process,  $\nu_{\text{sgn}}^p$ , can be seen ( $\hat{\nu}_{\text{sgn}}^p = 8.7^{+5.1}_{-4.2}$ ). However, this is not sufficient to claim a discovery, and, instead, an upper limit on the signal contribution is set.

7) Note that the global mode and the set of modes of the marginal distributions do not have to coincide.



**Figure 10.1** (a) Output distributions (normalised to unity) of a neural network for signal (dashed line) and background events (solid line). (b) Distribution observed in data (points) and the sum of signal and background templates scaled to their best-fit parameter values.

#### 10.4.1 Template Shapes

In most applications, the probability densities  $f_j(x)$  are approximated by frequency distributions taken from simulation or data. The statistical uncertainty on these distributions is typically assumed to be negligible. If this assumption is not valid, alternatives like the *kernel density estimation* [7, 8] or a parameterisation of the distribution should be considered. Tools like the `TFractionFitter` [9, 10] consider the statistical uncertainty of the templates assuming Gaussian fluctuations.

These considerations are particularly important if a distribution has bins with zero entries because the probability to find events from the process under consideration in them will be zero. Consequently, if data are observed in a bin with zero predicted events, the template fit will return a zero probability for all fit parameters.

Care has to be taken when smoothing algorithms are applied to the frequency distributions since these can cause large deviations from the frequency distribution at the interval boundaries of the model parameters. Furthermore, one should also consider assigning a systematic uncertainty on the altered shape of the templates.

#### 10.4.2 Including Prior Knowledge

Prior knowledge about the individual contributions (as implemented in (10.20) in the form of the factor  $\pi_0$ ) can come from auxiliary measurements or from theoretical constraints. An example of the former is the estimation of the background contribution from a sideband region. A theoretical constraint could be a relation between several contributions; it might for example be the case that the total background to a process is known, but its composition is not.

The contribution of processes, and thus the parameters of the fit, should be constrained to be larger or equal to zero. Such a constraint can be modelled via the priors of the parameters. One should note that, when performing ensemble tests, a bias in the estimated contributions can occur for values close to zero since negative contributions are not allowed in the generation of ensembles and fluctuations will only be towards positive values.<sup>8)</sup> One solution could be to accept the bias. After all, it is expected that prior knowledge has an impact on the parameter estimate. However, as discussed in Chapter 1, one should make sure that the impact is not too large. Alternatively, the parameters could be treated as being of purely mathematical nature with negative values allowed, ignoring their physical meaning. Technically, this can be achieved by replacing the Poisson distribution with a Gaussian distribution in the likelihood. Note that a direct estimation of uncertainties and limits is not possible or meaningful in this case. The (potentially negative) estimated parameter values need to be mapped onto physical quantities in a second step. If the mapping is done by cutting off the negative tails, the bias in the physical quantities will reappear.

#### 10.4.3

##### Including Efficiencies

If one is interested in the number of events expected to be produced (in contrast to the number of events expected to be produced *and* observed), then one has to consider acceptance effects and efficiencies. This is, for example, the case in cross-section measurements or in cases where the relative contributions or fractions of several processes are of interest, for example fractions of differently polarised particles. The efficiencies for the different processes  $j$ ,  $\epsilon_j$ , can have different values, for example because of the underlying kinematics and the selection criteria. In general, the efficiencies can depend on the variable  $x$  under study so that  $\epsilon_j = \epsilon_j^{\text{eff}}(x)$ . Efficiencies can be included into the fit by the following replacement in (10.19):

$$\nu_i = \sum_{j=1}^{NP} \nu_j^p \cdot \int_{\Delta x_i} \epsilon_j^{\text{eff}}(x) \cdot f_j(x) dx . \quad (10.25)$$

#### 10.4.4

##### Including Systematic Uncertainties

Systematic uncertainties can be interpreted as uncertainties on the functions  $f(x)$  in (10.25) or, as done in the following, they can be treated as uncertainties on the efficiencies. For each source of systematic uncertainty,  $k$ , a nuisance parameter,  $\delta_k^{\text{syst}}$ , is introduced. Its prior is often assumed to be a Gaussian with a mean value of zero and a standard deviation of one. For small deteriorations, the uncertainties on

8) Note that this bias is a purely frequentist concept since in its evaluation the experiment is repeated under the same conditions, that is using the same prior. In contrast, a Bayesian would update the prior knowledge after each experiment and would therefore obtain only a weak dependence on the initial assumptions.

the efficiencies can be assumed to be linear functions of  $\delta_k^{\text{syst}}$ , that is for a process  $j$  the efficiency becomes

$$\epsilon_j(x, \boldsymbol{\delta}^{\text{syst}}) = \epsilon_j^{\text{eff}}(x) \cdot \left( 1 + \sum_{k=1}^{N_{\text{syst}}} \delta_k^{\text{syst}} \cdot \Delta\epsilon_{jk}^{\text{syst}}(x) \right), \quad (10.26)$$

where  $N_{\text{syst}}$  is the number of systematic uncertainties considered and  $\Delta\epsilon_{jk}^{\text{syst}}(x)$  is the uncertainty on the efficiency of process  $j$  stemming from systematic effect  $k$ . Note that the total efficiency has to be positive although  $\Delta\epsilon_j^{\text{syst}}(x)$  can be negative in a certain region of  $x$ . This is for example the case if the overall efficiency does not change but only the shape of the distribution under study is affected.

Equation 10.20 then becomes

$$p(\mathbf{v}^{\text{p}}; \mathbf{n}) = \int \frac{p(\mathbf{n}; \mathbf{v}^{\text{p}}, \boldsymbol{\delta}^{\text{syst}}) \cdot \pi_0(\mathbf{v}^{\text{p}}) \cdot \pi_0(\boldsymbol{\delta}^{\text{syst}})}{\int p(\mathbf{n}; \mathbf{v}^{\text{p}}, \boldsymbol{\delta}^{\text{syst}}) \cdot \pi_0(\mathbf{v}^{\text{p}}) \cdot \pi_0(\boldsymbol{\delta}^{\text{syst}}) d\boldsymbol{\delta}^{\text{syst}} d\mathbf{v}^{\text{p}}} d\boldsymbol{\delta}^{\text{syst}}. \quad (10.27)$$

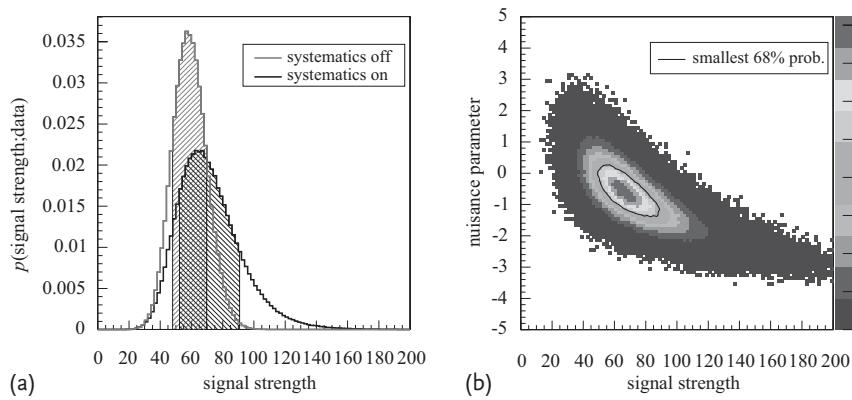
If the uncertainty on a parameter is estimated from the marginal distribution, then it now includes the systematic uncertainty. In this case, it is not possible to disentangle the statistical and systematic uncertainties although this is often requested during a review process (“What is the impact of this and that source of systematic uncertainty on the result?”). In order to assess the impact of a specific source of systematic uncertainty, it is best to evaluate (10.27) with and without a nuisance parameter and to compare the two resulting uncertainties (and estimators). Another option is to compare the outcome of the analysis including all uncertainties with the outcome of an analysis in which the systematic uncertainty under study was removed. The latter approach gives a hint on how the result would improve if one particular uncertainty would not be present. This can help to identify dominating sources of systematic uncertainty which should be worked on in further iterations of the analysis.

So far the impact of a source of systematic uncertainty was assumed to only depend linearly on the corresponding nuisance parameter. This may be a good approximation for sources of systematic uncertainty with a small impact on the shape of the efficiency. However, this approximation might fail for large uncertainties. In such cases, the efficiency can be parameterised as a suitable function of the nuisance parameter, for example a higher-order polynomial. For a linear parameterisation, the slope,  $\Delta\epsilon_{jk}^{\text{syst}}(x)$ , is usually obtained from Monte Carlo samples including a systematic variation of a parameter like the jet energy scale. In this case, the ‘scaled’ distribution should be retrieved from the ‘nominal’ one by setting the nuisance parameter  $\delta_k^{\text{syst}}$  to  $\pm 1$ .

For a non-linear parameterisation, one has to consider more complicated procedures, for example *template morphing* [11]. For an interesting discussion of systematic uncertainties for Poisson processes see also [12].

### Example 10.3 Systematic uncertainties and nuisance parameters

Let us assume that we are investigating the signal strength of a specific process in a sample of dimuon events with known background. Because of limited Monte Carlo statistics, the detection efficiency for the process is only known up to a certain level and is treated as a nuisance parameter. Figure 10.2a shows the difference in the posterior probability density function (pdf) obtained in an analysis with and without the uncertainty: the uncertainty on the signal strength is larger in the presence of the nuisance parameter (adding the uncertainty, the standard deviation of the pdf increases from 11 to 21); the mean value moves from 59 to 72. The correlation between the signal strength and the nuisance parameter is shown in Figure 10.2b. The solid curve represents the smallest interval containing 68% posterior probability. As expected, the dataset analysed has an impact on the knowledge of the signal strength *and* on that of the nuisance parameter.



**Figure 10.2** (a) Posterior pdf with and without the systematic uncertainty on the efficiency. The shaded region includes the central 68% probability. (b) The correlation between the signal strength and the nuisance parameter. The solid curve represents the smallest region covering 68% posterior probability.

#### 10.4.5

##### Systematic Uncertainties Due to the Fitting Procedure

The fitting procedure itself is fraught with several systematic uncertainties. These include the statistical fluctuations and/or smoothing bias from the templates which might be a particular problem for background samples if the potential signal is only found in the tails of the distributions. This should be kept in mind when generating the corresponding Monte Carlo samples. A further problem are badly modelled distributions. Such a mismodelling can arise from systematic uncertainties from theory (as discussed in Chapter 9), from a mismeasured detector or from neglected background processes. A goodness-of-fit test (see Section 3.8) can (but does not necessarily have to) be sensitive to these effects.

Another frequently discussed issue is the choice of the bin width for the variable under study. The bin width should be large enough to minimise the statistical fluctuations in each bin and small enough to resolve any structure which helps to discriminate the different processes from each other. The optimum bin width can be found using ensemble tests: a figure of merit, for example the expected sensitivity or total uncertainty, is calculated for ensembles generated with different settings, and the most suitable one is selected (see Section 10.5 for details). According to the discussion of the experimenter's bias (Section 10.6), such an optimisation should be performed *before* looking into the experiment's data.

#### 10.4.6

##### **Alternative Fitting Methods and Choice of Parameters**

Several alternatives to the likelihood defined in (10.21) exist and can be used for template fitting. If the Poisson terms are replaced by Gaussian terms, a classical  $\chi^2$  expression is restored. The standard deviation is usually chosen to be the square root of either the number of expected events or of the number of observed events (see the discussion of Pearson's [13] and Neyman's  $\chi^2$  definitions [14] in Section 2.5.4). Note that this ansatz is not recommended if the number of events in at least one bin is small since then the assumption of a Gaussian distribution of the number of observed events – at least in this one bin – is not justified.

In some cases, an alternative set of parameters to the number of expected events seems more natural, for example when studying the relative contributions of different subprocesses. One could use the sum of all processes,  $\nu$ , and the relative contributions  $f_j$  of the subprocesses as parameters. Note that only  $(N^p - 1)$  fractions can be considered free parameters due to the overall normalisation. Care must be taken when defining the likelihood and choosing the set of parameters. If the prior probabilities of the overall normalisation and the  $(N^p - 1)$  fractions are assumed to be flat, the probability distribution for the remaining fraction – calculated using propagation of uncertainties – will not be flat, and a bias is potentially introduced.

#### 10.4.7

##### **Extension to Multiple Channels and Multi-Dimensional Templates**

The above prescription can easily be extended to fit multiple channels simultaneously. An example could be simultaneous searches for resonances in dielectron and dimuon spectra. Usually, one would like to combine these two measurements to increase the overall sensitivity. If the two channels are uncorrelated, this can be achieved by multiplying likelihood terms of the type in (10.21):

$$p(\mathbf{n}; \boldsymbol{\nu}^p) = \prod_{i=1}^{N_{\text{bins}}} \prod_{j=1}^{N_{\text{ch}}} \frac{\nu_{ij}^{n_i}}{n_{ij}!} \cdot e^{-\nu_{ij}}, \quad (10.28)$$

where  $N_{\text{ch}}$  is the number of channels and all quantities with index pairs  $ij$  correspond to that quantity in bin  $i$  and channel  $j$ .

The template method can easily be extended to fit multi-dimensional distributions. In this case, the index  $i$  in (10.21) simply denotes the multi-dimensional bin.

## 10.5 Ensemble Tests

One is often interested in what the outcome of a future experiment could be. As an example, one might want to estimate the sensitivity of a future collider experiment to the detection of supersymmetric particles or the expected precision of a measurement of the mass of a certain particle. This sort of analysis is important in the design of experiments, that is when choosing one detector technology over another. Since an analytical approach, for example the minimum-variance bound as described in Chapter 2, is not always feasible, numerical methods are typically used for such predictions. One may also want to compare the outcome of an already conducted experiment with the expectation and evaluate  $p$ -values from the distribution of a test statistic to judge the validity of a certain model (see for example Chapter 3).

The questions mentioned above can be studied by performing *ensemble tests*, that is by repeating simulations of an experiment under the same conditions many times and analysing the outcome of these simulations. The procedure consists of three steps:

- In the first step, the experiment under consideration is represented by a statistical model,  $M$ . This model predicts possible outcomes of data,  $\mathbf{x}$ , given a set of parameters,  $\boldsymbol{\theta}$ , which belong to the model,  $p(\mathbf{x}; \boldsymbol{\theta}, M)$ . The latter expression is the same as the likelihood defined in Chapter 1.
- In the second step, ensembles are built. These are sets of possible outcomes,  $\mathbf{x}$ , of the experiment conducted under the same conditions. These outcomes are also referred to as *pseudo-data* as they result from *pseudo-experiments*. One ensemble could be represented, for example, by a single number, by a set of simulated events, or by a distribution of some quantity. When building ensembles, it is important to cover the whole available phase space. If this is not the case, then the conclusions drawn from the ensemble tests may be biased.
- The third step is the analysis of all pseudo-datasets. In everything that follows, these pseudo-data are analysed in the same way as real data. In most cases, either simple observables or complex, derived quantities are calculated – for example an observed number of events, the estimated invariant mass of a particle, or the limit on a cross section times branching ratio. The outcomes  $\mathbf{o}$  of the pseudo-experiments – for simplicity, assume that they are single numbers per pseudo-experiment – are typically presented as histograms. These frequency distributions,  $\hat{f}(\mathbf{o})$ , are an estimate of the corresponding probability density function  $f(\mathbf{o})$ . They can then be used to directly calculate expectation values, variances and other descriptive quantities, or to calculate  $p$ -values when comparing the distributions with real experimental data. A typical application is the test of the bias of an estimator and its predicted accuracy.

What do the ensemble tests tell us? They show the expectation we would get if we were to repeat an experiment under the same conditions many times. This directly relates to the definition of a probability (density) in the frequentist interpretation. Note, however, that ensemble tests are not done to estimate a parameter. Furthermore, it is noteworthy that ensemble tests can be done in frequentist and Bayesian analyses.

### 10.5.1 Generation of Ensembles

What are ensembles made of? In simple examples, the frequency distributions can be calculated analytically. The modelling of a real collider experiment is usually more difficult, and numerical simulation programs consisting of two parts are used: physics models are implemented in Monte Carlo generators that produce simulated events consisting of a set of final-state particles. These particles are then tracked through a simulated detector volume, and the detector response is simulated. The prediction from such a model can thus be written as

$$p(\mathbf{x}; \boldsymbol{\theta}, M) = \int d\gamma p(\mathbf{x}; \gamma, \boldsymbol{\theta}, M_{\text{det}}) \cdot p(\gamma; \boldsymbol{\theta}, M_{\text{phys}}), \quad (10.29)$$

where  $M_{\text{det}}$  and  $M_{\text{phys}}$  model the detector and the physical process, respectively.  $\gamma$  represents the potential ‘true’ (and unknown) values of a set of observables. Note that the parameters  $\boldsymbol{\theta}$  can belong to either of the two models. Detector parameters are often used as nuisance parameters and introduced in the context of systematic uncertainties. In the end, it is usually the parameters of the physics model that one is interested in. The outcome of these simulations represents a complete model of the real data.

How are ensembles built? The simulation of processes is usually limited by CPU time and storage space, so that only a finite number of events,  $N$ , is simulated. These events can then be partitioned into  $M$  statistically independent ensembles of  $n$  events each. The obvious drawback is that a large number of events is needed to make reliable statements about the underlying population [15].

An alternative to the partitioning approach is the following: ensembles are generated by randomly drawing a number of events,  $n$ , from the available Monte Carlo sample of  $N$  events. The procedure of drawing subsets from a finite sample is known as *resampling* or *bootstrapping* [16]. Here,  $n$  can either be a fixed number, that is all ensembles consist of the same number of events, or it can fluctuate, for example when a certain run time or luminosity is fixed. In the latter case, the number of events in each ensemble varies according to a Poisson distribution. It is important not to remove events already drawn from the Monte Carlo population; one should draw events *with replacement* in order not to bias the composition of the ensembles. If  $n$  is small compared to  $N$ , the ensembles can be assumed to be independent. If, conversely,  $n$  is of the same order of magnitude as  $N$ , the ensembles are not independent since one particular event may occur in several ensembles or an event can occur in one ensemble several times. This can cause biases in the expectation

values and variances. In practice, one should always generate Monte Carlo samples that contain a multiple of the observed (or expected) number of events. The bias from using the same events multiple times vanishes for a large number of events  $N$ .

In some cases it is possible to generate ensembles from a parameterisation. In the simple case of a counting experiment, for example, there is no need to go through the full detector simulation if the detection efficiency is known. Instead of counting the number of selected events in each ensemble (each corresponding to a defined luminosity times cross section), it is sufficient to generate random numbers according to a Poisson distribution with an expectation value given by the product of the efficiency, the luminosity and the cross section times branching ratio. The advantage of using a parameterisation is that ensembles can usually be generated more easily and without the constraint of small statistics. The parameterisation has to be good enough to describe all relevant features of the data. Features in parameterised distributions, such as small bumps or slopes, need not necessarily be of statistical nature but could be due to effects which are not considered in the parameterisation. Therefore, a thorough comparison of the distribution from simulations and the parameterisation should be performed. If the agreement is not satisfactory, a systematic uncertainty has to be assigned to cover the effect.

If a model has parameters, it might not be obvious which numerical values these should take when generating ensembles. If these parameter values are predicted or asserted, for example a certain cross section, then they can be fixed. Conversely, if experimental data are available and used for inference, the best estimates of the parameters can be used instead. If experimental data are not available, then in the context of Bayesian inference one could also vary the parameters according to their priors during the generation of ensembles. Technically, this is best done by giving a weight to each ensemble based on the prior probability. This weight can be changed if different priors are tested. Integrating over the prior probability then not only models the experiment but also the belief of the analyser in the model and in the corresponding parameter values.

### 10.5.2

#### Results of Ensemble Tests

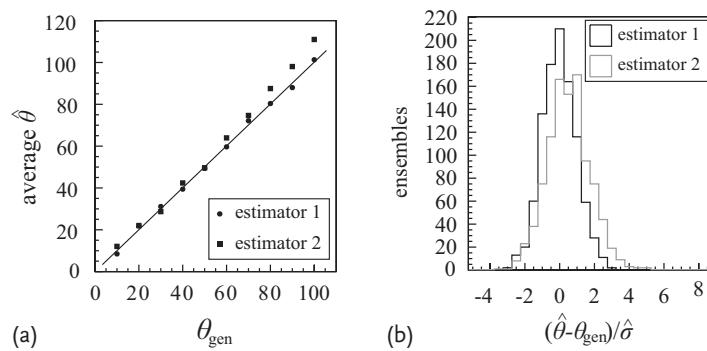
What are the concrete results of ensemble tests? Ensemble tests can be used to test the properties of an estimator, such as consistency, efficiency and bias. This procedure is sometimes referred to as a *closure test*. The most prominent tools for investigation are the *pull distribution* and the linearity of an estimator: for a set of ensembles generated using a parameter value  $\theta_{\text{gen}}$ , the pull is defined as the difference between the estimator  $\hat{\theta}$  and  $\theta_{\text{gen}}$ , divided by the estimated uncertainty,  $\hat{\sigma}(\hat{\theta})$ . For estimators with Gaussian uncertainties, the distribution of the pull is expected to be a Gaussian distribution itself, with a mean of zero and a standard deviation of one. If the mean value deviates from zero the estimator is biased. If the standard deviation deviates from unity, the estimator has a variance which is either too small or too large compared to the statistical fluctuations. In both cases it

is important to investigate the cause of this behaviour. Note that if the model is not Gaussian, the expected shape of the pull distribution will most likely differ from this Gaussian expectation, see [17] for further details.

For the *closure* or *linearity test* of the estimator, the expectation value of an estimator,  $E[\hat{\theta}]$ , can be calculated for sets of ensembles generated using different parameter values  $\theta_{\text{gen}}$ . Typically, this expectation value is plotted as a function of  $\theta_{\text{gen}}$  using the standard deviation of the average  $\hat{\theta}$  as the uncertainty. If the estimator is unbiased, a fit through the set of points is expected to be consistent with a straight line with an offset of zero and a slope of one within uncertainties. Deviations from this behaviour should be investigated and understood, and can be used for an empirical correction. For these plots, it is important to note that the correlation among the different ensembles can cause a deterioration from the expectation.

#### Example 10.4 Properties of estimators

The properties of two different estimators are studied. Figure 10.3a shows the average estimators as a function of  $\theta_{\text{gen}}$ . While estimator 1 (full circle) is unbiased, estimator 2 (squares) shows a bias for large values of  $\theta_{\text{gen}}$ . Such a bias will have to be corrected for in a real analysis. Figure 10.3b shows the pull distribution of the two estimators for a fixed value of  $\theta_{\text{gen}}$ . The pull of estimator 1 (black histogram) has a mean value of  $-0.02 \pm 0.03$  and a standard deviation of  $0.98 \pm 0.02$  – the estimator is unbiased and the estimated statistical fluctuations correspond to those observed among the different ensembles. In contrast, the pull of estimator 2 has a mean value of  $0.50 \pm 0.04$  and a standard deviation of  $1.22 \pm 0.03$  – the estimator has a significant bias and, even worse, underestimates the statistical uncertainty.



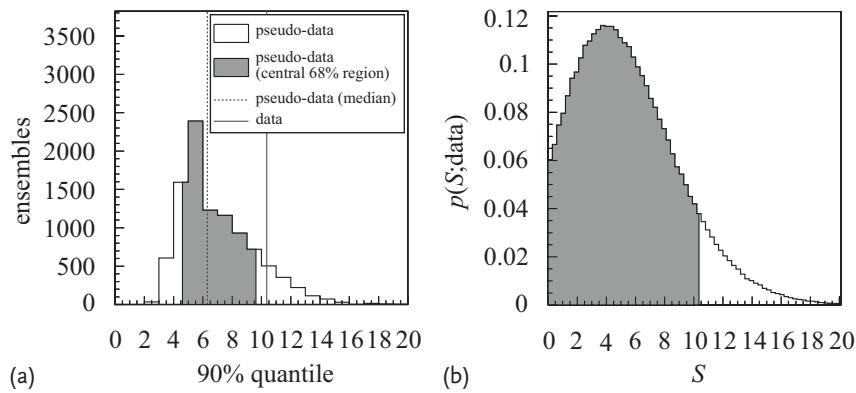
**Figure 10.3** (a) Average estimator  $\hat{\theta}$  of two different estimators as a function of  $\theta_{\text{gen}}$ . To guide the eye, the black line with a slope of one and an offset of zero was added. While estimator 1 (full circle) is unbiased, estimator 2 (squares) shows a bias for large values of  $\theta_{\text{gen}}$ . (b) The pull distribution for the two

estimators. The pull for estimator 1 (black histogram) has a mean of zero and a width of unity, while the pull for estimator 2 (grey histogram) has a shift to larger values and a standard deviation significantly larger than one.

When inferring parameters, one usually wants to compare the value of an estimator obtained from data,  $\hat{\theta}(\mathbf{x} = \mathbf{D})$ , with the distribution of possible outcomes. Here,  $\mathbf{x}$  can be any (pseudo-)data and  $\mathbf{x} = \mathbf{D}$  represents the data obtained from the (real) experiment. The distribution is typically calculated by building ensembles using as true parameter values  $\theta = \hat{\theta}(\mathbf{D})$ , that is the best-fit parameter values found in the real data. Typical distributions include the distribution of the estimator itself,  $\hat{\theta}(\mathbf{x})$ , the distribution of the uncertainty of the estimator,  $\hat{\sigma}(\hat{\theta})$ , and, if applicable, the distribution of the limit set on  $\theta$ , for example the 90% probability upper limit,  $\hat{\theta}_{90}(\mathbf{x})$ , for ensembles  $\mathbf{x}$  generated under the null hypothesis, that is in the absence of a signal. Typically, the values obtained from data are indicated in these distributions. It is then possible to compare the expected distribution with the results from data by eye or by calculating  $p$ -values. This will help to judge if the estimator and its properties are consistent with the expectation. If this is not the case, one should try to understand the origin of the discrepancy. An example of such a deviation is a significant discrepancy between observed and expected limits in searches for new particles. Such a discrepancy could hint towards a contribution to the data not modelled in the pseudo-data.

#### Example 10.5 Search for a new particle

Let us assume that a new particle and its decays into two muons are predicted by some theory to have a certain cross section. An experiment is built to test the theory, and one expects ten background events. Before looking at the data collected by the experiment, ensemble tests are performed: pseudo-datasets containing only



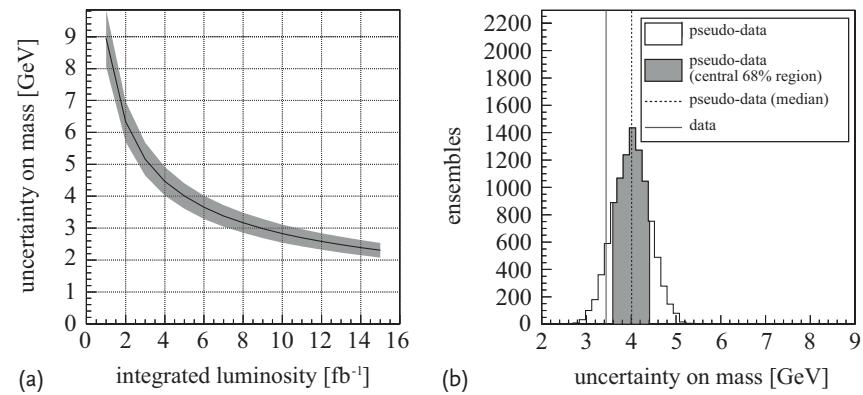
**Figure 10.4** (a) Distribution of the 90% quantiles on the signal contribution  $S$  extracted from 10 000 pseudo-datasets containing only background events (with expectation value 10). The expected limit on the signal is  $6.2^{+3.5}_{-1.6}$  events. The shaded area represents the central 68% probability region and the dashed line indicates the median. The ob-

served limit is represented by the solid line and agrees with the expectations. (b) The posterior probability  $p(S; \text{data})$  of the signal contribution  $S$  for 14 observed events. The shaded area shows the posterior probability region up to the 90% quantile, corresponding to 10.4 events.

background events (drawn from a Poisson distribution with ten expected events) are generated and analysed using a Bayesian approach. A limit on the signal contribution  $S$ , for example the 90% quantile of the posterior pdf, is calculated and histogrammed for all pseudo-datasets. Figure 10.4a shows such a distribution. The median of this distribution (shown as a dashed line) is typically used to define the *expected limit*. The shaded band shows the central 68% probability region, that is the typical variation of the limit from ensemble to ensemble. In this case a limit of  $6.2^{+3.5}_{-1.6}$  is expected. The analysis is then repeated for the data collected with the experiment (14 events observed), resulting in a limit on the signal contribution of 10.4 events based on the posterior probability shown in Figure 10.4b. This *observed limit* is also shown in Figure 10.4a as a solid vertical line and is a bit higher than the expectation from the background-only ensemble test.

When planning an experiment or analysis, it is often useful to estimate the expected statistical precision. In analogy with the linearity test, one often plots the expected uncertainty  $E[\sqrt{\hat{V}(\hat{\theta})}]$  and, if applicable, the expected limit on  $\theta$  for different scenarios, for example luminosities, assumptions on the detector performance, or values of the parameter  $\theta_{\text{gen}}$ . One usually shows the expected systematic uncertainties as well as the total uncertainties in one plot to illustrate the relative importance of the two. Comparing the distributions for different estimators, the expectation value of these distributions can help to find the most precise one. For limits, the mean value or the median are often used to represent the expected limit, and the standard deviation or a suitable set of quantiles are used to represent the expected fluctuations. A concrete example of such expected limits and ensemble tests is given in [18].

#### Example 10.6 Measuring the mass of a resonance



**Figure 10.5** (a) Expected uncertainty on the mass of the new particle as a function of the integrated luminosity. The shaded band indicates the statistical uncertainty on the prediction. (b) The distribution of uncertainties derived from 10 000 ensembles, for an inte-

grated luminosity of  $5 \text{ fb}^{-1}$ . The shaded area represents the central 68% probability region and the dashed line indicates the median. The observed uncertainty is represented by the solid line and agrees with the expectation within a bit more than one standard deviation.

Let us assume that the particle searched for in Example 10.5 was found at a new collider and that now its mass is being measured. An estimate of the predicted uncertainty as a function of the luminosity was performed before the discovery of the particle, see Figure 10.5a. The shaded band indicates the uncertainty on the estimate. In this example, the expectation values of the (rather complicated) mass estimator itself and the uncertainty estimate cannot be calculated analytically and thus ensemble tests were used instead. The new collider collected data corresponding to  $5 \text{ fb}^{-1}$ . Figure 10.5b shows the distribution of possible observed uncertainties on the mass together with the median and the actually observed value. It turns out that the experimentalists were lucky and observed a slightly smaller uncertainty than expected (although the difference is not worrisome).

A further important use case of ensemble tests is the estimation of systematic uncertainties. One possible approach is to re-run the analysis on ensembles generated under different conditions, for example with a different jet energy scale, with an alternative signal model, or with different event weights. The difference between the expectation values of the estimator obtained with the original and the altered ensembles can be quoted as (the expected) systematic uncertainty.

One can also use ensemble tests to investigate the correlation of two estimators when combining results. If two analyses are applied on datasets which are not statistically independent, for example the measurement of the top-quark production cross section with and without the requirement of a  $b$ -tag, then the results will be correlated. Frequently used combination methods, like the *best linear unbiased estimator* (BLUE) [19], use this correlation as an input. It can be estimated by repeating the two analyses on the same pseudo-datasets and then calculating the correlation coefficient of the two estimators.

## 10.6

### The Experimenter's Role and Data Blinding

An important part of every measurement is the evaluation of statistical and systematic uncertainties. While the former are defined by the statistical model, the latter ones are not necessarily easy to spot and quantify. A thorough discussion of systematic uncertainties and how to circumvent or estimate them is given in Chapters 8 and 9. This section focuses on a unique subset of systematic uncertainties: biases due to the experimenter's preconception. These biases can (mostly unintentionally) be caused by the experimenter knowing the result or favouring one result over the other. For example, it is more appealing to publish a result if it describes the discovery of a new phenomenon rather than to only set a limit. On the contrary, new measurements of well-known quantities might not be published directly if the result obtained differs significantly from an established world average.

One strategy to avoid an experimenter's bias is to *blind* (or *mask*) the data, or to artificially alter the data at hand. The procedure is referred to as performing a *blind*

*analysis.* There is no silver-bullet technique – the choice of a particular blinding procedure strongly depends on the analysis itself.

The probably best-known example of blinding is that chosen for drug testing in medical sciences. The effect of a drug is tested by treating one group of patients with the drug itself and another group with a placebo. One speaks of *single-blind tests* if the patient does not know to which of the two groups he or she belongs. In a *double-blind test*, neither patients nor researchers know whether the drug or a placebo was administered.

This section describes several methods of data blinding and gives examples of its usage in high energy physics.

#### 10.6.1

##### The Experimenter's Preconception

In high energy physics, the following sources of experimenter's bias might be encountered:

- data selection: A cut-based analysis requires a cut optimisation. This process is predestined to allow experimenter's bias if the optimisation is done when the data have already been analysed. This step might include the re-binning of histograms which can have a significant impact on the statistical significance of a result;
- cross-checks of the analysis;
- omission or addition of systematic uncertainties if the measured value does or does not correspond to the one obtained in previous experiments;
- prejudices concerning theoretical models or calculations;
- personal preferences towards or against the results of one experiment over that from another experiment.

The following section describes different blinding techniques used in high energy physics that might be helpful in avoiding these biases and gives working examples.

#### 10.6.2

##### Variants of Blind Analyses

**Hiding the answer** The simplest, and probably most idealistic, procedure of performing a blind analysis is to simply not calculate the result until the very last step. Analysers should also avoid producing intermediate results that might lead to building an anticipation of the final outcome. The *hidden signal box method* is a special representative of this approach.

**The hidden signal box method** In searches for rare processes, it might be known in which part of the phase space the signal should occur. One can then blind this phase-space region (the *signal box*). Optimisation studies should be performed us-

ing only simulated signal and background samples, and tests of the background simulation should be carried out in signal-free sideband or control regions. The signal box should only be opened after the analysis procedure has been completely frozen.

A simple example are experiments searching for neutrinoless double- $\beta$  decay, a rare process predicted to occur if the neutrino is its own antiparticle. In most such searches, the  $Q$ -value of the decay is known and the expected signal is a mono-energetic line in an energy spectrum. The signal region could thus be defined as the region around the expected peak with a width equal to a multiple of the energy resolution of the detector.

The hidden signal box method was used for example by the XENON10 collaboration [20]. The experiment was built to search for *weakly interacting massive particles* (WIMPs) by simultaneously measuring scintillation light and ionisation of liquid xenon produced by the elastic scattering of WIMPs off atomic nuclei. The two-dimensional signal box was defined before the analysis was performed and it was only looked at after the analysis procedure was established.

Blinding the signal region is not an option for analyses in which the location of a signal is not a priori known, for example when searching for bumps in an invariant mass spectrum. Blinding methods applicable to such cases are described below.

**Shifting the answer** In some cases, removing the answer and thus completely blinding the data is not necessary or even possible, for example in precision measurements. In such examples, one way to blind the sensitive part of the data is to shift the estimator of the parameter under study by an amount unknown to the analyser, that is by adding a constant offset. This offset should be generated by a Gaussian pseudo-random number generator with fixed seed and centred around zero with a width close to the estimated uncertainty on that parameter. This procedure allows the analysis to be tuned using data. It also allows two independent analyses to be compared without unblinding the results.

#### Example 10.7 CP violation in $B^0$ decays

An example of this procedure is the measurement of a CP-violating asymmetry in the time distributions of decays of  $B^0$  and  $\bar{B}^0$  mesons to CP eigenstates performed by the BABAR collaboration [21] and discussed in [22, 23]. The asymmetry is built from the difference of the decay rates of the  $B^0$  and  $\bar{B}^0$  as a function of the decay time,  $\Delta t$ . By construction,  $\Delta t$  can also be negative because it is defined as the difference between the time at which one of the two mesons decays into a CP eigenstate and the time at which the other one is flavour-tagged. The two distributions of  $\Delta t$  are not centred around  $\Delta t = 0$  in the presence of CP violation, and they are approximately equal, but opposite in the sign of  $\Delta t$ .

During the design of the analysis procedure, the individual decay-time distributions needed to be carefully studied, and a potential asymmetry in the data would have already been observed by visual inspection of the measured  $\Delta t$  distributions.

As this could have caused an unwanted bias due to the analysers, the asymmetry

was hidden by modifying the observable to

$$\Delta t_{\text{blind}} = s_{\text{tag}} \cdot \Delta t_{\text{measured}} + x , \quad (10.30)$$

where  $s_{\text{tag}}$  is either  $+1$  or  $-1$ , depending on the flavour tag. This reflects one of the two distributions onto the other and significantly reduces the measured asymmetry. Indications of CP violation in the individual  $\Delta t$  distributions were hidden by adding an offset  $x$  which moved the  $\Delta t = 0$  mark to an arbitrary position.

**Removing or adding data** In some data analyses, signal properties play an essential role in the analysis optimisation. The described blinding methods do not work under such circumstances. One possibility of blinding the data can then be to add or remove parts of the data. For the measurement of the solar neutrino flux by the SNO collaboration, a certain fraction (unknown to the analysers) of the data [24] was removed. In this way the analysis could be conducted using the signal information while the actual result, the final neutrino event count, was blinded.

## 10.7 Exercises

### Exercise 10.1 Properties of an estimator

Generate 100 000 sets of ten random numbers drawn from a flat distribution between 0 and 1. Estimate the (true) mean value and the (true) standard deviation with and without Bessel corrections. Show that the Bessel corrections are needed to obtain an unbiased estimator of the standard deviation. Repeat the exercise with 100 events and compare the effect of the Bessel corrections.

### Exercise 10.2 Neutrinoless double- $\beta$ decay

Neutrinoless double- $\beta$  ( $0\nu\beta\beta$ ) decay is a rare process predicted to occur if the neutrino is its own antiparticle. The experimental signature is a sharp peak in the energy spectrum of the sum of the two final-state electrons. The peak position is known and the lineshape is typically a Gaussian with known width, usually not much larger than the energy resolution of the detector. For that reason, one typically counts the number of observed events in a fixed energy range in which most signal events would appear. Let us assume that a hypothetical experiment expects 1.5 background events per month in that energy range.

- a) Calculate the expected exclusion limit on the signal process for durations between 1 month and 5 years based on ensemble tests: build the ensembles under the background-only hypothesis assuming that the observed numbers of events fluctuate according to a Poisson distribution. Calculate the (Bayesian)

95% probability limit on the hypothetical signal process for each ensemble and determine the ensemble average, mode and standard deviation of this limit.

- b) The experiment runs for 4 years and 100 events are observed. Calculate the observed 95% probability limit on the hypothetical signal process. Compare the observed number of events with the distribution of the expected number of background events. Is the observed number of events still in agreement with the null (i.e. background-only) hypothesis? Calculate the  $p$ -value.
- c) Re-calculate the limit in (b) assuming that the expected background is only known up to 10% precision. Include this uncertainty as a Gaussian prior probability in your counting experiment. What is the correlation between the signal and background parameters? This correlation can be estimated from the two-dimensional posterior probability.

### Exercise 10.3 Estimating a signal strength

We now want to estimate the signal strength for the example given above. This is done by a fit of signal and background templates to the observed data. The signal is assumed to be a Gaussian distribution with a mean value of 2039 keV – the position of the  $0\nu\beta\beta$  signal in  $^{76}\text{Ge}$  – and a standard deviation of 5 keV – the energy resolution of a typical germanium detector. The background is assumed to be distributed uniformly over the analysed mass range with a total expectation value of 1.5 events per month. Again, the experiment runs for 4 years. The data are binned in 5 keV bins and given below.

Bin center	2014	2019	2024	2029	2034	2039
Events	2	7	5	11	3	5
<hr/>						
Bin center	2044	2049	2054	2059	2064	
Events	10	6	5	8	10	

Perform the template fit assuming a 10% uncertainty on the background. Use three different likelihoods for the fit and compare the 95% limit on the signal contribution:

- a) A product of Poisson distributions, that is  $L = \prod_i (\nu_i^{n_i} / n_i!) e^{-\nu_i}$ .
- b) A least-squares ansatz assuming a per-bin uncertainty of  $\sigma_i = \sqrt{\nu_i}$ .
- c) A least-squares ansatz assuming a per-bin uncertainty of  $\sigma_i = \sqrt{n_i}$ .

## References

- 1 Straessner, A. and Schott, M. (2010) A new tool for measuring detector performance in ATLAS. *J. Phys. Conf. Ser.*, **219**, 032023.
- 2 ATLAS Collab. (2012) Performance of the ATLAS muon trigger in 2011, ATLAS-CONF-2012-099.
- 3 ATLAS Collab. (2012) Performance of the ATLAS electron and photon trigger in  $p\text{-}p$  collisions at  $\sqrt{s} = 7\text{ TeV}$  in 2011, ATLAS-CONF-2012-048.
- 4 Casadei, D. (2012) Estimating the selection efficiency *JINST*, **7**, P08021.
- 5 ATLAS Collab. (2011) Measurement of the charge asymmetry in top quark pair production in  $p\text{-}p$  collisions at  $\sqrt{s} = 7\text{ TeV}$  using the ATLAS detector, ATLAS-CONF-2011-106.
- 6 Caldwell, A., Kollar, D., and Kröninger, K. (2009) BAT: The Bayesian Analysis Toolkit. *Comput. Phys. Commun.*, **180**, 2197.
- 7 Rosenblatt, M. (1956) Remarks on some non-parametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832.
- 8 Parzen, E. (1962) On the estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065.
- 9 Barlow, R.J. and Beeston, C. (1993) Fitting using finite Monte Carlo samples. *Comput. Phys. Commun.*, **77**, 219–228, doi:10.1016/0010-4655(93)90005-W.
- 10 Brun, R. and Rademakers, F. (1997) ROOT: An object oriented data analysis framework. *Nucl. Instrum. Methods A*, **389**, 81.
- 11 Read, A.L. (1999) Linear interpolation of histograms. *Nucl. Instrum. Methods A*, **425**, 357.
- 12 Conway, J.S. (2011) Incorporating nuisance parameters in likelihoods for multisource spectra. arXiv:1103.0354.
- 13 Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.*, **50**, 157.
- 14 Neyman, J. (1949) *Contribution to the Theory of the  $\chi^2$  Test*. Proc. Berkeley Symp. Math. Stat. Probab., Berkeley and Los Angeles, University of California Press, p. 29.
- 15 Barlow, R. (2000) Application of the Bootstrap resampling technique to particle physics experiments. *Manchester Part. Phys.*, MAN/HEP/99/4.
- 16 Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1.
- 17 Lyons, L. and Demortier, L. (2002) Everything you always wanted to know about pulls, CDF Note-5776.
- 18 Caldwell, A. and Kröninger, K. (2006) Signal discovery in sparse spectra: a Bayesian analysis. *Phys. Rev. D*, **74**, 092003.
- 19 Lyons, L., Gibaut, D., and Clifford, P. (1988) How to combine correlated estimates of a single physical quantity. *Nucl. Instrum. Methods A*, **270**, 110.
- 20 XENON Collab., Angle, J. et al. (2008) First results from the XENON10 dark matter experiment at the Gran Sasso national laboratory. *Phys. Rev. Lett.*, **100**, 021303.
- 21 BABAR Collab., Aubert, B. et al. (2001) Measurement of CP violating asymmetries in  $B^0$  decays to CP eigenstates. *Phys. Rev. Lett.*, **86**, 2515.
- 22 Roodman, A. (2003) Blind analysis in particle physics. *eConf*, C030908, TU-IT001.
- 23 Klein, J. and Roodman, A. (2005) Blind analysis in nuclear and particle physics. *Ann. Rev. Nucl. Part. Sci.*, **55**, 141.
- 24 SNO Collab., Aharmim, B. et al. (2008) An independent measurement of the total active  ${}^8\text{B}$  solar neutrino flux using an array of  ${}^3\text{He}$  proportional counters at the Sudbury neutrino observatory. *Phys. Rev. Lett.*, **101**, 111301.

## 11

### Analysis Walk-Throughs

*Aart Heijboer and Ivo van Vulpen*

#### 11.1

##### Introduction

The goal of this chapter is to apply some of the analysis concepts described in the earlier chapters using two simple and semi-realistic analysis examples: a *search* for a hypothetical particle and a *measurement* of its properties once it has been established. Although the analysis is simplified, it provides a skeleton to address a few of the numerous statistics issues and provides references to more complex concepts as discussed in earlier chapters. We hope that the discussions and exercises will help you to grasp some of the basics of statistics used by your physicist colleagues to report their measurements.

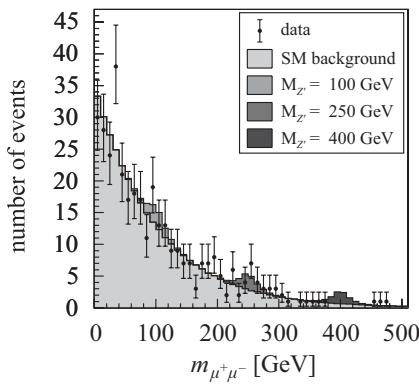
As reading about statistics is really no substitute for working out a few concrete examples, each section comes with a set of exercises. They address the basics by reproducing the results in the text, but also let you struggle with some of the more complex issues. Although most of the modelling and statistics tools used in high energy physics is coded in the `root` analysis package (in particular in `ROOFIT` [1] and `ROOSTARS` [2]), the examples used here are based on `root` macros. Working on some of the steps yourself will make you appreciate the speed and sophistication of `ROOFIT`.

#### 11.2

##### Search for a $Z'$ Boson Decaying into Muons

Several extensions of the Standard Model (SM) predict the existence of a new gauge group [3]. The aim of the analysis presented here is to establish the presence of the associated gauge boson in the data by searching for a  $Z' \rightarrow \mu^+ \mu^-$  resonance on top of a SM background in the spectrum of the dimuon mass,  $m_{\mu^+ \mu^-}$ . For this purpose we have simulated (pseudo-)datasets<sup>1)</sup> for the SM background (we left out the  $Z$ -boson peak) and  $Z'$  signals at various masses. Similar to the Higgs boson,

1) The histograms are available in the file `DataSample_search.root`.



**Figure 11.1** Distribution of the dimuon invariant mass spectrum from SM processes and from three hypothetical signals with  $Z'$  masses of 100, 250 and 400 GeV.

although its mass is unknown, the cross section of  $Z'$  production is known as a function of its mass. We will refer to this as the *nominal*  $Z'$  cross section,  $\sigma_{\text{nom}}$ , in the text. The natural width of the  $Z'$ -boson is assumed to be small compared to the experimental resolution. The simulated backgrounds together with the simulated signal for three different  $Z'$  masses and an assumed integrated luminosity of  $1 \text{ fb}^{-1}$  are shown in Figure 11.1. Besides the mass information, the events are, of course, characterised by additional properties (like angular information) that will be investigated in Section 11.3 and in the exercises.

The next sections will address the following questions: how can we quantify the compatibility with the SM ('background-only', ' $b$ -only') and SM +  $Z'$  ('signal-plus-background' or ' $s + b$ ') hypotheses? How can we optimise the sensitivity to a possible signal? How can we interpret the excess at 250 GeV? Can we exclude  $Z'$  hypotheses at other masses? How can we summarise the results from this search?

### 11.2.1 Counting Experiment

If a  $Z'$ -boson of 250 GeV is present in the data, it is expected to show up as an excess of events in the dimuon mass distribution around 250 GeV. As a first step we will assume this mass to be known and will count events in a window around it.

#### 11.2.1.1 Quantifying the Sensitivity: $p$ -Values and Significance

A measure to judge the deviation of the observed number of events from the expectation according to the  $b$ -only hypothesis is the  $p$ -value. It is the probability to observe as many events as were found in the data  $n_{\text{obs}}$ , or more, assuming that only SM processes contribute. In a counting experiment with a perfectly known

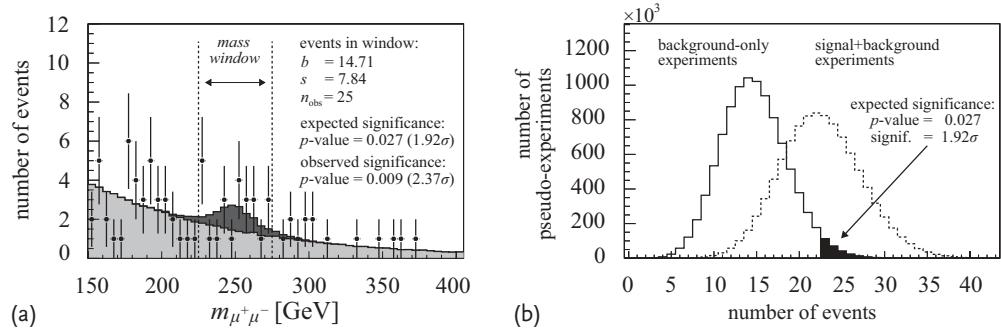
expected background  $b$  the  $p$ -value is given by

$$p = \sum_{n=n_{\text{obs}}}^{\infty} f(n; b), \quad (11.1)$$

where  $f(n; b)$  is the Poisson distribution introduced in Chapter 1. It is common practice to convert the  $p$ -value into an *observed significance*  $Z$  corresponding to a unit Gaussian:<sup>2)</sup>

$$\int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = p. \quad (11.2)$$

The observed significance is expressed in units of sigma ( $\sigma$ ), that is the number of standard deviations of the unit Gaussian. The *expected significance*, or expected deviation of a signal from the  $b$ -only expectation, is defined as the significance associated with the median number of expected events under the  $s + b$  hypothesis. It can be used to express an analysis' or experiment's strength to separate the two hypotheses using a single number. As an example, Figure 11.2a shows the number of expected and observed events as a function of the dimuon mass. In a 50-GeV mass window around 250 GeV (indicated by the two vertical dashed lines), one expects on average 14.71 background events and 7.84 signal events; 25 data events are observed. In a next step, pseudo-data are generated using the expected numbers of events for the  $b$ -only and  $s + b$  hypotheses (the concept of pseudo-data and ensemble tests is explained in Chapter 10). Figure 11.2b shows the resulting distributions of the numbers of events for the two hypotheses. The median number



**Figure 11.2** (a) The dimuon mass distribution. The dotted lines indicate the mass window around 250 GeV. A summary of the number of expected and observed events in the mass window is also shown. (b) Distribution of the number of events for the  $b$ -only (solid line) and  $s + b$  (dashed line) hypotheses derived from pseudo-data. A graphical representation of the expected significance is also indicated.

bution of the number of events for the  $b$ -only (solid line) and  $s + b$  (dashed line) hypotheses derived from pseudo-data. A graphical representation of the expected significance is also indicated.

- 2) The definition of observed significance comes as single- and double-sided, that is it sometimes involves an additional factor 2. This is subtle but important when comparing

results. Here, the single-sided version is used, as we search for an excess over the expectation.

of events for the  $s + b$  hypothesis is 22.55; this corresponds to a  $p$ -value of 0.0271 (indicated by the filled region in the plot), which translates into an expected significance of  $1.92\sigma$ . If a signal is present, the significance that will be observed in experiments will fluctuate around this median value. In our data sample, 25 events are observed, which is more than was expected from a  $Z'$  signal at 250 GeV. Because of this fluctuation, the observed significance of  $2.37\sigma$  is slightly higher than the expected significance.<sup>3)</sup>

### 11.2.1.2 Optimising the Mass Window

The choice of the mass window of 50 GeV around a  $Z'$  mass of 250 GeV in the example above was arbitrary. Using the expected numbers of signal and background events we can search for the mass window that optimises the expected significance. The size of the mass window should be chosen to yield the optimal compromise between signal efficiency and background rejection. This optimal size depends on the background uncertainty, the luminosity, and so on. An important remark to make here is that the optimal window should be determined, and fixed, based on the expectation *before* looking at the data (see also Section 10.6). Exercise 11.1 addresses some of the issues when searching for the optimal mass window.

### 11.2.1.3 Estimating the Background from Data Using Sidebands

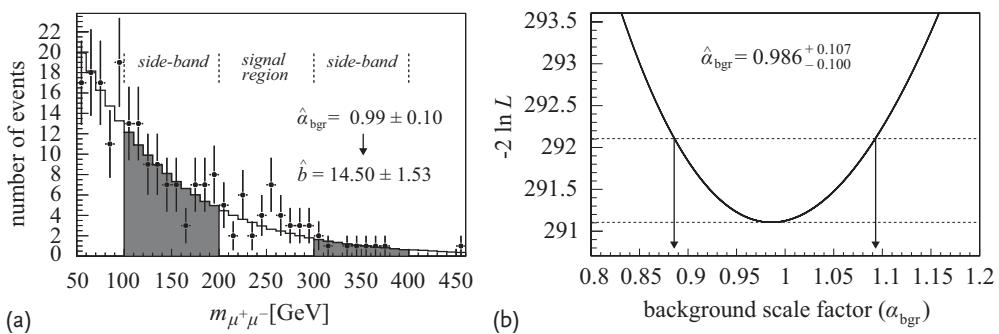
The expected  $Z'$  signal sits on top of a large SM background which could be estimated using Monte Carlo (MC) event generators. However, it might be preferable to extract a background estimate from the data themselves – especially if the uncertainties on the simulation are large or the simulations do not describe the data sufficiently well. In our example we assume we understand the shape of the SM background, but we will estimate its normalisation from data using a signal-free region (in our case  $100 < m_{\mu^+\mu^-} < 200$  GeV and  $300 < m_{\mu^+\mu^-} < 400$  GeV, see Figure 11.3a). In our model the scale factor for the SM contribution with respect to the MC expectation is denoted by  $\alpha_{\text{bgr}}$ . Similarly, the signal scale factor is denoted by  $\mu_s$ . Now, if we trust the shape of the background, we can obtain an estimate of  $\alpha_{\text{bgr}}$  by maximising the likelihood of the observed number of events in the bins that make up the sideband regions:

$$L(\mu_s, \alpha_{\text{bgr}}) = \prod_{\text{bins } i} f(n_i; \alpha_{\text{bgr}} b_i + \mu_s s_i) , \quad \text{with } \mu_s = 0 , \quad (11.3)$$

where  $f$  is again the Poisson distribution and  $b_i$  and  $s_i$  are the predicted numbers of background and signal events in bin  $i$ , respectively. The result of the sideband fit using (11.3) is shown in Figure 11.3b. The estimate of the scale factor,  $\hat{\alpha}_{\text{bgr}} = 0.99 \pm 0.10$ , shows that the data are in good agreement with the expectation. This value leads to an estimate of the expected number of background events in the signal region of  $\hat{b} = 14.50 \pm 1.53$ .

- 3) An estimate of the observed significance in a counting experiment, where  $n_{\text{obs}}$  events were observed and given an expectation for background  $b$  and signal  $s$ , is given by

$Z = \sqrt{2n_{\text{obs}} \ln(1 + s/b) - 2s}$ . The expected significance can be computed by replacing  $n_{\text{obs}}$  by  $s + b$  (the ‘Asimov data value’).



**Figure 11.3** (a) Dimuon mass spectrum in the signal and the sideband regions. (b) The  $-2 \ln L$  function near its minimum; the estimated value of  $\alpha_{\text{bgr}}$  and its uncertainty are also indicated.

An uncertainty on the expected number of background events, like any other (systematic) uncertainty, translates into a deteriorated separation between the  $b$ -only and  $s + b$  hypotheses. In our counting experiment, the distributions of the numbers of events for both hypotheses will be formed by the Poisson distributions (like Figure 11.2b), but now convoluted with a Gaussian distribution to take into account the 10% uncertainty on the background estimate. The expected and observed significances, now computed using the estimated background estimation  $\hat{b}$ , drop to  $1.63\sigma$  and  $2.24\sigma$ , respectively.

The fact that the fit result for the background scale factor  $\alpha_{\text{bgr}}$  is very close to the expectation (i.e. one), although it has a large statistical uncertainty, opens an interesting discussion as to whether to see the use of it merely as a ‘sanity check’ and trust the results from the Monte Carlo simulation instead. Strategies for data-driven background estimates can be found in Chapter 10.

**Considerations for a more realistic example** In real-life examples, there are typically more issues to be addressed. One of them is to find a region of phase space which is populated only by events from one specific background source and which therefore allows the estimate to be translated into a number of background events in the signal region. In addition, if the sideband region is not completely free of signal events, the estimate will be biased. It is important to understand and control such biases (see Exercise 11.2). To address this last issue we will perform a simultaneous fit to the signal and the sideband regions in Section 11.2.2.

#### 11.2.1.4 Scanning over the Full Dimuon Mass Range: The ‘Look-Elsewhere Effect’

When searching for an excess from a hypothetical resonance somewhere in a mass range, the largest observed significance has to be corrected: when scanning a large mass region, one has to consider excesses from random upwards fluctuation in the number of observed events in the absence of a signal. This is the so-called *look-elsewhere effect*. The correction can be sizable and has to be taken into account when quoting observed significances. The effect is further discussed in [4] and in Section 3.5.4. Exercise 11.1 will try to give you a feeling for this issue.

### 11.2.2

#### Profile Likelihood Ratio Analysis

Using the knowledge that the background has a different shape in the dimuon mass spectrum than the signal (a falling exponential versus a resonance), we can construct a test statistic that is more powerful than counting events in a mass window. Exploiting differences in other event characteristics beyond mass information can be implemented in a similar manner and will enhance the separation power. The combination of information is discussed in a general way in Chapter 5 and in the exercises.

##### 11.2.2.1 Profile Likelihood Test Statistic

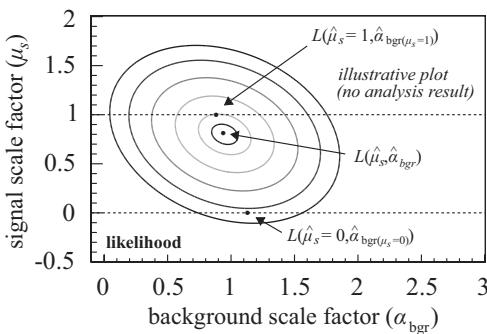
A crucial role in any search is played by the so-called *test statistic*  $t$ . This is a function of the data which characterises a full dataset in a single number. The distribution of  $t$  for a given hypothesis  $H$  is denoted as  $g(t; H)$  and can be obtained from ensemble tests. In the previous section the test statistic was simply the number of events in a mass window. For searches, it should be constructed such that the distributions  $g(t; b\text{-only})$  and  $g(t; s + b)$  are maximally separated. The test statistics can also be the output from a multivariate analysis tool like a boosted decision tree or an artificial neural network (see Chapter 5). These tools are in principle more powerful in separating different hypotheses as they take into account more event characteristics simultaneously. In this section we will use the dimuon mass information to define the test statistic as the *profile likelihood ratio* (see Section 3.2) between the two hypotheses,  $s + b$  and  $b\text{-only}$  (however, much of what follows is valid for any other test statistic as well):

$$t = -2 \ln(Q), \quad \text{with} \quad Q = \frac{L(\mu_s = 1, \hat{\alpha}_{\text{bgr}}(\mu_s=1))}{L(\mu_s = 0, \hat{\alpha}_{\text{bgr}}(\mu_s=0))}, \quad (11.4)$$

where the likelihood  $L$  is the one defined in (11.3), and the mass range in the fit spans the signal and sideband regions:  $100 < m_{\mu^+ \mu^-} < 600 \text{ GeV}$ . The quantity  $Q$  is a profile likelihood ratio which means that for each of the two values of  $\mu_s$  ( $\mu_s = 0, 1$ ), a fit is performed over the nuisance parameter  $\alpha_{\text{bgr}}$  to find the one value  $\hat{\alpha}_{\text{bgr}}(\mu_s=\mu_s)$  that maximises the likelihood. Figure 11.4 shows equi-likelihood contours in the  $\alpha_{\text{bgr}} - \mu_s$  plane and indicates the global maximum of the likelihood as well as the results for the two different constraints on the signal scale factor  $\mu_s = 0$  and  $\mu_s = 1$ .<sup>4)</sup>

Note that although the background scale factor  $\alpha_{\text{bgr}}$  is a free parameter in the fit, it is more common to include the uncertainties on nuisance parameters in the likelihood. A typical model of the uncertainty is a Gaussian constraint.

4) Many analyses at the LHC use a variation of the test statistic defined in (11.4), involving  $L(\hat{\mu}_s, \hat{\alpha}_{\text{bgr}})$ , the maximum of the likelihood when scanning over all model parameters,  $\mu_s$  and  $\alpha_{\text{bgr}}$  [5].



**Figure 11.4** Contours of equal likelihood near the maximum of the fit to a pseudo-dataset. The most likely value of  $\alpha_{\text{bgf}}$  for two different constraints on  $\mu_s$  ( $\mu_s = 0, 1$ ) are also indicated, as is the global maximum of the likelihood.

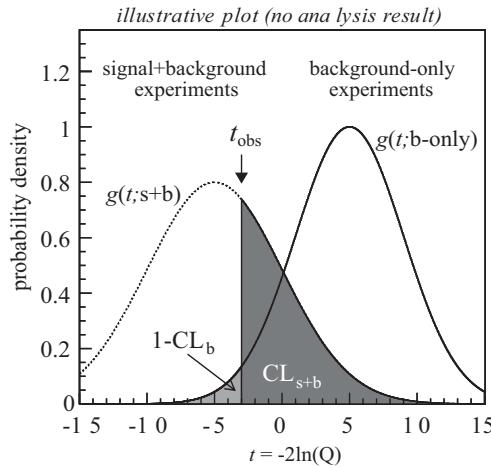
### 11.2.2.2 Properties of the Test Statistic Distributions for the $b$ -only and $s + b$ Hypotheses

The pseudo-datasets used to generate the distributions of the test statistic (discussed below) for the  $b$ -only and  $s + b$  hypotheses were obtained by drawing events from the dimuon mass templates shown in Figure 11.1. The number of events in each pseudo-dataset is drawn from a Poisson distribution. The expectation value of the Poisson distribution is determined separately for each pseudo-data sample by considering all uncertainties – in our case the uncertainty on the background normalisation. More complex (correlated) systematic uncertainties can also be considered at this stage. Generating pseudo-datasets is an important ingredient in the evaluation of the performance of any analysis. It is discussed in detail in Section 10.5, and a practical example related to our specific problem is part of Exercise 11.2.

**Confidence levels** Evaluating the compatibility of a dataset with a specific hypothesis means calculating the probability of observing the obtained test statistic value or a more extreme one, assuming the hypothesis to be true. Figure 11.5 shows the test statistic distribution for the  $b$ -only and  $s + b$  hypotheses generated using large numbers of pseudo-datasets. It also shows the test statistic  $t_{\text{obs}}$  that is observed in the data. The compatibility of  $t_{\text{obs}}$  with the two hypotheses is quantified by two *confidence levels* –  $1 - \text{CL}_b$  and  $\text{CL}_{s+b}$  – which are defined as follows:

- $1 - \text{CL}_b = \int_{-\infty}^{t_{\text{obs}}} g(t; b\text{-only}) dt$  (background  $p$ -value): the probability for the test statistic  $t$  to be as small as the observed test statistic  $t_{\text{obs}}$ , or even smaller (more signal-like), under the  $b$ -only hypothesis. Note that this is the  $p$ -value defined in Section 11.2.1.1.<sup>5)</sup>

5) This is true only for the case that the median of the  $t$  distribution for the  $s + b$  hypothesis is smaller than that for the  $b$ -only distribution. Therefore, in the example discussed in Section 11.2.1, the limits of integration are  $t_{\text{obs}} = n_{\text{obs}}$  and  $+\infty$ .



**Figure 11.5** An illustrative example of the test statistic distribution for the  $b$ -only (solid line) and  $s + b$  (dotted line) hypotheses, and a representation of the confidence levels used to quantify the compatibility of the observation,  $t_{\text{obs}}$ , with either hypothesis.

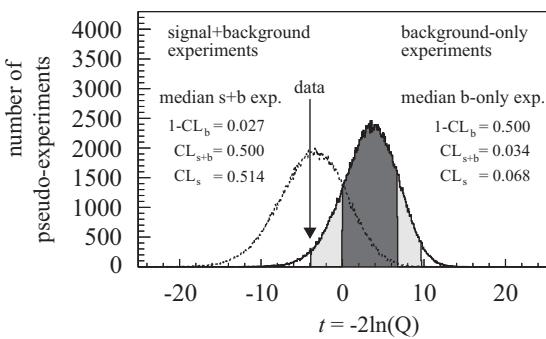
- $\text{CL}_{s+b} = \int_{t_{\text{obs}}}^{+\infty} g(t; s + b) dt$ : the probability for the test statistic  $t$  to be as large as  $t_{\text{obs}}$  or larger (more background-like), under the  $s + b$  hypothesis.

**Expected significance of the signal** By studying the distribution of a large number of pseudo-experiments under the two different hypotheses ( $b$ -only and  $s + b$ ) for our analysis at 250 GeV (Figure 11.6), we can now define the expected confidence levels. As before, these are defined as the confidence levels associated with the median value of the test statistic for each of the two hypotheses.

The expected significance of the 250-GeV  $Z'$  signal, for example, is computed from the  $p$ -value (or  $1 - \text{CL}_b$ ) of the median  $s + b$  experiment. At the test statistic value of the median  $s + b$  experiment ( $t = -3.57$ ) the value of  $1 - \text{CL}_b$  is 0.027, which translates into an expected significance of  $1.93\sigma$ . Similarly, the median  $b$ -only experiment has a test statistic value of  $t = 3.49$  which gives an expected confidence level in the signal+background ( $\text{CL}_{s+b}$ ) of 0.034. This means that, in case there is no signal, the typical experiment will be quite incompatible with the  $s + b$  hypothesis. That is: only 3.4% of the  $s + b$  experiments produce such a value of the test statistic or a more background-like one. A few of the characteristic numbers for the median  $b$ -only and  $s + b$  hypotheses are shown in Figure 11.6 and listed in Table 11.1.

#### 11.2.2.3 Rules for Discovery and Exclusion

Before calculating the test statistic  $t_{\text{obs}}$  from the data, we need to define a set of ‘rules’ that help us to decide whether – based on the confidence levels associated with our measurement – to accept or reject a given hypothesis.



**Figure 11.6** The distribution of the test statistic for a large number of  $b$ -only (solid line) and  $s + b$  (dotted line) pseudo-datasets from our example analysis. For the  $b$ -only hypothesis, the  $1\sigma$  and  $2\sigma$  regions are represented by the

dark and light shaded areas, respectively. The test statistic observed in the data, which is close to the expectation for the median value from  $s + b$  experiments, is also indicated.

**Table 11.1** Characteristic numbers (confidence levels and  $CL_s$  ratio) for the median  $b$ -only experiment, the median  $s + b$  experiment, and the data.

Confidence level	Median $b$ -only experiments	Median $s + b$ experiments	Data
$1 - CL_b$	0.500	0.027	0.021
$CL_{s+b}$	0.034	0.500	0.542
$CL_s$	0.068	0.514	0.554

**Discovery** To claim the discovery of a signal, the  $b$ -only hypothesis (our precious SM) has to be rejected. As we like to be quite certain that the effect we observe can not be explained by the SM, it is customary in high energy physics to only claim a discovery if the  $p$ -value is smaller than  $5.73 \cdot 10^{-7}$ , the famous ‘ $5\sigma$  effect’.<sup>6)</sup> Note that one should also account for the look-elsewhere effect described in Section 11.2.1.4.

**Exclusion** A signal hypothesis can be excluded if the compatibility with the  $s + b$  hypothesis is ‘small’. Several limit-setting methods exist and are a source of intense discussions among physicists, to say the least (see also Section 4.6). Although it might seem natural to define a signal as excluded at 95% confidence level if  $CL_{s+b} < 5\%$ , there are some undesirable consequences associated with this choice. Near the sensitivity limit, where the test statistic distribution for the  $b$ -only and  $s + b$  hypotheses are not well separated (either because the signal is small or the analysis is not powerful enough to separate signal and background), a downward fluctuation in the data with respect to the  $b$ -only expectation will result in the exclusion of a signal while the analysis has no real sensitivity. One of the common solutions

6) Note that this is a double-sided definition of the  $p$ -value – a convention that is often, but not always, used in searches.

experiments invoke to address this issue is to correct for a downward fluctuation by using the so-called  $\text{CL}_s$  method [6], where a signal is called ‘excluded’ at 95% confidence level if  $\text{CL}_s < 0.05$ , where  $\text{CL}_s \equiv \text{CL}_{s+b}/\text{CL}_b$ . Note that since  $\text{CL}_s$  is not a classic confidence level but only a ratio of confidence levels, the interpretation of the result and the comparison to other search results should be done with caution. Another way is to use so-called *power-constrained limits* (PCL), where the signal is indeed excluded using  $\text{CL}_{s+b} < 5\%$ , but where the limit is kept at the more conservative expected limit if there is a large downward fluctuation.

#### 11.2.2.4 Results from Data

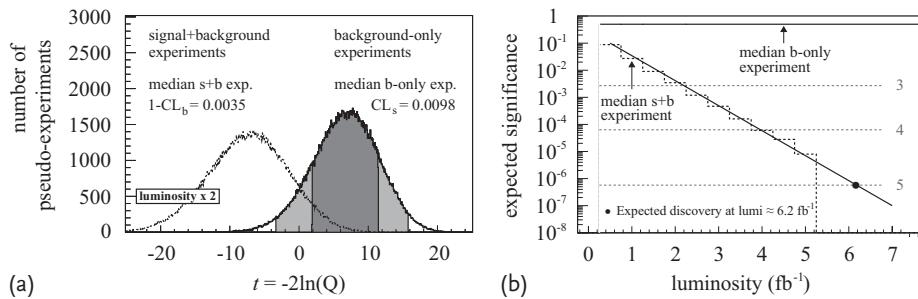
The distribution of the test statistic for  $b$ -only (solid line) and  $s + b$  experiments (dotted line) and the results obtained in the data are shown in Figure 11.6. The dark and light regions indicate the  $1\sigma$  and  $2\sigma$  fluctuations on the  $b$ -only hypothesis (containing 68% and 95% of the pseudo-datasets, respectively). The measurement will only result in one value of the test statistic,  $t_{\text{obs}}$ : in our case  $t = -3.99$ . The confidence levels for that value of the test statistic are listed in Table 11.1 together with those derived from  $b$ -only and  $s + b$  pseudo-experiments.

Although we observe an excess in the data ( $1 - \text{CL}_b = 0.021$ , corresponding to  $2.03\sigma$ ), the excess is not significant enough to reject the  $b$ -only hypothesis. Disappointing as that might be, we also did not really expect to be able to claim a discovery with this amount of integrated luminosity because the expected significance  $1 - \text{CL}_b$  is only 0.027. From Figure 11.6 we see that the value observed in the data is consistent with the presence of a signal, and since  $\text{CL}_s > 0.05$  in the data we cannot exclude the signal hypothesis. Like for the discovery, with this luminosity and this signal cross section we did not expect to be able to exclude a signal since the expected  $\text{CL}_s$  for the  $b$ -only hypothesis is 0.068. The excess observed in the data is compatible with the expectation from a 250 GeV  $Z'$ , and it would therefore be interesting to see whether a larger dataset rules out the signal or leads to its discovery.

#### 11.2.2.5 Probing the Sensitivity Limits: Enhanced Luminosity and Signal Cross Sections

The search for a 250 GeV  $Z'$  in our dataset of  $1\text{ fb}^{-1}$  resulted in neither a discovery nor an exclusion. We can thus try to see what the sensitivity limits of our analysis are by investigating with which luminosity we expect to be able to make a discovery and what  $Z'$  signal cross section we *can* exclude.

**Enhanced luminosity** As the luminosity increases beyond  $1\text{ fb}^{-1}$ , the difference between the test statistic distributions for the two hypotheses becomes more pronounced. Figure 11.7a shows these distributions with twice the luminosity as before. By artificially increasing the luminosity in our pseudo-datasets, we can study the evolution of the expected significance of the  $Z'$  signal as a function of the luminosity. This distribution, shown in Figure 11.7b, reveals that we only expect to be able to claim a discovery in a dataset roughly six times the size of the current one.



**Figure 11.7** (a) Distribution of the test statistic under the  $b$ -only (solid line) and  $s + b$  (dotted line) hypotheses, where the luminosity is scaled by a factor 2 with respect to the previous example. The dark and light shad-

ed regions represent the  $1\sigma$  and  $2\sigma$  regions under the  $b$ -only hypothesis. (b) Expected significance as a function of the integrated luminosity.

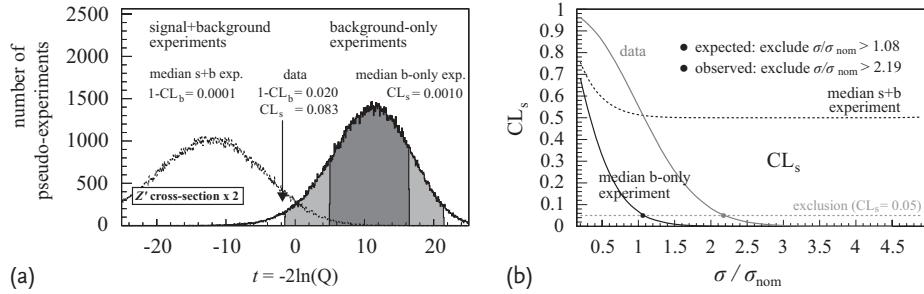
**Enhanced signal cross section** Although with this dataset we could not exclude a  $Z'$  signal with its nominal cross section ( $\sigma_{\text{nom}}$ ), we can determine what cross section we *can* exclude. Such an exercise can be useful if the cross section of the signal contains a free parameter. It also provides an alternative way to summarise the sensitivity and result of the experiment.<sup>7)</sup> If the predicted signal cross section  $\sigma$  increases, it is easier to separate the  $b$ -only and  $s + b$  hypotheses, as can be seen from Figure 11.8a. The figure shows the distribution of the test statistic under the  $b$ -only (solid line) and  $s + b$  (dotted line) hypotheses, where the signal cross section was increased by a factor 2. By further increasing the predicted signal cross section, we can study the evolution of the  $\text{CL}_s$  ratio for the signal versus  $\sigma/\sigma_{\text{nom}}$  and determine at what scale factor the  $\text{CL}_s$  ratio drops below 0.05 for the median  $b$ -only experiment and the data. From Figure 11.8b we see that the cross section *expected* to be excluded is 1.08 times that of the nominal one. However, because we see an excess of events in the data, we can only exclude a signal cross section that is a factor 2.19 or more higher than  $\sigma_{\text{nom}}$ .

#### 11.2.2.6 Scanning the Full Mass Region

As described in the previous section, in our dataset of  $1 \text{ fb}^{-1}$  the expected and observed cross sections relative to the nominal one that can be excluded are 1.08 and 2.19, respectively, assuming  $M_{Z'} = 250 \text{ GeV}$ . However, as the mass of the  $Z'$  is an unknown parameter, our example analysis that aims at searching for a  $Z'$  with a mass of 250 GeV should be repeated over a wide range of hypothetical  $Z'$  masses. To avoid missing the signal, the mass steps in such a scan should be separated by not more than the mass resolution.

Optimistic as we are, we will first cover the discovery potential and then discuss the potential for excluding  $Z'$  masses and cross sections.

7) This approach is for example used to summarise the results of Higgs searches performed by the various experiments at LEP, Tevatron and the LHC.

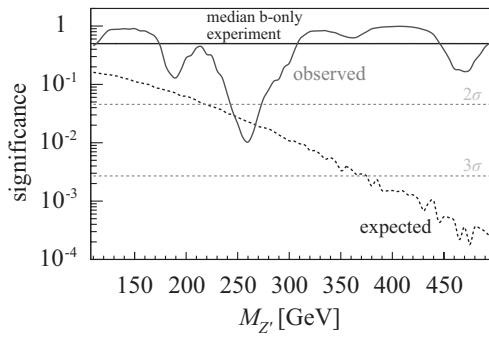


**Figure 11.8** (a) Distribution of the test statistic for the  $b$ -only (solid line) and  $s + b$  (dotted line) hypotheses, where the signal cross section is scaled by a factor 2. The dark and light bands represent  $1\sigma$  and  $2\sigma$  regions for the  $b$ -only hypothesis. Note that also the pseudo-experiments with  $-2 \ln(Q) < -25$  (not shown

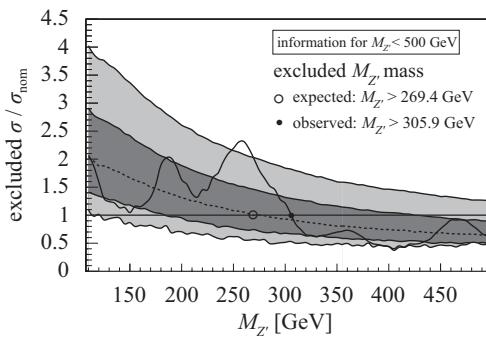
in the plot) are taken into account when computing confidence levels. (b) The  $CL_s$  ratio for the median  $b$ -only experiment, the median  $s + b$  experiment and the data as a function of the ratio of the signal cross section over the nominal one,  $\sigma/\sigma_{\text{nom}}$ .

**Discovery** Figure 11.9 shows the expected significance as a function of the  $Z'$  mass hypothesis. No sufficiently large excess was found in the data to claim a discovery ( $1 - CL_b < 5.73 \cdot 10^{-7}$ ) for any of the masses. The most significant excess is observed at a mass of 260 GeV and is even larger than that expected from a 260 GeV  $Z'$ .

**Exclusion** Similar to what was done for  $M_{Z'} = 250$  GeV, by artificially increasing the cross section of the  $Z'$  signal relative to the nominal one, the observed (expected) result of the search can be presented as an observed (expected) cross-section exclusion. In addition to the expected exclusion from the *median b-only* experiment, we can also evaluate the value of  $CL_s$  for  $\pm 1\sigma$  or  $\pm 2\sigma$  fluctuations of the background (the borders of the dark and light regions in Figure 11.6). The values of the signal cross-section scale factors for which these characteristic points cross



**Figure 11.9** Distribution of the expected and observed significances as a function of the  $Z'$  mass. The local  $p$ -values shown here have not been corrected for the look-elsewhere effect.



**Figure 11.10** Excluded cross-section scaling factors versus the  $Z'$  mass. The dotted (solid) line represents the expected (observed) excluded cross section, and the dark and light bands represent the  $\pm 1\sigma$  and  $\pm 2\sigma$  fluctuations of the background.

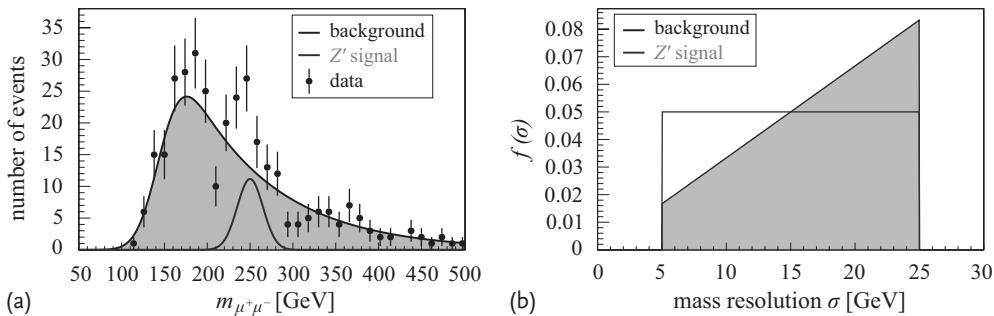
$CL_s = 0.05$  define the boundaries of the dark and light regions in Figure 11.10. The summary of excluded scale factors as a function of  $M_{Z'}$  is shown in Figure 11.10.

The mass region that we expect to be able to exclude with the available luminosity of  $1 \text{ fb}^{-1}$  is defined as the region where the median  $b$ -only experiments lead to an excluded cross-section scaling factor ( $\sigma/\sigma_{\text{nom}}$ ) smaller than unity. In Figure 11.10 we see that this is the case for  $Z'$  masses larger than 269.4 GeV (open bullet). Because of the excess in data, however, the excluded mass region is significantly smaller (filled bullet). The summary of our measurement will be: ‘If a  $Z'$  exists and has a cross section as predicted, then  $Z'$  masses between 305.9 and 500 GeV are excluded at 95% confidence level.’

## 11.3 Measurement

### 11.3.1 Introduction

Once we have established the presence of a signal, we can turn to the determination of its properties. There are many common points between searches and measurements; in both cases the likelihood plays a crucial role. In what follows, we will use an unbinned likelihood, which allows more information than only the invariant mass of the dimuon pairs to be easily used. In particular, our events are not only characterised by the measured dimuon mass, but also by the experimentally determined uncertainty on this number, which can be significantly different from event to event. We will make use of this information in order to improve the precision of the measurement. The data used here are different from those used in the previous section also in another respect: we assume that the detector acceptance is a function of the event mass. This typically is a consequence of the trigger criteria used to collect the events. As a result there are no events at very low masses in



**Figure 11.11** (a) Invariant dimuon mass distribution in the data sample, together with the pdfs for the signal and background that are used for the generation of the sample and for

the likelihood computation. (b) Distributions  $f^{\text{sig}}$ ,  $f^{\text{bgr}}$  of the mass resolution  $\sigma$  for signal and background.

our dataset. Again, we assume the signal and background shapes to be known perfectly, but not the normalisations. The distribution of the dimuon mass is shown in Figure 11.11a along with the signal and background probability density functions (pdfs),  $h^{\text{sig}}$  and  $h^{\text{bgr}}$ , that were used to generate it. The generated data contain 35 signal events in addition to 300 background events.<sup>8)</sup> The  $Z'$  mass is assumed to be 250 GeV, and we assume that the experimental width is fully due to the resolution of the mass measurement; that is the natural width of the  $Z'$  is negligible. This uncertainty on the mass measurement is known,<sup>9)</sup> and we assume that it correctly describes the true event-by-event resolution. In real life, this can be checked using a known (preferably narrow) resonance. For the signal events, the mass is distributed following a Gaussian with standard deviation  $\sigma$ , which is different for each event. Because of differences in the event topology and track quality, the distributions of  $\sigma$  for signal and background events,  $f^{\text{sig}}$  and  $f^{\text{bgr}}$ , can be different. The two assumed distributions are shown in Figure 11.11b.

In the following sections, we will set up a measurement of both the  $Z'$  mass and the number of signal events,  $\nu^{\text{sig}}$ . Assuming that the branching fraction, the acceptance of the detector and the selection efficiency are known from simulations, the measurement of  $\nu^{\text{sig}}$  will allow a direct determination of the cross section.

### 11.3.2 Unbinned Likelihood

The workhorse of the analysis is again the likelihood, which quantifies – as a function of the parameters of interest to us – the probability of the observed dataset to occur. We use an *extended unbinned* likelihood (see Section 2.5.2), meaning that the logarithm of the likelihood is computed by a sum over the events themselves,

- 8) The relevant information for the property determinations are available in the file `DataSample_measurement.root`.
- 9) The mass resolution of the event is typically computed from the covariances of the track momenta using error propagation.

rather than over the bins of some histogram of their properties:

$$\begin{aligned} \ln L(\text{events}; \nu^{\text{sig}}, \nu^{\text{bgr}}, M_{Z'}) \\ = -(\nu^{\text{sig}} + \nu^{\text{bgr}}) + \sum_i \ln \left[ \nu^{\text{sig}} \cdot p_i^{\text{sig}} + \nu^{\text{bgr}} \cdot p_i^{\text{bgr}} \right], \end{aligned} \quad (11.5)$$

where  $\nu^{\text{sig}}$  and  $\nu^{\text{bgr}}$  denote the expected numbers of signal and background events, respectively;  $\nu^{\text{sig}}$  is directly related to the production cross section. Formally,  $p_i^{\text{sig}}$  and  $p_i^{\text{bgr}}$  denote multi-dimensional probability density functions of the observables that we want to use in the analysis, evaluated at the event characteristics of event  $i$ . In our dataset, the events are characterised by a reconstructed dimuon mass  $m_i$  and its resolution  $\sigma_i$ . In our case, the likelihood has three parameters: the signal and background normalisations,  $\nu^{\text{sig}}$  and  $\nu^{\text{bgr}}$ , and the  $Z'$  mass  $M_{Z'}$ .

As we are interested in a precise measurement of the  $Z'$  mass, we want to take advantage of the known uncertainty on the dimuon invariant mass for each event. This information can be used in the likelihood, where events with a better mass resolution pose stronger constraints on the parameters. To achieve this, the width of the signal pdf for the reconstructed dimuon mass is assumed to be a function of the experimentally determined mass resolution  $\sigma_i$  for each event  $i$  in the dataset:

$$p_i^{\text{sig}} = p^{\text{sig}}(m_i, \sigma_i; M_{Z'}) = h^{\text{sig}}(m_i; M_{Z'}, \sigma_i) \cdot f^{\text{sig}}(\sigma_i) \quad (11.6)$$

and

$$p_i^{\text{bgr}} = p^{\text{bgr}}(m_i, \sigma_i) = h^{\text{bgr}}(m_i; \sigma_i) \cdot f^{\text{bgr}}(\sigma_i). \quad (11.7)$$

It is noteworthy that the terms  $f^{\text{sig}}(\sigma_i)$  and  $f^{\text{bgr}}(\sigma_i)$  can not be omitted in general. Both  $m_i$  and  $\sigma_i$  are random variables, and the likelihood should reflect this by using the full pdf. The  $\sigma_i$  are distributed differently for signal and background events, and neglecting these terms can lead to biases in the estimated parameters. The information not only helps in the mass measurement, but also allows a more precise cross-section measurement to be performed.

### 11.3.2.1 Likelihood Ingredients

A major step in implementing the likelihood function is obtaining the various probability density functions. In our simple example, we have the luxury of using the same simple functions that we have assumed to describe reality: a Gaussian signal on top of an exponential background. While using a set of simple functions is often a reasonable choice, it may also happen that the signal mass distribution is complicated and needs to be modelled by an MC simulation: samples are generated for a limited number of  $Z'$  masses, which are then interpolated to provide the likelihood for all values of  $M_{Z'}$ . This procedure is known as *template morphing*, and there are various algorithms around for it (see e.g. [7]).

## 11.3.3

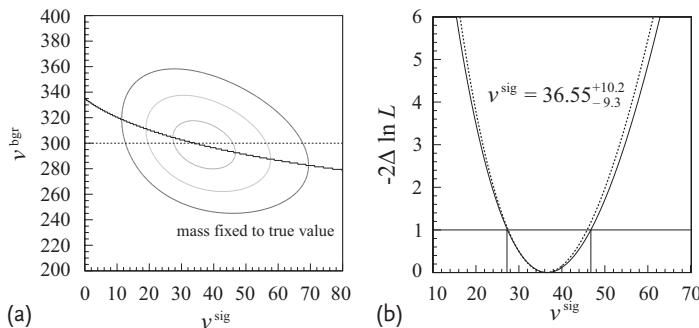
## Extracting a Measurement in the Presence of Nuisance Parameters

There are three parameters in the likelihood:  $\nu^{\text{sig}}$ ,  $\nu^{\text{bgr}}$  and  $M_{Z'}$ . We are not very interested in the number of background events,  $\nu^{\text{bgr}}$ , which is a so-called *nuisance parameter*. Nuisance parameters can be dealt with using a *profile likelihood* (see Chapter 2).

To illustrate likelihood profiling, we first estimate  $\nu^{\text{sig}}$  from the data for the case where  $M_{Z'}$  is fixed in the likelihood to its true value. Figure 11.12a then shows the likelihood contours in the  $\nu^{\text{bgr}}-\nu^{\text{sig}}$  plane. The two-dimensional likelihood function is reduced to a one-dimensional likelihood for  $\nu^{\text{sig}}$ . This is achieved by taking, for every value of  $\nu^{\text{sig}}$ , the best (i.e. maximum) likelihood that we can find as a function of  $\nu^{\text{bgr}}$ . As Figure 11.12a shows contours in  $-2 \ln L$ , this means we find the minimum along each line of constant  $\nu^{\text{sig}}$ , which then corresponds to the maximum-likelihood estimate of  $\nu^{\text{bgr}}$ . The points found in this way are indicated by the solid line in the figure. The result is the profile likelihood,  $-2 \ln L^{\text{prof}}$ , as a function of  $\nu^{\text{sig}}$ , which is shown as a solid line in Figure 11.12b.

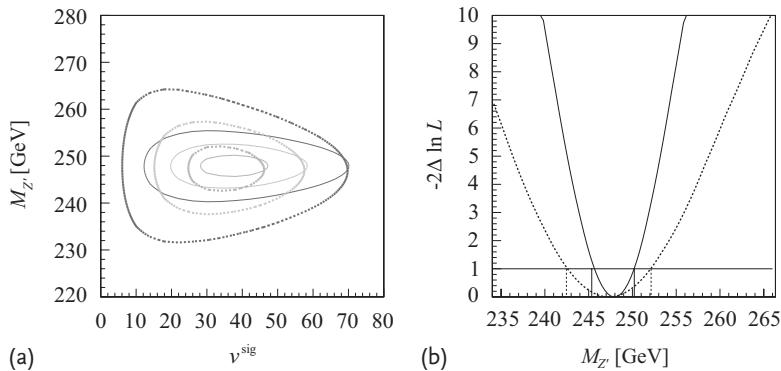
From the profile likelihood curve, we extract the maximum-likelihood estimate for  $\nu^{\text{sig}}$  by finding the minimum value of  $-2 \ln L^{\text{prof}}$ . The result in our dataset is  $\nu^{\text{sig}} = 36.6^{+10.2}_{-9.3}$ . The  $1\sigma$  confidence interval corresponds to the range of values for which  $-2 \ln L^{\text{prof}}$  differs from its minimum value by less than 1.

We can compare the profile likelihood result to the unrealistic case where we have perfect knowledge of  $\nu^{\text{bgr}}$ . This corresponds to the dashed lines in Figure 11.12. The resulting likelihood function is only marginally narrower than the profile likelihood discussed above, which indicates that the uncertainty on the background normalisation is not a severely limiting factor in this particular measurement.



**Figure 11.12** (a) Contour plot of  $-2 \ln L$  as a function of the numbers of signal and background events,  $\nu^{\text{sig}}$  and  $\nu^{\text{bgr}}$ . The solid contours indicate the  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  confidence intervals. The dashed line shows the true number of background events. (b) Profile likelihood plot for  $\nu^{\text{sig}}$ . The dashed line shows the likelihood for  $\nu^{\text{bgr}} = 300$ .

dence intervals. The dashed line shows the true number of background events. (b) Profile likelihood plot for  $\nu^{\text{sig}}$ . The dashed line shows the likelihood for  $\nu^{\text{bgr}} = 300$ .



**Figure 11.13** (a) Contours of the profile likelihood as a function of the signal normalisation  $\nu^{\text{sig}}$  and the  $Z'$  mass  $M_{Z'}$ . The contours indicate the  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  confidence intervals obtained using the average mass resolution

(dashed) and the event-by-event resolution (solid). (b) Profile likelihood for  $M_{Z'}$  using the average mass resolution (dashed line) and the event-by-event resolution (solid line).

### 11.3.4

#### Mass Measurement

Having introduced the concept of the profile likelihood, we now turn to measuring the  $Z'$  mass. As before, the profile likelihood is found by evaluating the likelihood function at the maximum-likelihood estimate of  $\nu^{\text{bgr}}$  for each value of  $\nu^{\text{sig}}$  and the  $M_{Z'}$ . The resulting profile likelihood contours are shown in Figure 11.13a. The dashed lines in the figure correspond to a likelihood function that does not use the event-by-event mass resolution information, while the solid lines correspond to (11.6) and (11.7). Figure 11.13b shows the profile likelihood for  $M_{Z'}$ . It is clear from the figure that using the event-by-event mass resolution has significantly improved the measurement of  $M_{Z'}$ . We measure a value of  $M_{Z'} = 247.8 \pm 2.5 \text{ GeV}$ . Since the likelihood function is reasonably symmetric, we have opted to quote just a single value for the uncertainty here. The measured value deviates from the true one by about  $1\sigma$ . The uncertainty would be about a factor of two larger if we had not used the event-by-event mass resolution – in this case the result is  $M_{Z'} = 247.2 \pm 4.8 \text{ GeV}$ .

### 11.3.5

#### Testing for Bias and Coverage

The method and machinery for making a measurement should be tested to ensure that it yields, on average, the input values and appropriate uncertainties. This is all the more important if the method is complicated. As detailed in Chapter 10, such tests can be done by running the full method on pseudo-experiments generated using either simplified or fully fledged Monte Carlo simulations. In each of the pseudo-experiments, the measured mass and the associated confidence interval

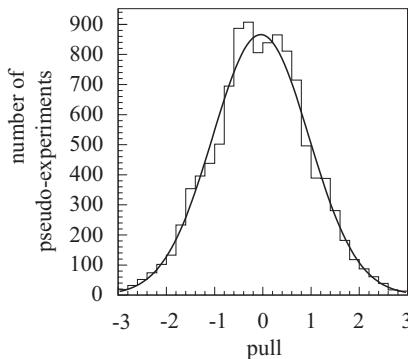


Figure 11.14 The pull distribution of our  $Z'$  mass measurement.

can be compared with the true mass that was the input for the simulation.

In this way, it can be checked if the method is unbiased, which for our example means that the distribution of the difference between measured and true  $Z'$  mass is centred around zero. One should also test whether the confidence interval contains the true value of the parameter in the required number of experiments. That is: in 68% of the experiments, the  $1\sigma$  confidence interval should contain the known, true value. This is called a *coverage check*.

If the negative of the logarithm of the likelihood function is quadratic near the minimum, the tests for bias and coverage are often performed simultaneously by studying the distribution of the so-called *pull*, which is defined as the difference between measured and true value, divided by the estimated uncertainty. In case of asymmetric uncertainties, one can take the upper or lower uncertainty depending on the sign of the pull. The pull distribution for our example analysis is shown in Figure 11.14. An unbiased pull distribution is centred around zero; a reliable error estimate is indicated by a Gaussian-shaped pull distribution with a unit width.

A bias in the pull distribution may point to some shortcoming in the method. One reason could be that the pdfs used in the likelihood are not a perfect description of the Monte Carlo simulations – for example because certain simplifications are made. It is not uncommon to simply correct the method for the bias by subtracting it from the central value of the result. The same is true for the uncertainties provided by the method. If the pull is too wide by a small fraction, it can be acceptable to scale the uncertainties by this factor. However, one should be careful not to mask true mistakes in the method in this way. A valuable sanity check of the likelihood is to generate the pseudo-experiments according to exactly the same pdfs that are used in the parameter-estimation process.

### 11.3.6

#### Systematic Uncertainties

Every measurement suffers from the imperfect knowledge of input quantities. One should therefore study the related effects on the measurement and evaluate the re-

sulting *systematic uncertainties*. Systematic uncertainties are discussed at length in Chapters 9 and 8. In the specific case of our  $Z'$ -mass measurement, one has to consider – among other things – if there is a systematic uncertainty on the reconstructed dimuon mass. One example is an uncertainty on the mass scale which would immediately impact the main result. Assuming our experiment had an uncertainty on the dimuon mass scale of 0.5%, we could quote the main result as  $M_{Z'} = 247.8 \pm 2.5(\text{stat}) \pm 1.2(\text{syst}) \text{ GeV}$ .

In most cases, the impact of the uncertainty on the final result is not so easy to evaluate. In a mass measurement, one might for example need to propagate the uncertainties on the detector alignment to the final result. In such cases, the systematic uncertainty on the result is usually determined by generating multiple (often just two) Monte Carlo datasets in which the uncertain input parameters have been varied (by their  $1\sigma$  uncertainties). The effect of the uncertainty can be evaluated by performing the measurement on the simulated samples and observing the variation in the output.

In case the analysis uses Monte Carlo simulations as direct input, for example to compute the detector acceptance in a cross-section measurement, one can repeat the measurement using the different inputs derived from the varied simulation.

If there is good reason to assume that two systematic uncertainties are uncorrelated, it is justifiable to add them in quadrature. Otherwise, they should be varied consistently when generating the Monte Carlo datasets.

In some cases, the variations in the input lack a good probabilistic meaning. It is, for example, quite common to quote the variations observed when using two different Monte Carlo programs as a systematic uncertainty. While this may not be very rigorous, it can nevertheless be useful to know if such variations have a large or small impact on the result.

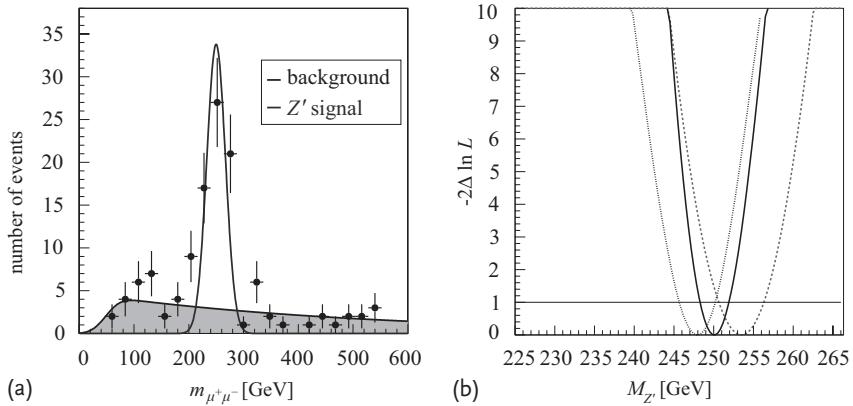
### 11.3.7

#### **Constraints and Combining Measurements**

Combining different measurements is, in principle, a simple matter of adding the logarithms of the likelihood functions. To illustrate this, we suppose we also measure the  $Z'$  particle in a dielectron sample. The measured dielectron mass distribution is shown in Figure 11.15a.<sup>10)</sup> The profile likelihood for  $M_{Z'}$  in this sample is derived in an analogous way to what was discussed before for the muon channel. The result is the dashed curve in Figure 11.15b. In this sample alone we measure  $M_{Z'} = 253.5 \pm 3.0 \text{ GeV}$ .

The combination proceeds by adding the logarithms of the profile likelihood curves; the resulting combined likelihood is shown as a solid curve in Figure 11.15b. From this combined profile likelihood, we obtain  $M_{Z'} = 250.2^{+2.1}_{-1.7} \text{ GeV}$ . Hence, by adding the electron sample, the uncertainty has been reduced by  $\sim 20\%$ .

<sup>10)</sup> It is not very realistic that the electron sample is so much cleaner than the muon sample, but it will suffice as an example.



**Figure 11.15** (a) Invariant dielectron mass distribution of the electron sample. (b) Profile likelihood curves for the measurement in the muon sample (dotted curve), the electron sample (dashed curve) and the combined likelihood (solid curve).

Note that we have implicitly assumed that the  $\nu^{\text{sig}}$  in the two samples are independent: each of them was ‘profiled away’ independently. This can be done if we assume, for example, that the branching fractions of the  $Z'$  are unknown. If the branching fractions are known, however, the two  $\nu^{\text{sig}}$  parameters are fully correlated and one should add the two-dimensional likelihood distributions of  $\alpha \nu^{\text{sig}}$  (where  $\alpha$  is the acceptance which is different for each sample) versus  $M_{Z'}$  and extract the combined measurement from the sum of the distributions.

If there are many common parameters in the two likelihoods, it might not be practical to produce such likelihood distributions for combining the results as they would become multi-dimensional. In such a case, it is good to realise that one can also combine the event-by-event likelihoods from (11.5) directly and simply loop over the events in both samples each time the likelihood is computed. This then allows for as many common parameters as required in both data samples.

The subject of combining measurements is closely related to the notion of *constraints* in the likelihood. Constraints are discussed extensively in Chapter 7. In the example above one could regard one measurement in the electron sample as a constraint on the result in the muon sample. There is no fundamental difference between *combining* measurements and *constraining* one measurement with the other, or vice versa, and both can be implemented in the same way. However, the term ‘constraints’ is usually used when one adds the additional information directly to the likelihood formula rather than *a posteriori* as we have done here. For example, the logarithm of the likelihood in the electron sample can be described by a parabola:

$$\ln L^{(e)} = -\frac{1}{2} \cdot \left( \frac{M_{Z'} - 253.5}{3.0} \right)^2. \quad (11.8)$$

This term can simply be added to (11.5) in order to constrain the measurement in the muon channel. If the likelihood curve is not available, it is often assumed to be Gaussian in shape.

## 11.4 Exercises

### Exercise 11.1 Counting significance and simple mass windows

In the counting experiment from Section 11.2.1 we assumed that the background  $b$  was perfectly known, that is, its uncertainty was  $\Delta b = 0$ , and we tried to optimise the expected significance. For a 250 GeV  $Z'$  and using a symmetric mass window:

- Find the optimal window and list the corresponding  $p$ -value and significance.
- Find the optimal window for a five times higher luminosity.
- Find the window that optimises the *observed* significance.
- Plot the expected significance as a function of the luminosity. At what luminosity do you expect to make a  $5\sigma$  discovery? What if  $\Delta b = 10\%$ ?
- If you could also change the mass window, what is the first luminosity at which you expect to be able to claim a discovery?
- Scan over all mass windows and masses to find the most significant excess. How significant is the excess you found in question (a) if you take into account the look-elsewhere effect?

### Exercise 11.2 Monte Carlo experiments and sideband fit

In Section 11.2.1.3 we performed a sideband fit to estimate the number of background events in the signal region.

- Perform the sideband fit to the data and estimate the number of background events in the signal region and its uncertainty.
- Generate pseudo-datasets for the  $b$ -only and  $s + b$  hypotheses.
- Plot the pull distribution to check if your estimate of the number of background events in the signal region is unbiased and that the uncertainty estimate is correct. Pull =  $(N_{\text{fit}}^{\text{predicted}} - N_{\text{MC}}^{\text{truth}})/\sigma_{N_{\text{fit}}^{\text{predicted}}}$ . What is the mean and standard deviation of this distribution? What does that tell you?

### Exercise 11.3 Pseudo-datasets and test statistic distributions

In Section 11.2.2.1 we described the test statistic  $t$  and rules for limit setting.

- Check the influence of an uncertainty on the luminosity of 10% on the sensitivity by comparing  $1 - \text{CL}_b$  (or an expected  $\sigma/\sigma_{\text{SM}}$  exclusion) for the median  $s + b$  experiment.

- b) Calculate a limit by requiring  $CL_{s+b} < 0.05$  rather than  $CL_s < 0.05$ . Is it more strict than  $CL_s$ ? What are the differences?
- c) Implement the LHC test statistic mentioned in Section 11.2.2.1. Study its effect by comparing  $1 - CL_b$  for the median  $s + b$  hypothesis or an expected exclusion on  $\sigma/\sigma_{\text{SM}}$ .

#### Exercise 11.4 Poisson errors on data points

The subtleties of uncertainty regions enter at many stages of an analysis: although it is customary to assign an uncertainty of  $\sqrt{n}$  to an event count of  $n$  events, this is not the most natural way to summarise the measurement. While  $\sqrt{\nu_b}$  is a measure for the expected spread of the number of observed events from a Poisson process with a well-known mean  $\nu_b$ , the uncertainty on an observed event yield is expected to reflect information on what we infer on the expected number of events. Although there are many ways to define a region summarising the measurement, the uncertainty interval assigned to data points in Figure 11.1 (and the default in `ROOFIT`), is the region  $(\mu_{\text{low}}, \mu_{\text{up}})$  defined by

$$\begin{aligned}\nu_{\text{up}} : \quad &\text{smallest } \nu \quad \text{for which} \quad P(n \geq n_{\text{obs}}|\nu) = 0.159 , \\ \nu_{\text{low}} : \quad &\text{largest } \nu \quad \text{for which} \quad P(n \leq n_{\text{obs}}|\nu) = 0.159 .\end{aligned}$$

Other constructions, each with their specific properties, use the likelihood directly or integrate the posterior probability density function for  $\nu$  (Bayesian). To get a feel for the different choices we compute several confidence-level interval regions for  $n_{\text{obs}} = 3$  as presented in Table 1 of [8]. Note that the various choices have also been coded in `PoissonError.C`.

- a) Classical central:  $(\mu_{\text{low}}, \mu_{\text{up}}) = (1.37, 5.92)$ .
- b) Likelihood ratio ( $\Delta \ln(L) = 1$ ):  $(\mu_{\text{low}}, \mu_{\text{up}}) = (1.58, 5.08)$ .
- c) Bayesian: use the likelihood and a flat prior in  $\mu$  to construct the posterior pdf for  $\mu$  and construct a confidence region by integrating the pdf, either with symmetric error regions ‘Bayesian central’ (2.09, 5.92) or boundaries that have equal probability (1.55, 5.14).
- d) Irritate and confuse people at your institute by discussing this over coffee.

As a final remark on this exercise we note that a similar and regularly returning discussion is on the estimate of the uncertainty on a selection efficiency [9].

#### Exercise 11.5 Likelihood for a measurement

- a) Show that a change of units used for the probability densities occurring in (11.5) has the effect of adding a constant to  $\ln L$ . This means that units are irrelevant.
- b) Show that (for the one-dimensional case) (11.5) can be derived from the expression for the binned log-likelihood by making the bin size arbitrarily small.

- c) Generate a large sample of events and check the effect of omitting the  $f(\sigma_i)$  terms in (11.6) and (11.7). Are the mass and  $N^{\text{sig}}$  measurements biased?
- d) Suppose another experiment has measured  $M_{Z'} = 151 \pm 2 \text{ GeV}$ . Combine this measurement with our likelihood by (1) adding a term to the likelihood function in the code, (2) combining the likelihood curves. Assume the likelihood of the other experiment is Gaussian in both cases. Compare the results.

## References

- 1** Verkerke, W. and Kirkby, D.P. (2003) The RooFit toolkit for data modeling *eConf C0303241*, MOLT007.
- 2** Moneta, L. *et al.* (2010) The RooStats project. *PoS, ACAT2010*, 057.
- 3** Erler, J. *et al.* (2009) Improved Constraints on  $Z'$  Bosons from electroweak precision data. *JHEP*, **0908**, 017.
- 4** Gross, E. and Vitells, O. (2010) Trial factors or the look elsewhere effect in high energy physics. *Eur. Phys. J. C*, **70**, 525.
- 5** Cowan, G. *et al.* (2011) Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, **71**, 1554.
- 6** Read, A.L. (2002) Presentation of search results: the CLs technique. *J. Phys. G: Nucl. Part. Phys.*, **28**, 2693.
- 7** Read, A.L. (1999) Linear interpolation of histograms. *Nucl. Instrum. Methods A*, **425**, 357.
- 8** Cousins, R.D. (1995) Why isn't every physicist a Bayesian? *Am. J. Phys.*, **63**, 398.
- 9** Casadei, D. (2012) Estimating the selection efficiency *JINST*, **7**, P08021.

## 12

### Applications in Astronomy

*Harrison B. Prosper*

#### 12.1

##### Introduction

As noted in Chapter 1, probability theory was inspired, in part, by the needs of gamblers. But, as Tom Loredo observes [1], a compelling case can be made that the development of statistics was inspired by loftier, if somewhat less colourful, needs: solving problems in astronomy [2]. It is therefore fitting that, after more than a century of relative neglect, astronomers find it increasingly necessary to embrace statistical reasoning, as evidenced by the growing number of conferences in astronomy dedicated to statistics, such as the conference series *Statistical Challenges in Modern Astronomy* [3].

One reason for the resurgence of interest in formal statistics is the rapidly increasing size of the datasets in astronomy. For example, the anticipated dataset from the Large Synoptic Survey Telescope (LSST) [4] exemplifies what astronomers – indeed anyone who is interested, since these data will be made public – will face. When the LSST becomes fully operational, it will produce a 6 GB image every 20 s every night for ten years!<sup>1)</sup> Traditional astronomy will still be possible in which individual objects are explored in detail. However, it is generally accepted that fundamental advances in astronomy, astrophysics, and cosmology will require the large-scale application of statistical methods in much the same way as is done routinely in high energy physics.

The range of statistical analysis problems in astronomy, astrophysics and cosmology is large, but most fall into the following broad (but overlapping) categories:

- *Statistical inference*: An example is extracting statistical summaries from multi-TB image libraries, or inferring the parameters of an astrophysical model, for example the strength of the magnetic field in a core-collapse supernova.
- *Multivariate analysis*: An example is characterising the structure of galaxy clustering in 3-dimensional space, or perhaps also in spectral space, taking into ac-

1) Eventually, the LSST will make 30 TB of data per night, available on a routine basis to anyone who wants them. This is clearly the future of *Big Science*.

count the *heteroscedastic*<sup>2)</sup> nature of astronomical measurements and the bias of observational data towards brighter sources, as first noted by Swedish astronomer Gunnar Malmquist in 1920. Moreover, astronomical data may be incomplete. This is typically the case for observed spectra because they are viewed in spectral bands.

- *Multivariate time series analysis:* An example of growing importance is the use of time series of Type Ia supernova light curves, in multiple spectral bands, to constrain supernova models. Another example, which we hope will one day become routine, is the analysis of periodic and burst gravitational wave signals.
- *Likelihood-free analysis:* Astronomers, like high energy physicists, have recognised the utility of creating simulations from models. If the models are complicated, and the data are multi-dimensional, it may not be possible to write down a likelihood function explicitly. This has motivated approaches, generically referred to as *Approximate Bayesian Computation* (ABC), which do not require an explicit calculation of the likelihood function. ABC is a somewhat better term because ‘likelihood-free’ is misleading as the likelihood function is implicit in the ensemble of realisations of the model. The challenging questions in this approach are: how to explore the parameter space of the model and how to compare real and simulated data?

With the exception of time series analysis, there is clearly considerable overlap between the categories of problems in astronomy, astrophysics and cosmology and the categories of problems in high energy physics. However, it is not possible in a single chapter to do justice to the wide range of applications of statistics in these related fields. Instead, we discuss a few somewhat disparate examples in order to illustrate the diversity of problems. The unifying theme in our choice of problems is that each uses statistical methods that are either directly applicable to high energy physics, or applicable with some field-specific modifications. The examples also have a decidedly Bayesian bias, reflecting the growing importance of Bayesian methods in astronomy. For ease of exposition, we shall use the word ‘astronomy’ as shorthand for astronomy, astrophysics, and cosmology.

## 12.2

### A Survey of Applications

Statistics in astronomy, ‘astrostatistics’ as it has come to be known, covers a wide range of statistical problems, which include the analysis of Poisson point processes (typically, images), heteroscedastic measurements, time series data (often in multiple spectral bands and over multiple channels), parameter estimation of astro-

2) In heteroscedastic data, the variance associated with a measurement may vary from one measurement to the next. An analogous situation arises in the measurement of the proper decay times of particles in high energy physics. Typically, the measurement error is assumed to be Gaussian-distributed, with a variance that can vary from one lifetime measurement to the next.

physical and cosmological models, hypothesis testing, model selection and object classification.

Astronomical data are of course quite different from those in high energy physics. However, the underlying statistical methods are often mathematically identical, albeit described with different jargon. In this section, we examine three applications: the *on/off problem*, *image reconstruction*, and the *fitting of cosmological parameters*, which is a problem of parameter estimation (see Chapter 2 for a detailed discussion). This is then followed by an overview of *nested sampling*, an interesting and powerful algorithm that was developed within the cosmology community.

Unless clarity would be impaired, we denote probability densities by lower-case letters  $p(\dots)$ , probabilities by upper-case letters,  $P(\dots)$ , and  $\pi(\dots)$  denotes prior probability densities.

### 12.2.1

#### The On/Off Problem

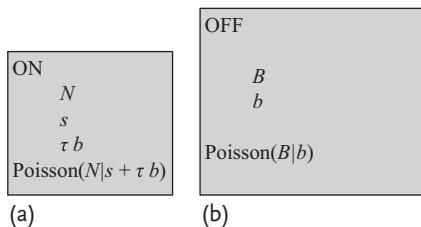
The on/off problem in astronomy [5] consists in determining whether an observed photon count,  $N_{\text{on}}$ , in the direction of a source, say a potential gamma-ray emitter, differs sufficiently from the observed photon count,  $N_{\text{off}}$ , in a direction away from the source to justify claiming a statistically significant photon signal from the source. The on-source count,  $N_{\text{on}}$ , is the number of photons counted during a given time interval  $t_{\text{on}}$ , while  $N_{\text{off}}$ , the off-source count, is the number of photons counted away from the source during a time interval  $t_{\text{off}}$ . By construction, the ratio  $\alpha = t_{\text{on}}/t_{\text{off}}$  is known. The basic assumption is that, *in the absence of a signal*, the expected on-source count is given by  $\mu = \alpha n_{\text{off}}$ , where  $n_{\text{off}}$  is the *unknown* expected off-source count of which  $N_{\text{off}}$  is an estimate. More generally, assuming an additive signal, the expected on-source count is given by  $n_{\text{on}} = s + \mu$ , where  $s$  is the expected signal. We shall follow the convention of denoting parameters, such as  $s$  and  $\mu$ , with lower-case symbols, and we shall use upper-case symbols for observed values.<sup>3)</sup>

The on/off problem is mathematically identical to the high energy physics problem [6, 7] in which a signal region, corresponding to the on-source direction, yields a count  $N$  that is to be compared with count  $B$  in a sideband region, corresponding to the off-source direction. Here,  $\tau = \mu/b$  is the ratio of the expected background count  $\mu$  in the signal region to the expected count  $b$  in the sideband region, assuming the signal there to be negligible. This is illustrated in Figure 12.1.

For notational simplicity, we shall discuss the on/off problem using the high energy physics notation and formulation. To go from the astronomy on/off problem to the signal/sideband problem requires merely a change of notation:  $N_{\text{on}} \rightarrow N$ ,  $N_{\text{off}} \rightarrow B$ ,  $n_{\text{on}} \rightarrow n$ ,  $n_{\text{off}} \rightarrow b$ , and  $\alpha \rightarrow \tau$ .

As has been emphasised in earlier chapters, the likelihood function is (or ought to be) the starting point for any serious statistical analysis (the so-called likelihood-

3) This is the reverse of the convention used by statisticians.



**Figure 12.1** The standard on/off problem comprises two regions, labelled ON (a) and OFF (b) in the cartoon, with data consisting of observed counts  $N$  and  $B$ , respectively, associated with expected (that is, mean) counts,  $s + \tau b$  and  $b$ . The scale factor  $\tau$  is assumed to be accurately known. A key assumption

is that any signal in the OFF region (i.e. the sideband) is negligible. An alternative formulation, which has the virtue of de-cluttering the likelihood function in the ON (that is, signal) region is to write the expected count in that region as  $s + \mu$  and that in the OFF region as  $b = \beta\mu$ , where  $\beta = 1/\tau$ .

free method mentioned above is a misnomer). The on/off problem is no exception. Assuming that the counts are Poisson-distributed, the likelihood function for the on/off problem is

$$p(D|s, b) = \text{Poisson}(N|s + \tau b)\text{Poisson}(B|b), \quad (12.1)$$

where the data ( $N$  and  $B$ ) are denoted generically by  $D$  and where  $\tau$  is assumed to be a known constant.<sup>4)</sup>

Given the likelihood, the goal is to quantify the statistical significance of a putative signal, a problem that has been considered in detail by Li and Ma [5] in the context of gamma-ray astronomy. Li and Ma's solution to this problem uses Wilks' theorem (see Chapter 2), which for the on/off problem may be stated as follows: if the (null) hypothesis – here  $s = 0$  – is true, then for  $\Lambda = p(D|s, \hat{b}(s))/p(D|\hat{s}, \hat{b})$  the quantity  $-2 \ln \Lambda$  will asymptotically (that is, as the number of counts goes to infinity) be distributed as a  $\chi^2$  variate with one degree of freedom. In the numerator,  $\hat{b}(s)$  denotes the maximum-likelihood estimate (MLE) of the parameter  $b$  for a fixed value of  $s$ , while in the denominator,  $\hat{s}$  and  $\hat{b}$  are the unrestricted MLEs.<sup>5</sup>

4) If the scale factor  $\tau$  is not known precisely, but is itself based on some auxiliary observations  $X$ , then (12.1) should be augmented with the likelihood function of these measurements. For example, suppose that  $\tau$  is estimated by the ratio of two additional independent integers  $Q$  and  $M$ , for which the likelihood is

$$p(Q, M | m, \tau) = \text{Poisson}(Q | \tau \cdot m) \\ \cdot \text{Poisson}(M | m),$$

then, for this more general problem, the complete likelihood is given by

$$p(D|s, b, m, \tau) = p(N, B|s, b)p(Q, M|m, \tau),$$

which depends on the four parameters  $s$ ,  $b$ ,  $m$  and  $\tau$ . The scale factor  $\tau$  can sometimes be estimated using a data sample that is disjoint with respect to the sample of interest, but which is otherwise kinematically identical to it. Suppose that, for example, the goal is to estimate the number of  $t\bar{t} \rightarrow e^+e^- + X$  events in the signal region. It may be possible, depending on the nature of the cuts that have been applied, to use a sample of  $t\bar{t} \rightarrow e\mu + X$  events and count how many of them fall in both the sideband and signal regions.

5) Sometimes, the restricted maximum-likelihood estimate  $\hat{b}(s)$  is denoted by  $\hat{\bar{b}}$ .

Li and Ma proposed to quantify the statistical significance of a gamma-ray event using  $Z_L = \sqrt{-2 \ln A}$ . If one found  $Z_L = 3$ , this would be reported as a ‘3 standard deviation event’. For the on/off problem with a known constant  $\tau$ , Li and Ma’s expression for  $Z_L$  (in our notation) may be written as

$$Z_L = \sqrt{2} \left\{ N \ln \left[ \frac{1 + \tau}{\tau} \left( \frac{N}{N + B} \right) \right] + B \ln \left[ (1 + \tau) \left( \frac{B}{N + B} \right) \right] \right\}^{1/2}. \quad (12.2)$$

Equation 12.2, can be considered a standard frequentist solution to the estimation of signal significance in the on/off problem. More details of frequentist solutions can be found in [6, 7].

### 12.2.1.1 A Bayesian Approach

The on/off problem can, of course, be treated using the Bayesian approach. The solution, in principle, is straightforward and direct: one computes the probability of the hypothesis  $H_1$  that  $s > 0$ ,

$$P(H_1|D) = \frac{B_{10} P(H_1)}{B_{10} P(H_1) + P(H_0)}, \quad (12.3)$$

where  $H_0$  denotes the hypothesis  $s = 0$ ,  $B_{10}$  is the Bayes factor

$$B_{10} \equiv \frac{p(D|H_1)}{p(D|H_0)} \quad (12.4)$$

and  $P(H_1)$  and  $P(H_0)$  are the prior probabilities of the associated hypotheses<sup>6</sup>. If one is prepared to make the choice  $P(H_1) = P(H_0)$ , then the probability  $P(H_1|D)$  reduces to

$$P(H_1|D) = \frac{B_{10}}{B_{10} + 1}. \quad (12.5)$$

It is not overly controversial to argue that this choice favours neither hypothesis. The difficulty arises in the calculation of the Bayes factor,  $B_{10}$ , which is computed from the following functions:

$$p(D|H_1) \equiv \int ds \int db p(D|s, b) \pi(s, b), \quad (12.6)$$

$$p(D|H_0) \equiv \int db p(D|b) \pi(b), \quad (12.7)$$

where we denote by  $\pi(s, b)$  and  $\pi(b)$  the prior densities for the parameters associated with the two hypotheses  $H_1$  and  $H_0$ , respectively. A necessary condition

6) Note that these are probabilities, not probability densities.

for a Bayes factor to make sense, and hence the probability  $P(H_1|D)$ , is that it must not depend on arbitrary constants. For this to be true, it is sufficient that the prior densities  $\pi(s, b)$  and  $\pi(b)$  be proper, that is, integrate to one. Ideally,  $\pi(s, b)$  and  $\pi(b)$  should be proper *evidence-based* priors.<sup>7)</sup> If one were to use improper priors, which, by definition, do not integrate to one and therefore are known only to within *arbitrary* scale factors,<sup>8)</sup> the Bayes factor would likewise be defined only to within an arbitrary scale factor. Consequently, the probability  $P(H_1|D)$  would be undetermined.

However, the requirement that the priors be proper can be weakened slightly if the likelihoods contain parameters that are in common, as is the case here for the off-source expected background  $b$ . To see this, it is convenient to factorise  $\pi(s, b) = \pi(b|s)\pi(s)$  – and to calculate the marginal likelihood

$$p(D|s) = \int_0^\infty p(D|s, b)\pi(b|s)db. \quad (12.8)$$

Observe that  $p(D|b) = p(D|s = 0, b)$ . Furthermore, in many applications  $\pi(b|s)$  and  $\pi(b)$  are the same function, in which case we can write (12.6) and (12.7) as

$$p(D|H_1) = \int_0^\infty p(D|s)\pi(s)ds, \quad (12.9)$$

$$p(D|H_0) = p(D|s = 0). \quad (12.10)$$

Here, it is permissible to use an improper prior for  $b$ , for example, the Jeffreys prior  $\pi(b|s) = \pi(b) = C/\sqrt{b}$ , because this prior appears in both  $p(D|H_1)$  and  $p(D|H_0)$ . Therefore, the arbitrary scale factor cancels in the ratio  $B_{10} = p(D|H_1)/p(D|H_0)$ . However, since the signal prior  $\pi(s)$  appears in  $p(D|H_1)$  only, it is absolutely essential that it does not contain an arbitrary scale factor – that is,  $\pi(s)$  must be proper. The quantities  $p(D|H_i)$ ,  $i = 0, 1$ , are often referred to as *evidences*. The larger the value of  $p(D|H_i)$  the greater the evidence it provides in support of hypothesis  $H_i$ . In the following example we apply these ideas to the on/off problem.

### Example 12.1 The on/off problem

Here, we compute the Bayes factor  $B_{10}$  for the standard on/off problem with the likelihood function given in (12.1). This requires the calculation of the evidences  $p(D|H_1)$  and  $p(D|H_0)$  for the signal-plus-background and background-only hypotheses, respectively. It is convenient to compute first the marginal (or

7) The more common term is a ‘subjective’ prior. However, we prefer ‘evidence-based’ because this terminology more closely reflects how these priors are constructed in practice.

8) In spite of their name, improper priors are not intrinsically problematic and can rigorously be incorporated into probability theory, see [8].

integrated) likelihood

$$\begin{aligned} p(D|s, H_1) &= \int_0^\infty \text{Poisson}(N|s + \tau b) \text{Poisson}(B|b) \pi(b) db \\ &= C \frac{1}{(1 + \tau)^2 B} \\ &\times \sum_{r=0}^N \text{Beta}\left(\frac{\tau}{1 + \tau}; N - r + 1, B\right) \text{Poisson}(r|s), \end{aligned} \quad (12.11)$$

where

$$\text{Beta}(\theta; \alpha, \beta) \equiv \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad (12.12)$$

and  $\Gamma(z) \equiv \int_0^\infty e^{-t} t^{z-1} dt$  is the gamma function. For simplicity, we have used the improper prior  $\pi(b) = C$ , where  $C$  is an arbitrary constant, such as  $C = 1$ . (The Jeffreys prior would be a better choice.) This choice for  $C$  yields the interesting sum rule

$$\sum_{N=0}^\infty \sum_{r=0}^N \text{Beta}\left(\frac{\tau}{1 + \tau}; N - r + 1, B\right) \text{Poisson}(r|s) = (1 + \tau)^2 B. \quad (12.13)$$

**Evidence for hypothesis  $H_0$ :** The evidence for the background-only hypothesis,  $H_0$ , is given by

$$\begin{aligned} p(D|H_0) &= p(D|s = 0, H_1) \\ &= C \frac{1}{(1 + \tau)^2 B} \text{Beta}\left(\frac{\tau}{1 + \tau}; N + 1, B\right). \end{aligned} \quad (12.14)$$

**Evidence for hypothesis  $H_1$ :** Suppose we are given a prediction for the signal,  $S \pm \delta S$ . We can model this prediction with a gamma density,

$$\pi(s) = q \exp(-qs)(qs)^Q / \Gamma(Q + 1), \quad (12.15)$$

where  $Q \equiv (S/\delta S)^2$  and  $q \equiv Q/S$ . This model can be motivated as follows. We apply Bayes' theorem to a Monte Carlo prediction of the number of signal events. Suppose this yields a signal count  $Q$  with likelihood function  $\text{Poisson}(Q|a)$ , where the parameter  $a$  is the mean count. The use of a flat or Jeffreys prior for  $a$  leads to a posterior density for  $a$  that is a gamma density. We then write  $qs = a$ , where  $q$  is a known scale factor (for example, the ratio of the data to Monte Carlo integrated luminosities), and arrive at (12.15). Note that for large  $Q$ , the signal prior  $\pi(s)$  approaches a Gaussian. Inserting (12.11) and (12.15) into (12.9), we obtain the evidence for the signal:

$$\begin{aligned} p(D|H_1) &= C \frac{q^2}{(1 + q)^2 Q} \frac{1}{(1 + \tau)^2 B} \\ &\times \sum_{r=0}^N \text{Beta}\left(\frac{\tau}{1 + \tau}; N - r + 1, B\right) \text{Beta}\left(\frac{1}{1 + q}; r + 1, Q\right). \end{aligned} \quad (12.16)$$

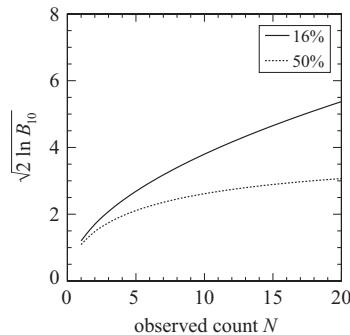
As expected, the evidences for the two hypotheses are defined only to within an arbitrary constant due to the impropriety of the background prior. However, because the constant  $C$  is the same for both hypotheses, the Bayes factor  $B_{10} = p(D|H_1)/p(D|H_0)$  is nevertheless well defined.

As an illustration of how these factors work in practice, consider the results published by the DØ collaboration in support of their top-quark discovery claim [9]. DØ observed  $N = 17$  events, with a background estimate in the signal region of  $\hat{\mu} = 3.8 \pm 0.6$  events. From this background estimate, we can compute an effective sideband count of  $B = (\hat{\mu}/\delta\mu)^2 = (3.8/0.6)^2 = 40.1$  events and an effective scale factor of  $\tau = \mu/b \approx \hat{\mu}/B = 0.0947$ . For simplicity, we set  $C = 1$  and assume the *simple* signal hypothesis  $s = 14$  events. In the limit of simple signal hypotheses, (12.16) reduces to (12.11) with  $s$  set to 14 events. This yields  $p(D|H_0) = 3.0 \cdot 10^{-6}$  and  $p(D|H_1) = 9.3 \cdot 10^{-2}$ , that is, a Bayes factor of  $B_{10} = 3.1 \cdot 10^4$ .

The Bayes factor can be more readily interpreted if mapped to the scale  $Z_{\text{BF}} \equiv \sqrt{2 \ln B_{10}} = 4.5$ . The quantity  $Z_{\text{BF}}$  is a Bayesian analog of an ‘ $n\sigma$ ’. Figure 12.2 illustrates the general behaviour of the Bayes factor as a function of the number of observed events  $N$  and the uncertainty on the background estimate.

Unfortunately, an evidence-based prior for the signal is often not available in astronomical applications. However, over the decades, several formal procedures for constructing suitable priors have been developed by statisticians. One such procedure is the *intrinsic* (or *expected-posterior*) prior construction of Berger and co-workers [10], which for the on/off problem proceeds as follows:

- Let  $p_0(s|D) = p(D|s)\pi_0(s)/p_0(D)$ , where  $p_0(D) = \int p(D|s)\pi_0(s)ds$ ,  $p(D|s)$  is given in (12.11), and  $\pi_0(s)$  is a prior that works well for parameter estimation in the sense that it provides estimates that have desirable properties (see, for example, [11] and references therein).



**Figure 12.2** Plots of  $\sqrt{2 \ln B_{10}}$  as a function of the observed count  $N$ , for two values of the relative background uncertainty, 16% and 50%. As expected, the Bayes factor is lower for the higher background uncertainty. The signal hypothesis is taken to be  $s = N - \hat{\mu}$ .

2. Then, provided that  $p_0(D) < \infty$ , the intrinsic prior for  $s$  is

$$\pi_I(s) = \overline{p_0(s|D)} \equiv \sum_D p_0(s|D) p(D|s=0),$$

which is proper by construction and where the average is with respect to the background-only (that is, null) distribution,  $p(D|s=0)$ . The intuition here is that given data  $D$ , from some subsidiary measurements, it would be perfectly reasonable to use the posterior density  $p(s|D)$  as a proper evidence-based prior  $\pi(s)$  to compute  $p(D|H_1)$ . However, since we do not have such data and  $D$  is consequently *unknown*, we follow standard Bayesian practice and marginalise over the unknown quantity  $D$ . Doing so using the null distribution,  $p(D|s=0)$ , is more cautious than using the distribution  $p(D|H_1)$  of the alternative hypothesis. We illustrate the procedure in the following example.

### Example 12.2 The intrinsic prior for the on/off problem

The calculation of an intrinsic prior for the on/off problem entails marginalising over the space of possible observations. This necessarily requires specification of this space, that is, the *ensemble*. In this example, we assume an ensemble in which the background count  $B$  is fixed. Again, for simplicity, we take  $\pi_0(s) = 1$ . We find that

$$\begin{aligned} p_0(D) &= \int p(D|s)\pi_0(s)ds, \\ &= \frac{1}{(1+\tau)^2 B} \sum_{r=0}^N \text{Beta}\left(\frac{\tau}{1+\tau}; N-r+1, B\right) \end{aligned} \quad (12.17)$$

is finite and yields the expression

$$p_0(s|N) = \sum_{r=0}^N p_{N,r} \text{Poisson}(r|s), \quad (12.18)$$

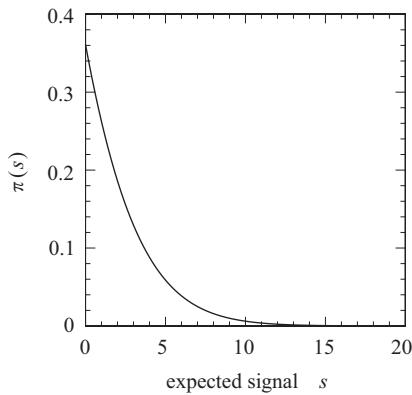
for a posterior density that would be suitable for estimating the signal, where

$$p_{n,r} \equiv \frac{\text{Beta}\left(\frac{\tau}{1+\tau}; n-r+1, B\right)}{\sum_{j=0}^n \text{Beta}\left(\frac{\tau}{1+\tau}; n-j+1, B\right)}. \quad (12.19)$$

Therefore, the intrinsic prior  $\pi_I(s)$  for the on/off problem is given by

$$\begin{aligned} \pi_I(s) &= \overline{p_0(s|n)} = \sum_{n=0}^{\infty} p_0(s|n) p(n|s=0) \\ &= \frac{1}{(1+\tau)^2 B} \sum_{n=0}^{\infty} p_0(s|n) \text{Beta}\left(\frac{\tau}{1+\tau}; n+1, B\right), \end{aligned} \quad (12.20)$$

which, by construction, satisfies  $\int_0^\infty \pi_I(s)ds = 1$ . Figure 12.3 shows the intrinsic prior assuming  $B = 40.1$  events and  $\tau = 0.0947$ , that is, the background data from the DØ top-quark discovery results.



**Figure 12.3** The intrinsic prior, (12.20), for the sideband count  $B = 40.1$  and a ratio of signal to sideband count of  $\tau = 0.0947$ .

#### 12.2.1.2 Conclusion

It is important to note and keep in mind that the frequentist and Bayesian solutions for the on/off problem, sketched above, are conceptually quite different. The first, in effect, answers the question: ‘What is the probability to obtain a count equal to  $N$  or larger?’ The second, in contrast, answers the question: ‘What is the probability that  $s > 0$ ?’ While these are quite different questions, it is still helpful to have a way to calibrate the two answers. This is most easily achieved by computing the number

$$Z_{\text{BF}} = \sqrt{2 \ln B_{10}}, \quad (12.21)$$

introduced in the example above. It is readily shown that in the limit in which the expected on-source background  $\mu = \tau b$  is known exactly,  $Z_{\text{BF}}$  and  $Z_L$  approach  $s/\sqrt{\mu}$  when the on-source signal to background ratio  $s/\mu$  obeys:  $s/\mu \ll 1$ .

#### 12.2.2 Image Reconstruction

Data in astronomy are often in the form of images, which today are obtained with large-format digital cameras. The standard practice is to remove instrumental effects from these images using the known *point spread functions* (PSF) of the relevant telescopes. The point spread function describes the blurring that the telescope imposes on the image that enters the telescope. For ground-based telescopes, there is, in addition, the blurring due to the atmosphere. An analog of the PSF in high energy physics is the *jet response function* of a particle detector, which, in the simplest but common case, describes the transverse momentum smearing imposed by the detector and the jet reconstruction algorithm.

In astronomy, the removal of instrumental effects from images is referred to as *image reconstruction*, while the analogous procedure in high energy physics is called

*unfolding* (see Chapter 6 and, for example, the proceedings of PHYSTAT 2011 [12]). Mathematically, they are precisely the same procedure. The main difference between the two fields is the ubiquitous use of *maximum entropy methods* (MEM) in astronomy and their relatively infrequent use in high energy physics. Maximum entropy methods have enjoyed considerable success in image reconstruction, not only in astronomy but in many other fields including the field of medical imaging. They may therefore be of interest to particle physicists who need to solve mathematically identical analysis problems.

We begin with the statement of the problem and the motivation for MEM. We end with a brief discussion of some practical issues.

#### 12.2.2.1 Of Monkeys and Kangaroos

Consider an image comprising  $M$  pixels. In standard applications, the pixels are all of the same size. However, in more sophisticated approaches the pixel size may vary in size according to the level of detail in a given part of the image. An example of a commercialised version of a high-performance adaptive method, the so-called *Pixon* method, is described in [13].

For many astronomical instruments (and the same is true of high energy physics) the relationship between the expected data  $d(x)$ , an image  $f(y)$ , and the point spread function  $R(x, y)$ , is a *Fredholm equation* of the first kind:

$$d(x) = \int dy R(x, y) f(y) . \quad (12.22)$$

Typically, (12.22) is discretised so that the (unknown) expected mean  $d_i$  per data pixel is given by

$$d_i = \sum_{j=1}^N R_{ij} f_j , \quad (12.23)$$

where  $N$  is the number of pixels in the image space.

The goal is to infer a discretised image  $f \equiv f_1, \dots, f_N$  given observed data  $D \equiv d_1, \dots, d_M$ . Since these data comprise a finite number of counts and are always noisy, they are consistent with a truly enormous number of possible images  $f$ . The inverse problem is therefore ill-posed. The only way to arrive at an acceptable solution is to impose, *a priori*, a sufficient number of conditions on the vast collection of images with which the data are compatible. The problem of course is to decide what those conditions should be.

Given the necessity of imposing prior conditions, it is natural to consider image reconstruction from a Bayesian perspective. The task is to compute

$$p(f|D) = p(D|f) \frac{\pi(f)}{\pi(D)} , \quad (12.24)$$

where  $\pi(f)$  is the prior that imposes constraints on the image and  $p(D|f)$  is the likelihood of the data. If the crosstalk between pixels is negligible, then  $p(D|f)$

could be modelled as the product of  $M$  Poisson distributions,

$$p(D|f) = \prod_{i=1}^M \text{Poisson}(D_i|d_i), \quad (12.25)$$

which may be replaced by Gaussians if the photon counts are large enough.

We now consider the prior  $\pi(f)$ . In the pioneering work of Gull and Daniell [14], the prior probability of an image,  $\pi(f)$ , is taken to be proportional to the number of ways the image can be formed. There are many different ways to render this choice plausible. One way, introduced by Gull and Daniell, involves monkeys and kangaroos.

**Monkeys** Imagine a team of monkeys who draw pictures by randomly throwing  $n$  tiny darts at a screen divided into  $M$  pixels. The number of darts per pixel is taken to be the intensity of the picture at that point. Every time a picture matches the data it is saved otherwise the picture is rejected and the monkeys try again. The most likely outcome, consistent with the data, is the picture that can be formed in the largest number of ways. The number of ways to obtain the monkeys' masterpieces is the multinomial coefficient

$$\Omega = \frac{n!}{\prod_{i=1}^M n_i!}, \quad (12.26)$$

where  $n_i$  is the number of darts in pixel  $i$ . From the definition of entropy,  $S \propto \ln \Omega$ , and in the large  $n$  limit, we arrive at the *Shannon entropy* [15, 16],

$$S \propto - \sum_{i=1}^M P_i \ln P_i, \quad (12.27)$$

for the probability distribution  $P_i$ , where  $P_i = f_i/n \approx n_i/n$ . This form of the entropy was first used by Frieden [17] in the context of optical image restoration.

**Kangaroos** Suppose you are given the following information: 1/3 of all kangaroos have blue eyes ('BE'), and 1/3 of all kangaroos are left-handed ('LH'). On the basis of this information alone you are to estimate the proportion of kangaroos  $P$  that are both left-handed and blue-eyed. We can draw up a  $2 \times 2$  contingency table, shown below, that is parameterised using the unknown probability  $P$ .

As required, the probabilities in the left column sum to 1/3; likewise for those in the first row. Gull and Daniell suggest that we favour the solution with the maximum entropy. The entropy of the probabilities in Table 12.1 is given by

$$S \propto P \ln P + 2(1/3 - P) \ln(1/3 - P) + (1/3 + P) \ln(1/3 + P). \quad (12.28)$$

Maximising  $S$  with respect to  $P$  yields the result  $P = 1/9$ .

Of all the possible solutions, it can be argued that this is the most reasonable, not because it is the most probable; it may, or may not, be. Rather it is reasonable

**Table 12.1** Probability of blue-eyed, left-handed kangaroos.

	Blue-eyed	Not blue-eyed
Left-handed	$P$	$1/3 - P$
Not left-handed	$1/3 - P$	$1/3 + P$

because it is the *simplest* solution consistent with the available information, namely, the two probabilities  $P(\text{BE}) = 1/3$  and  $P(\text{LH}) = 1/3$ , and the implicit constraint  $P \leq 1/3$ . Note that *any* other solution would require the use of a non-zero correlation between the prevalence of blue eyes and left-handedness. But, using a value for the correlation coefficient would impose an unwarranted assumption.

One may object that a value for the correlation coefficient has been assumed, namely, zero. The counter-argument is that, in fact, no such assumption has been made, indeed can be made, because the correlation coefficient is an *additional* parameter that is not identifiable from the information provided. Consequently, it simply does not enter the problem.

#### 12.2.2.2 Maximum Entropy Method in Practice

These gedankenexperiments, and many different lines of reasoning, suggest an image prior of the form

$$\pi(f|\alpha) = \exp\left(-\alpha \sum_{j=1}^N f_j \ln f_j\right), \quad (12.29)$$

where  $\alpha$  is a nuisance parameter with prior  $\pi(\alpha)$ . The posterior density of the image  $f$ , and the parameter  $\alpha$ , can therefore be written as

$$\begin{aligned} p(f, \alpha | D) &\propto \left[ \prod_{i=1}^M \text{Poisson}\left(D_i | \sum_{j=1}^N R_{ij} f_j\right) \right] \\ &\times \exp\left(-\alpha \sum_{j=1}^N f_j \ln f_j\right) \pi(\alpha). \end{aligned} \quad (12.30)$$

The maximum entropy method solution for the image is the mode of the density  $p(f|D) = \int p(f, \alpha | D) d\alpha$ . This of course is only one of several different estimates of the image that could be computed from the posterior density  $p(f|D)$ . But, in order to compute  $p(f|D)$ , it is necessary to specify the prior  $\pi(\alpha)$ . Typically, however, an evidence-based prior for  $\alpha$  is not available. In this case, a reasonable approach would be first to compute the marginal likelihood  $p(D|\alpha)$  by integrating over the parameters  $f_j$ . Having reduced the likelihood to the single parameter  $\alpha$ , we then compute the Jeffreys prior for  $\alpha$  from  $p(D|\alpha)$ .

The key characteristic of the MEM solution is that it favours images that are as flat as possible, consistent with the data. Therefore, it tends to suppress noise and

eliminate structure where none is needed, which is good. But the smoothing occurs uniformly over every part of the image. Therefore, a region rich in detail will be smoothed as much as one devoid of features. This will tend to over-smooth the information-rich regions and under-smooth regions that are relatively featureless. Clearly, what is needed is to adapt the pixel size to the ‘information density’ suggested by the data. This is precisely what the Pixon method attempts to do [13].

### 12.2.3

#### Fitting Cosmological Parameters

The use of Bayesian methods has become quite common in astronomy during the past decade [18]. Indeed, their application in astronomy is arguably more extensive and sophisticated than in high energy physics.

It is generally accepted that the Bayesian approach provides a coherent framework in which to think about statistical analysis problems. It also provides useful solutions. But, like any purely mathematical theory, they are no substitute for the judicious use of scientific insight, such as that which led to the discovery of the accelerating expansion of the Universe [19, 20] in 1998, and the awarding of the 2011 Nobel prize in physics to Perlmutter, Schmidt and Riess. Interestingly, Bayesian methods were used in these discoveries to fit the cosmological parameters, which nicely illustrates the promise and challenge of these methods. We use this example in order to examine how Bayesian methods are being used in astronomy. For a good pedagogical introduction, we recommend the review by Trotta [18].

For pedagogical completeness, we begin with a sketch of the standard model of cosmology. Then we consider how its parameters have been determined using both frequentist and Bayesian methods, where for the latter we focus on the issue of priors.

##### 12.2.3.1 Cosmology in a Nutshell

The standard model of cosmology makes two important assumptions: (1) On a sufficiently large scale, the Universe is isotropic and homogeneous, and (2) Einstein’s general relativity is the correct (classical) theory of space-time. The homogeneity and isotropy assumption is spectacularly accurate for the cosmic microwave background (CMB) and is approximately true for matter on scales above  $\sim 100$  Mpc. Given those two assumptions, one arrives at a beautifully simple cosmology [21] based on Einstein’s equations, which yield the Friedmann equation,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3}, \quad (12.31)$$

the energy-momentum conservation law,

$$\dot{\epsilon} + 3(p + \epsilon)\frac{\dot{a}}{a} = 0, \quad (12.32)$$

and the Friedmann–Lemaître–Robertson–Walker (FLRW) space-time metric,

$$ds^2 = c^2 dt^2 - a^2(t) \left[ \frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2 \theta d\phi^2) \right], \quad (12.33)$$

where  $a(t)$  is the dimensionless scale factor,  $t$  is the time since the Big Bang,  $\dot{a} \equiv da/dt$ ,  $G$  is the gravitational constant,  $\epsilon = \rho c^2$  is the energy density,  $p = w\epsilon$  is the pressure,  $-\infty < K < \infty$  is the spatial curvature, and  $\Lambda$  is the cosmological constant. Equation 12.33 is expressed in spherical coordinates  $(r, \phi, \theta)$  and the radial coordinate  $r$  is defined so that, at the present time  $t = t_0$ , the proper area of a sphere centred at any arbitrarily chosen origin is  $4\pi r^2$ . Conventionally, symbols with subscript 0 denote quantities evaluated today (at  $t = t_0$ ).

The proper distance  $d(t)$  – that is the distance between two spatially separated points at the same cosmic time  $t$  – scales as

$$d(t) = a(t)\chi, \quad (12.34)$$

where the comoving distance  $\chi$  is related to the radial coordinate  $r$  by

$$\chi = \int_0^r \frac{da}{\sqrt{1 - Ka^2}} = \sin^{-1} \frac{\sqrt{K}r}{\sqrt{|K|}}, \quad (12.35)$$

that is,

$$r = \frac{\sin(\sqrt{|K|}\chi)}{\sqrt{|K|}}. \quad (12.36)$$

By construction,  $\chi$  coincides with the proper distance  $d(t)$  between any two points, today. If the spatial curvature  $K = 0$ , then  $r = \chi$ . For  $K < 0$ , (12.36) becomes  $r = \sinh(\sqrt{|K|}\chi)/\sqrt{|K|}$ .

With the usual definitions [21] for the critical density,  $\rho_{c0} \equiv 3H_0^2/8\pi G$ , the cosmological constant density,  $\rho_\Lambda \equiv \Lambda c^2/8\pi G$ , and the matter density today,  $\rho_0$ , and defining the Hubble parameter  $H \equiv \dot{a}/a$  of which the Hubble constant  $H_0$  is its present-day value, we can write the Friedmann equation (12.31) in the standard form

$$H^2 = H_0^2 \left[ \frac{\Omega_M}{a^3} + \frac{\Omega_K}{a^2} + \Omega_\Lambda \right], \quad (12.37)$$

where  $\Omega_M \equiv \rho_0/\rho_{c0}$ ,  $\Omega_K \equiv -Kc^2/H_0^2$ ,  $\Omega_\Lambda \equiv \rho_\Lambda/\rho_{c0}$  and  $\Omega_M + \Omega_K + \Omega_\Lambda = 1$ .<sup>9</sup>

Finally, since a light ray travels on a null geodesic, defined by  $c dt = a(t)d\chi$ , the comoving distance  $\chi$  is related to the redshift  $z$  by

$$\begin{aligned} \chi(z) &= c \int_{1/(1+z)}^1 \frac{da}{a\dot{a}} \\ &= \frac{c}{H_0} C(z; \Omega_M, \Omega_\Lambda). \end{aligned} \quad (12.38)$$

9) We have omitted the contribution from radiation because it is important only in the early Universe.

The dimensionless function  $C$  is given by

$$C \equiv \int_{1/(1+z)}^1 \frac{dx}{x^2 \sqrt{\Omega_M/x^3 + (1 - \Omega_M - \Omega_A)/x^2 + \Omega_A}}, \quad (12.39)$$

where we have changed the integration variable to  $x$  in order to avoid possible confusion with the scale factor  $a = 1/(1+z)$ .

**Supernova cosmology** In the absence of an assumption about the spatial curvature, the standard cosmological model is defined by two independent parameters,  $\Omega_M$  and  $\Omega_A$ . If one assumes  $\Omega_K = 0 = 1 - \Omega_M - \Omega_A$ , that is a flat Universe, the model depends on a single parameter, which may be taken to be  $\Omega_M$ . These parameters were determined independently by the High- $z$  Supernova Search Team [19] and the Supernova Cosmology Project (SCP) [20] in their Nobel prize-winning discovery of the accelerating expansion of the Universe. The discovery was based on the analyses of measurements of the distances and redshifts of Type Ia supernovae. These statistical analyses serve as interesting case studies of a high-profile discovery in which parameter estimation was critical.

The flux of energy received from a Type Ia supernova is given by

$$\begin{aligned} f &= \frac{L}{4\pi d_L^2} \\ &= f_0 10^{-2m/5}, \end{aligned} \quad (12.40)$$

where  $L$  is the luminosity of the supernova,  $d_L \equiv (1+z)r$  is its luminosity distance,  $f_0$  is the flux from objects of magnitude zero, and  $m$  is the apparent magnitude. As noted above, the radial coordinate  $r$  is defined so that the proper area of a sphere (centred at the supernova) is  $4\pi r^2$  at the present time. The absolute magnitude  $M$  is defined by the flux  $f_M = f_0 10^{-2M/5}$  from an object at a distance of 10 pc. Given  $f_M/f = 10^{2\mu/5} = (d_L/10^5)^2$ , where  $\mu = m - M$  is the distance modulus, we arrive at the distance modulus–redshift relation

$$\begin{aligned} \mu &= 5 \log_{10}[(1+z)r(z)] + 25 \\ &= 5 \log_{10}[(1+z)H_0 r(z)/c] + 25 - 5 \log_{10}[H_0/c]. \end{aligned} \quad (12.41)$$

In practice, we write this as

$$\mu(z; \Omega_M, \Omega_A, Q) = 5 \log_{10}[(1+z)D_L(z; \Omega_M, \Omega_A)] + Q, \quad (12.42)$$

where the constant  $Q$  depends on how  $\mu$  has been corrected for effects such as extinction due to dust and finite filter bandwidth. The quantity  $D_L \equiv H_0 r(z)/c$  is a dimensionless function, independent of the Hubble constant. From (12.38) and (12.39), and noting that  $i\sqrt{\Omega_K} = \sqrt{K}c/H_0$ , we can write  $D_L$  as

$$\begin{aligned} D_L &= \frac{\sin(i\sqrt{\Omega_K} C)}{i\sqrt{\Omega_K}} \\ &= \frac{\sinh(\sqrt{\Omega_K} C)}{\sqrt{\Omega_K}}. \end{aligned} \quad (12.43)$$

Equation 12.43 goes smoothly to the solution for  $\Omega_K = 0$ , and to  $D_L = \sin(\sqrt{|\Omega_K|}C)/\sqrt{|\Omega_K|}$  for  $\Omega_K = 1 - \Omega_M - \Omega_A < 0$ .

### 12.2.3.2 Statistical Analysis of Supernovae Data

In the papers announcing the discovery of the accelerating expansion [19, 20], the Type Ia supernovae data were analysed using both frequentist and Bayesian methods. Perlmutter *et al.* fitted (12.42) using both the Feldman–Cousins method [22] and a Bayesian method, while Riess *et al.* used a Bayesian method. In both approaches, the likelihood function was the same – a product of Gaussians,

$$p(D|\Omega_M, \Omega_A, Q) = \left[ \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \right] \exp\left(-\frac{\chi^2}{2}\right), \quad (12.44)$$

where

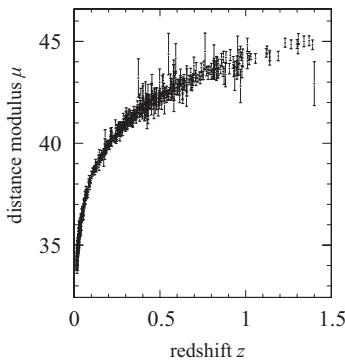
$$\chi^2 = \sum_i [\mu_i - \mu(z_i; \Omega_M, \Omega_A, Q)]^2 / \sigma_i^2, \quad (12.45)$$

and  $D$  denotes the data  $\mu_i \pm \sigma_i$  and  $z_i$ , which are the measured distance modulus and redshift, respectively, of supernova  $i$ . Figure 12.4 shows a compilation of distance modulus–redshift data for 557 Type Ia supernovae. A maximum-likelihood fit of these data using ROOT, without constraints, yields the results

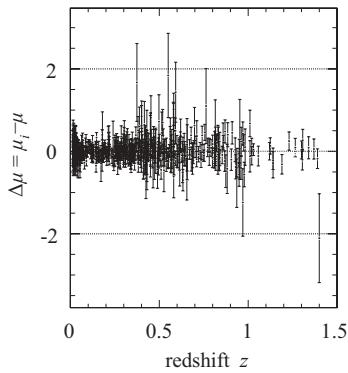
$$\begin{aligned} \Omega_M &= 0.30 \pm 0.05, \\ \Omega_A &= 0.77 \pm 0.09, \end{aligned} \quad (12.46)$$

with a  $\chi^2$  per degree of freedom of 0.979, indicating an excellent fit. These results imply that  $\Omega_K = -0.07 \pm 0.10$ , which is perfectly consistent with  $\Omega_K = 0$ . The residuals for this fit are shown in Figure 12.5.

A maximum-likelihood analysis of these data is straightforward. Moreover, one can learn more about the fit by drawing contours in the  $\Omega_M - \Omega_A$  plane after first



**Figure 12.4** The Union2 compilation of distance modulus–redshift data for 557 Type Ia supernovae [23].



**Figure 12.5** The residuals from a maximum-likelihood fit of the Union2 Type Ia supernova dataset [23]. The  $\chi^2$  per degree of freedom is 0.979.

profiling the likelihood function with respect to the nuisance parameter  $Q$ . However, a likelihood analysis does not permit a *direct* probabilistic interpretation of these contours. The permissible interpretation is the usual *indirect* frequentist one: a 68% region in the  $\Omega_M - \Omega_A$  plane is a member of an infinite ensemble of regions that cover the true value of the parameters of interest; that is a 68% region in the  $\Omega_M - \Omega_A$  plane is expected to be an approximate 68% confidence region. It is worth noting that, just as is true of intervals, the definition of a confidence region is not unique. Indeed, infinitely many such regions can be defined, a mathematical fact that is sometimes overlooked. As mentioned earlier, Feldman–Cousins [22], which uses a specific protocol for choosing intervals and regions, that is the ‘ordering rule’, was one of the methods used in the SCP analysis.

For a *direct* probabilistic interpretation, a Bayesian perspective is needed, which seems to be the perspective favoured by both supernova teams. This approach, however, requires the specification of a prior density for the parameters  $\Omega_M$ ,  $\Omega_A$ , and  $Q$ , which occasions the usual Bayesian controversies.

**Cosmological priors** The standard choice for the cosmological parameters is a flat prior, with the restriction of the parameters to the region considered physical, for example  $\Omega_M > 0$ . Flat priors for the parameters of interest are used both in the analysis of supernova data as well as the analysis of the cosmic microwave background (CMB) data from the Wilkinson Microwave Anisotropy Probe (WMAP) [24].

This has led to a paradoxical situation. On the one hand, the choice of a multi-parameter flat prior can be assailed by all the usual arguments against the use of flat priors. A flat prior does not necessarily imply weak prior information; flat priors do not necessarily guarantee proper posterior densities; and the chief question remains: why should the prior be flat in the parameters  $\Omega_M$ ,  $\Omega_A$  and  $Q$  and not, for example, in their logarithm? These are all legitimate criticisms. On the other hand, the results from the High- $z$  Team, from SCP, WMAP and other projects like the Sloan Digital Sky Survey (SDSS) are not obviously crazy. On the contrary, a beauti-

fully coherent picture of the standard model of cosmology has emerged from these projects. So what is the problem?

For parameter estimation, the results obtained by the various research groups appear to be reasonable in spite of their use of flat priors. The problem arises in the growing industry of Bayesian cosmological model selection. Here the choice of priors matters more because model selection results are more sensitive to the prior than is the case for parameter estimation. This leads us to our next topic.

#### 12.2.3.3 Model Complexity

It may seem obvious that a model with ten free parameters contains ten free parameters. However, suppose the data, together with whatever prior information is at hand, are unable to say anything useful about eight of them. It would seem then more sensible to regard the model as one with effectively two parameters only. Trotta [18] gives the following instructive example: consider the measurement of a periodic signal, modelled as follows:

$$f(t) = A[1 + \theta \cos(t + \delta)] . \quad (12.47)$$

In principle, this is a 3-parameter problem, with parameters  $A$ ,  $\theta$ , and  $\delta$ . But what if the parameter  $\theta$  is so small that the oscillatory term barely registers above the noise? In this case we can say almost nothing precise about the value of  $\theta$ , and much less about the phase  $\delta$ . Essentially, only the equilibrium value  $A$  can be measured and consequently the model has only one free parameter. It seems reasonable (at least at first) to consider the model with a single effective free parameter to be less complex than that with three. But to be useful this intuitive notion must be made precise.

Consider the *Kullback–Leibler (KL) divergence* [25]

$$\begin{aligned} \kappa(p, \pi) &\equiv \int p(\theta|D, M) \ln \frac{p(\theta|D, M)}{\pi(\theta|M)} d\theta \\ &= -\ln p(D|M) + \overline{\ln p(D|\theta, M)} \end{aligned} \quad (12.48)$$

between the posterior density  $p(\theta|D, M)$ , given some data  $D$  and some model  $M$ , and the prior  $\pi(\theta|M)$ . Here the evidence  $p(D|M)$  for model  $M$  is given by

$$p(D|M) = \int p(D|\theta, M) \pi(\theta|M) d\theta , \quad (12.49)$$

and the average  $\overline{\ln p(D|\theta, M)}$  is with respect to  $p(\theta|D, M)$ .

The KL divergence has many interpretations. For example, when the posterior density is close to the prior, it can be interpreted as the distance between the densities in the space of probability densities. More usefully, it can be regarded as a measure of the *expected* amount of information in the data. Defining

$$\hat{\kappa} \equiv \ln p(D|\hat{\theta}, M) - \ln p(D|M) , \quad (12.50)$$

which may be interpreted as an *estimate* of the amount of information in the data, Spiegelhalter *et al.* [26] propose

$$k_{\text{eff}} \equiv \hat{\kappa} - \kappa(p, \pi) = \ln p(D|\hat{\theta}, M) - \overline{\ln p(D|\theta, M)} \quad (12.51)$$

as a Bayesian measure of model complexity. This quantity is the effective number of parameters, that is the number of parameters that are well-determined by the data. The complexity measure, together with the evidence  $p(D|M)$ , provides a useful diagnostic tool for model selections. For example, when the evidences for two models do not lead to a clear preference for one of them, one might decide to base a decision on the value of the effective number of parameters.

In addition to exact measures, such as  $k_{\text{eff}}$ , there are three widely used approximate measures of complexity: the AIC, SIC [18], and DIC [26]:

- *Akaike Information Criterion (AIC)*

$$\text{AIC} \equiv -2 \ln p(D|\hat{\theta}, M) + 2k , \quad (12.52)$$

where  $D$  denotes the data points,  $k$  the number of parameters,  $\hat{\theta}$  the maximum likelihood estimate of  $\theta$ , and  $M$  the model to which the parameters pertain;

- *Schwarz Information Criterion (SIC)* (also known as the *Bayesian Information Criterion (BIC)*)

$$\text{SIC} \equiv -2 \ln p(D|\hat{\theta}, M) + k \ln N , \quad (12.53)$$

where  $N$  is the number of data points;

- *Deviance Information Criterion (DIC)*

$$\text{DIC} \equiv -2 \ln p(D|\hat{\theta}, M) - 2 \ln p(D|M) + 2k_{\text{eff}} , \quad (12.54)$$

where  $k_{\text{eff}}$  is the effective number of parameters.

**Pitfalls of model selection** Presumably, the goal of model selection in physics is to select one model amongst a set of competing models. But, as emphasised by Linder and Miquel [27], model selection cannot rely solely on statistical techniques. Physics insight is crucial. Statistics is a guide, but not one to be followed blindly.

Interestingly, this sage advice is already implicit in (12.3), the probability of hypothesis  $H_1$ ,  $P(H_1|D)$ , given the observation of  $D = N$  events. What (12.3) reminds us is that the probability of a model,  $P(M|D)$ , depends not only on the evidences  $p(D|M_i)$  for the models  $M_i$  under consideration but also on the prior probabilities  $P(M_i)$  assigned to them. Clearly, the latter is a matter of judgement, presumably formed by physics insight.

An analysis using Bayes factors may indicate that model  $M_1$  is preferred to model  $M_2$ . However, model  $M_2$  may be much more compelling physically than model  $M_1$ , and it may well be that subsequent observations prove that the correct choice was indeed to have ignored the Bayes factor in favour of  $M_1$ . At present, the standard model of cosmology remains, by far, the most compelling of the numerous models that exist in the literature, even if several give equally good fits to the available data.

Consider, for example, the ad hoc model of Dungan and Prosper [28] in which  $\Omega_M = 1$ ,  $\Omega_A = 0$  and  $\Omega_K = 0$ , and in which the comoving mass density, or equivalently the strength of gravity, varies like  $\exp[b(a - 1)]$ , where  $b$  is a dimensionless

parameter. The model exhibits a ‘Big Rip’ at

$$t(a \rightarrow \infty) = \frac{1}{H_0} \exp\left(\frac{b}{2}\right) \frac{\sqrt{2\pi/b}}{b}, \quad (12.55)$$

that is a scale factor  $a(t)$  that diverges to infinity in a finite amount of time. A maximum-likelihood fit to this model yields  $b = 2.16 \pm 0.06$ , which, for  $H_0 = 70 \text{ km/s/Mpc}$ , predicts that the Universe will meet its catastrophic end about 20 billion years from now. This ad hoc single-parameter model and the standard cosmological model both fit the Union2 compilation data [23] equally well, with  $\chi^2/\text{ndf} = 0.98$  for both. Therefore, on the basis of these data *alone* and the evidences  $p(D|M)$ , there would be no *statistical* basis to choose one model over the other. However, the standard cosmological model is physically much more compelling.

An excellent example of the importance of maintaining a measure of skepticism regarding statistical statements is provided by the Nobel prize-winning work of the Supernova Cosmology Project (SCP). In a paper published in 1997 [29], based on the analysis of data from seven high-redshift Type Ia supernovae, the SCP team concluded that

$$\Omega_A < 0.51 \quad \text{at 95% CL}.$$

However, 15 years later the same team along with others have amassed convincing evidence that  $\Omega_A \sim 0.70$  with impressive precision!

The point we are making is simply this: the job of a scientist is to do science. Statistics is a very powerful tool, which is surely indispensable but, like any powerful tool, it should be used with care informed by insight and common sense. In model selection, in particular, as we noted earlier, it is crucial to think carefully about the choice of priors and to make sure they are proper, a point not always appreciated in current practice. Bayesian methods are compellingly elegant. But we should beware of the siren call of elegance lest it hinders skeptical thought.

### 12.3 Nested Sampling

The popularity of Bayesian methods is undoubtedly due in part to the coherence and elegance of these methods and their broad applicability. However, it is also due in part to the development of powerful algorithms that permit the routine application of Bayesian methods to multi-parameter problems. Efficient implementations of methods such as Markov chain Monte Carlo (MCMC) [30] have revolutionised large-scale Bayesian calculations. More recently, nested sampling, introduced by Skilling [31] in 2004, was specifically designed to speed up the calculation of the Bayesian evidence  $p(D|M)$ . Nested sampling is the foundation of `MultiNest`, a Bayesian inference tool developed by Feroz, Hobson and Bridges [32]. In this section, we review the nested sampling algorithm.

Bayesian parameter estimation is based on the posterior density

$$p(\theta|D, M) = \frac{p(D|\theta, M)\pi(\theta|M)}{p(D|M)}, \quad (12.56)$$

where  $D$  denotes the data,  $M$  the model for which the parameters are being estimated, and  $\pi(\theta|M)$  the prior density. For parameter estimation, the normalisation factor  $p(D|M)$ , which is given by (12.49), is usually of no interest. However, as the previous section discusses, the evidence  $p(D|M)$  plays a pivotal role in Bayesian model selection. Nested sampling was designed to provide an efficient way to approximate  $p(D|M)$ .

Nested sampling is based on the observation that (12.49) – in general a multi-dimensional integral – can be transformed to a 1-dimensional integral using the prior-weighted volume  $X$ , defined by

$$dX \equiv \pi(\theta|M)d\theta \quad (12.57)$$

and

$$X = \int I[p(D|\theta, M) - \lambda]\pi(\theta|M)d\theta, \quad (12.58)$$

where the indicator function  $I[x]$  is one if  $x > 0$ , and zero otherwise.<sup>10)</sup> In other words,  $X$  is the prior-weighted volume of the region enclosed by the surface  $p(D|\theta, M) - \lambda = 0$ . Equation 12.49 can then be written as

$$p(D|M) = \int_0^1 p(D|X, M)dX, \quad (12.59)$$

assuming that the prior has been normalised to unity over a compact domain in the parameter space of  $\theta$ . Since this is a 1-dimensional integral, it can be approximated using many standard quadrature methods, which take the general form

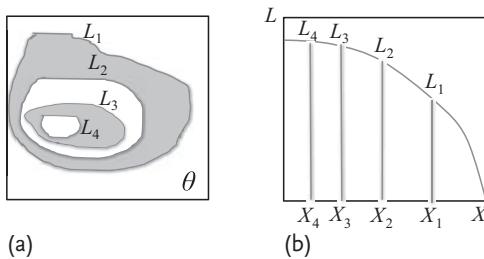
$$p(D|M) \approx \sum_{i=1}^M L(\theta_i)w_i, \quad (12.60)$$

where  $L(\theta) \equiv p(D|\theta, M)$  and  $w_i$  are the quadrature weights.

If we were handed a set of points sampled from the prior  $\pi(\theta|M)$ , we could compute the likelihood values  $L_i \equiv L(\theta_i)$  and sort them in decreasing order. Each likelihood value  $L_i$  is associated with a value  $X_i$  of  $X$ , which unfortunately is *unknown*. If, however, we knew the values of  $X$ , we could both plot the function  $L(X)$  and approximate its integral using (12.60). Figure 12.6 shows a cartoon of the relationship between the surfaces  $L(\theta) - \lambda = 0$ , which in two dimensions are contours, and the function  $L(X)$ , which is monotonically decreasing in  $X$ .

The observation that we can map the multi-dimensional evidence integral, (12.49), to a 1-dimensional integral is very clever. But for this to be useful, we need to solve the problem of finding the  $X$  values associated with each likelihood value.

<sup>10)</sup> The function  $I$  is also known as the Heaviside step function.



**Figure 12.6** (a) Iso-likelihood contours in a 2-dimensional parameter space.  $X_i$  is the prior-weighted volume enclosed by the  $i$ th contour. The value of the likelihood on the contour is  $L_i \equiv p(D|\theta_i, M)$ . (b) Plot of  $L(X)$ , which by construction is a monotonically *decreasing* function of  $X$ .

Nested sampling starts with a set of  $N$  points sampled from the prior,  $\pi(\theta|M)$ . All points are initially ‘active’. The point  $\theta'$  with the smallest likelihood  $L'$  is removed from the set of active points and placed in an initially empty ‘inactive’ set. Since  $L(X)$  is a monotonically decreasing function of  $X$ , this point will have the largest value of  $X$ , which we denote by  $X_{(N)}$ . By definition,  $X_{(N)}$  is an  $N$ th order statistic.<sup>11)</sup> Since we are sampling from the prior, the distribution of  $X$  will be uniform. Therefore,  $\Pr[X_{(N)} \in (t, t + dt)] = Nt^{N-1}dt$ . The mean of this distribution is  $N/(N+1)$ . And here is the key: we may take  $N/(N+1)$  as an *estimate* of the value of  $X' = X_{(N)}$  associated with  $L'$ . We therefore arrive at our first point  $(\theta', L', X')$  of the inactive set. A new point is sampled from the prior  $\pi(\theta|M)$ , subject to the constraint that  $L(\theta) > L'$ , so that we again have  $N$  points, but this time restricted to a smaller prior-weighted volume corresponding to the interval  $[0, X']$  rather than to the original interval  $[0, 1]$ . Since the sampling is now restricted to the interval  $[0, X']$ , the estimate of the new  $N$ th order statistic will be  $[N/(N+1)]^2$ . In general, for iteration  $i$ , the estimate will be  $[N/(N+1)]^i \approx \exp(-i/N)$ . The procedure is repeated until the entire space has been traversed. The sequence of points  $(\theta', L'(\theta), X')$  in the inactive set can be used in (12.60) to estimate  $p(D|M)$ . Moreover, by weighting each point  $\theta_i$  with  $L(\theta_i)w_i/p(D|M)$ , we can approximate any moment of the posterior density  $p(\theta|D, M)$  using these points. The pseudo-code for the algorithm is given below.

1.  $L' = -1$
2.  $\text{points} = \text{generate}(\pi(\theta|M), N, L')$
3.  $\text{active} = \text{computeLikelihood}(L(\theta), \text{points})$
4.  $\text{inactive} = \text{empty}()$
5. **for**  $i$  in  $1 \dots K$
6.      $L' = \text{minimumLikelihood}(\text{active})$
7.      $X' = \exp(-i/N)$

<sup>11)</sup> Given an ordered sequence of random variables  $X_{(1)}, X_{(2)}, \dots, X_{(k)}, \dots, X_{(N)}, X_{(k)}$  is called the  $k$ th order statistic. Since, by construction,  $X \sim \text{Uniform}(0,1)$ , the probability that exactly  $N-1$  values of  $X$  are smaller than  $t$  while the largest value lies in the interval  $(t, t+dt)$  is  $\propto t^{N-1}dt$ . When normalised, this becomes  $Nt^{N-1}dt$ .

```

8.    inactive.append((θ', L', X'))
9.    evidence = computeEvidence(inactive)
10.   if error(evidence) < small  break
11.   active.remove((θ', L'))
12.   point = generate(π(θ|M), 1, L')
13.   active.append(point)

```

As the algorithm proceeds up ‘Mount Likelihood’, it squeezes the same number of points into an ever smaller volume about the peak. The algorithm therefore places more points where they are needed most. The key to a successful implementation of this algorithm, such as MultiNest [32], is (1) having a fast generate() function, the job of which it is to generate one or more points from the prior under the restriction  $L(\theta) > L'$ , and (2) having an efficient way to locate the domain of each ‘mountain’ in a multi-modal landscape. Skilling’s clever idea has spawned an entire industry of nested samplers, which has added many new, and powerful, tools to our statistical analysis toolkit.

## 12.4

### Outlook and Conclusions

As this brief survey illustrates, astronomy is undergoing a revolution in its use of statistical methods that is largely driven by necessity: astronomical datasets are rapidly increasing in size, and in many cases the science lies in the collective properties of large collections of astronomical objects. The trend is certain to continue, as new observatories come online.

There was a time when astronomers actually peered through telescopes. Today, most peer at computer screens, much as their counterparts do in high energy physics. Another striking trend is the emergence of a new kind of astronomer, one who straddles the border between the disciplines of astronomy and statistics. Indeed, astrostatistics is emerging as a well-defined field that has already generated, and will surely continue to generate, ideas and methods that may be of interest to scientists in other fields. A good current example of broadly applicable methods that are being actively developed in astrostatistics is the set of methods called *Bayesian experimental design* (see e.g. [33]). This is being developed in the context of the search for extrasolar planets [34]. The basic idea is to apply Bayesian reasoning to the design of optimal adaptive planetary search strategies. In the context of high energy physics, this is analogous to the design of optimised search strategies for new physics. If the truth be told however, most so-called optimised analyses in high energy physics are, well, so-called! This is because analysis design in high energy physics is generally not formalised sufficiently to permit true optimisation. It will therefore be of interest to see how the ongoing astrostatistic efforts in Bayesian experimental design fare in practice and to assess whether they have something useful to teach us.

In this chapter, we tried to convey the diversity of applications of statistics in astronomy by reviewing a somewhat disparate set of examples, each of which has the virtue of containing ideas and methods that are directly applicable, or applicable with appropriate modifications, to high energy physics.

## 12.5 Exercises

For these exercises, use the DØ top discovery data [9]:  $D \equiv N = 17$ ,  $B = 40.1$  and  $\tau = 0.0947$ .

### Exercise 12.1 Calculating the likelihood; Bayes factor

Write a program to calculate  $p(D|s, H_1)$ , given in (12.11), and plot  $p(D|s, H_1)$  as a function of the expected signal  $s$ . Write a separate program to calculate  $p(D|H_0)$ . Check that  $p(D|H_0) = p(D|s = 0, H_1)$ . Verify the calculation of the Bayes factor, given in the Example 12.1.

### Exercise 12.2 Calculating the prior

Write a program to calculate the intrinsic prior  $\pi_I(s)$ , given in (12.20), and plot it as a function of  $s$ .

### Exercise 12.3 Comparison of Bayes factors

Using your programs, calculate the signal evidence  $p(D|H_1)$  using the intrinsic prior  $\pi_I(s)$ , then re-compute the Bayes factor  $B_{10}$ . Compare the new value with that calculated in Example 12.1. Study how the two calculations of  $B_{10}$  behave as a function of  $N$ .

You may wish to structure your programs as a class and use the `root` integrator class to perform the numerical integration over the expected signal  $s$ . A possible template can be found in file `integrator.h`.

## References

- 1 Lledo, T.J. (1994) The return of the prodigal: Bayesian inference in astrophysics, [www.astro.cornell.edu/staff/loredo/bayes/return.pdf](http://www.astro.cornell.edu/staff/loredo/bayes/return.pdf) (last accessed 1994).
- 2 Stigler, S.M. (2002) *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press.
- 3 Feigelson, E.D. and Babu, G.J. (eds) (2012) Statistical challenges in modern astronomy V, <http://astrostatistics.psu.edu> (last accessed 2012).
- 4 Wolff, S. The large synoptic survey telescope, [www.lsst.org/lst/](http://www.lsst.org/lst/) (last accessed 2013).

- 5 Li, T.P. and Ma, Y.Q. (1983) Analysis method for results in gamma-ray astronomy. *Astrophys. J.*, **272**, 313.
- 6 Linnemann, J.T. (2003) Measures of significance in HEP and astrophysics, in *Proc. Conf. Stat. Probl. Part. Phys. Astrophys. Cosmol.* (eds L. Lyons, R. Mount, and R. Reitmeyer), SLAC, PHYSTAT.
- 7 Cousins, R.D., Linnemann, J.T., and Tucker, J. (2008) Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. *Nucl. Instrum. Methods A*, **595**, 480.
- 8 Taraldsen, G. and Lindqvist, B.H. (2010) Improper priors are not improper. *Am. Stat.*, **64**, 154.
- 9 DØ Collab., Abachi, S. *et al.* (1995) Observation of the top quark. *Phys. Rev. Lett.*, **74**, 2632.
- 10 Berger, J. (2007) A comparison of testing methodologies, in *Proc. PHYS-TAT LHC Workshop Stat. Issues LHC Phys.* (eds H.B. Prosper, L. Lyons, and A. De Roeck), CERN, CERN-2008-001 in PHYSTAT, p. 8.
- 11 Demortier, L., Jain, S., and Prosper, H.B. (2010) Reference priors for high energy physics. *Phys. Rev. D*, **82**, 034002.
- 12 Prosper, H.B. and Lyons, L. (eds) (2011) Proc. PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, CERN-2011-006.
- 13 Puetter, R.C. and Amos, Y. (1999) The pixon method of image reconstruction, in *Astronomical Data Analysis Software and Systems VIII, ASP Conference Series*, vol. 172 (eds D.M. Mehringer, R.L. Plante, and D.A. Roberts), Astronomical Society of the Pacific, ASP, p. 307.
- 14 Gull, S.F. and Daniell, G.J. (1978) Image reconstruction with incomplete and noisy data. *Nature*, **272**, 686.
- 15 Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379.
- 16 Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 623.
- 17 Frieden, B.R. (1972) Restoring with maximum likelihood and maximum entropy. *J. Opt. Soc. Am.*, **62** (4), 511.
- 18 Trotta, R. (2008) Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemp. Phys.*, **49**, 71.
- 19 Riess, A.G. *et al.* (1998) Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.*, **116**, 1009.
- 20 Perlmutter, S. *et al.* (1999) Measurements of omega and lambda from 42 high-redshift supernovae. *Astrophys. J.*, **517**, 565.
- 21 Weinberg, S. (2008) *Cosmology*, 1st edn, Oxford University Press, New York.
- 22 Feldman, G. and Cousins, R.D. (1998) Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, **57**, 3873.
- 23 Amanullah, R. *et al.* (2010) Spectra and light curves of six type Ia supernovae at  $0.511 < z < 1.12$  and the union2 compilation. *Astrophys. J.*, **716**, 712.
- 24 Komatsu, E. *et al.* (2011) Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Cosmological interpretation. *Astrophys. J. Suppl.*, **192** (18), 1.
- 25 Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79.
- 26 Spiegelhalter, D.J. *et al.* (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, **64**, 583.
- 27 Linder, E.V. and Miquel, R. (2008) Cosmological model selection: Statistics and physics. *Int. J. Mod. Phys. D*, **17**, 2315.
- 28 Dungan, R. and Prosper, H.B. (2011) Varying-G Cosmology with type Ia supernovae. *Am. J. Phys.*, **79**, 57.
- 29 Perlmutter, S. *et al.* (1997) Measurements of the cosmological parameters  $\omega$  and  $\lambda$  from the first seven supernovae at  $z \geq 0.35$ . *Astrophys. J.*, **483**, 565.
- 30 Berg, B. (2004) *Markov Chain Monte Carlo Simulations and their Statistical Analysis*, 1st edn, World Scientific, Hackensack.
- 31 Skilling, J. (2004) Nested sampling, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*,

- AIP Conference Proceedings, vol. 735 (eds R. Fischer, R. Preuss, and U. v. Toussaint), AIP, p. 395.
- 32** Feroz, E., Hobson, M.P., and Bridges, M. (2009) Multinest: An efficient and robust Bayesian inference tool for cosmology and particle physics. *Mon. Not. R. Astron. Soc.*, **398**, 1601.
- 33** v. Toussaint, U. (2011) Bayesian inference in physics. *Rev. Mod. Phys.*, **83**, 943.
- 34** Loredo, T.J. *et al.* (2012) Bayesian methods for analysis and adaptive scheduling of exoplanet observations. *Stat. Methods*, **9**, 101.

## The Authors

**Roger Barlow** did his Ph.D. at Cambridge on one of the last bubble chamber experiments, and then worked on the DESY experiments TASSO and JADE, where he realised how much particle physics needed to understand statistics properly. He continued to develop statistical methods on the OPAL, BABAR and LHCb experiments. He worked at Manchester University for many years, where he taught several statistics courses and wrote an undergraduate textbook, before moving to his current position at Huddersfield University to start a new institute for accelerator physics.

**Olaf Behnke** is a staff physicist at DESY (Hamburg). He studied physics at the University of Hamburg, received his Ph.D. from ETH Zurich and habilitated at the University of Heidelberg. He has worked on the experiments CP-LEAR at CERN and on ARGUS, H1 and ZEUS at DESY. He currently holds the position of physics chair in ZEUS.

**Volker Blobel** studied physics in Brunswick and Hamburg, where he also obtained his Ph.D. in 1968. After post-doc positions at the University of Hamburg and at DESY, he became professor in Hamburg in 1977. He worked in several particle physics experiments (bubble chamber, various storage ring and neutrino experiments) at DESY and at CERN.

**Luc Demortier** studied physics at the University of Leuven (Belgium) and at Brandeis University in Massachusetts (USA). He collaborates on the CDF experiment at Fermilab and the CMS experiment at CERN, and served as member and chairman of the statistics committees of both collaborations. He is currently associate professor at the Rockefeller University in New York.

**Markus Diehl** studied physics in Göttingen, Paris, Heidelberg and Cambridge. He held post-doctoral positions in Palaiseau, Saclay, Hamburg, Stanford and Aachen and is currently a staff member in the theory group at DESY. His main interest is in Quantum Chromodynamics.

**Aart Heijboer** got his Ph.D. at the University of Amsterdam in 2004. He did a post-doc for the University of Pennsylvania at the CDF experiment, where he was responsible for combining the results from different decay channels into a single, precise, measurement of  $B_s$  oscillations. He also developed the method for and performed the evaluation of the statistical significance of this measure-

ment. This was followed by an analysis searching for the Higgs boson. He has received a research grant from the Netherlands Organisation for Scientific Research and holds a staff position at Nikhef since 2008, where he works on the analysis of data from the ANTARES neutrino telescope.

**Carsten Hensel** studied physics at the Universities of Münster and Hamburg. He held post-doc positions at DESY and the University of Kansas working for OPAL, ILC and DØ. He currently leads an Emmy Noether research group at the University of Göttingen working for ATLAS and DØ.

**Kevin Kröninger** studied at the Universities of Göttingen, Bonn and the North-eastern University, Boston. His Ph.D. work was on novel experimental techniques involving germanium detectors used in searches for neutrinoless double-beta decay with the GERDA experiment. He currently works at the University of Göttingen on the ATLAS experiment where he focuses his research on the measurement of top-quark properties.

**Benno List** studied physics at the Technical University of Berlin and the University of Hamburg, from which he received his Ph.D. in 1997. He held post-doc positions at CERN, ETH Zurich and in Hamburg and participated in the OPAL and H1 experiments. He is currently employed by DESY, where he works for the Global Design Effort of the International Linear Collider.

**Lorenzo Moneta** studied physics at the University of Pisa and Florence. He worked as a CERN research fellow in the ALEPH experiment and then as a post-doc at the University of Geneva in the CDF and ATLAS experiments. He has been a staff member at CERN since 2002, working on providing the common scientific software for the physics experiments. He is involved in the ROOT project with responsibility for the mathematical and statistical software for data analysis.

**Harrison B. Prosper** is a Distinguished Research Professor and the Kirby W. Kemper Professor of Physics at Florida State University, where he has taught since 1993. He holds a Ph.D. from the University of Manchester and is a Fellow of the American Physical Society. His main area of research is experimental high energy physics at hadron colliders and the development and application of advanced analysis methods.

**Thomas Schörner-Sadenius** studied physics at the Universities of Hamburg and Munich. He held post-doc positions in Munich, at CERN and in Hamburg, working on a number of different experiments (OPAL, H1, ATLAS, ZEUS, CMS). In 2008 he joined DESY (Hamburg) where he is currently the leader of the Analysis Centre of the German Helmholtz Alliance ‘Physics at the Terascale’.

**Grégory Schott** is a physicist employed at the Karlsruhe Institute of Technology (Germany). After completing his Ph.D. at CEA/Saclay (France) and a post-doc in the BABAR experiment, he joined the CMS experiment in 2007 where he has been working on Higgs searches. He looked at the possibility of combining the related measurements and worked on applying and comparing results of different statistical approaches. He is one of the authors of the ROOSTATS software package that is a general-purpose tool for statistical interpretation in data

analysis with various approaches used in high energy physics. He is currently a member of the statistics committee of the CMS collaboration.

**Ivo van Vulpen** received his Ph.D. at the University of Amsterdam after working on the DELPHI experiment at LEP (Higgs search and  $ZZ$  cross-section measurement). After his Ph.D. he worked as a CERN research fellow on the CMS experiment (ECAL testbeam) and as a post-doc at Nikhef working on the ATLAS experiment (top-quark physics). He received a personal grant from the Netherlands Organisation for Scientific Research to work on top-quark physics with his own research group and currently holds a staff position as a lecturer at the University of Amsterdam.

**Helge Voss** studied at the University of Bonn, obtaining his Ph.D. for the analysis of triple gauge-boson couplings at LEP with the OPAL collaboration. He has had post-doctoral positions at CERN, EPFL-Lausanne, Zurich and currently at the MPI-K in Heidelberg. Besides working on the design and construction of the LHCb silicon tracker he is a founding member of TMVA, a multivariate data analysis toolkit.

**Rainer Wanke** studied physics in Hamburg and Mainz, working for both the ARGUS and ALEPH experiments. He obtained his Ph.D. with one of the first observations of the time dependence in  $B^0\bar{B}^0$  mixing. As a post-doc he moved to Cornell University and CLEO, and then to the MPI Heidelberg, working for the HERA-B vertex detector. In 1999 he moved from  $B$  to  $K$  physics and now holds a staff position at the University of Mainz, where he leads the NA48/NA62 group and recently also joined the CALICE collaboration.

## Index

### **a**

Acceptance region 77  
 Activation function 169  
 Adaptive boost (AdaBoost) 179–180  
   – boost weight 180  
 Akaike information criterion (AIC) 400  
 Alternative hypothesis 76  
 Anderson–Darling test 101  
 Approximate Bayesian Computation (ABC) 382  
 Arithmetic mean 3  
 Armijo condition 253  
 Artificial neural network (ANN) 168–172  
   – activation function 169  
   – backpropagation 171  
   – batch learning 171  
   – feed-forward network 170  
   – learning rate 171  
   – multi-layer perceptron (MLP) 170  
   – neuron 169  
   – node 169  
   – online learning 171  
   – overtraining 172  
   – perceptron 170  
   – weight 169  
 Astrostatistics 382  
 Asymptotic distribution 83  
 Asymptotic likelihood-ratio test 130  
 Averaging data with inconsistencies 63–67  
   – outliers 65  
   – world average values 65–67  
 Averaging of measurements 32–33, 43–46, 63–67

### **b**

Backpropagation 171  
 Bagging 181  
   – bootstrapping 181  
 Bayes factor 92, 385

Bayes’ theorem 19, 67  
 Bayesian elimination 130  
 Bayesian experimental design 404  
 Bayesian inference 21–24  
 Bayesian information criterion (BIC) 400  
 Bayesian interval construction 133–140  
 Bayesian parameter estimation 67–69  
 Bayesian probability 18  
 Bayesian reference analysis approach 135  
 Bernoulli distribution 172  
 Bernoulli experiment 333  
 Best linear unbiased estimator (BLUE) 351  
 Bias  
   – experimenter’s bias 292–293, 354  
   – in classification 160, 166  
   – in hypothesis testing 81  
   – of an estimator 28  
 Bias-variance trade-off 159–161  
 Binned mass-peak fit: practical considerations 61–63  
 Binned maximum likelihood 59–60  
 Binning 272  
 Binomial distribution 9–10  
 Blind analysis 292–293, 351–354  
   – hidden signal box method 352  
 Blue noise 199  
 Boosted decision trees (BDT) 178–179  
   – boosting 179  
   – Gini index 179  
   – pruning 179  
 Boosting 179–181  
   – adaptive boost (AdaBoost) 179–181  
   – base classifiers 180  
   – base learner 180  
   – gradient boost 180  
 Bootstrap method (trigger) 332–333  
 Bootstrapping 110, 124–128, 346  
   – bootstrap sample 125  
   – bootstrap-*t* interval 125–127

- calibration 128
- percentile intervals 127–128
  - automatic percentile bootstrap 127
  - simple percentile interval 127
- Breit–Wigner distribution 10–11
  
- c**
- Cauchy distribution 10–11
- Central limit theorem (CLT) 6
- Central moments 3
- Characteristic function 5
- $\chi^2$  distribution 14–15, 37
  - number of degrees of freedom 15
- $\chi^2$  test 93–96
- Cholesky decomposition 197, 238, 258
- Classification 153–186
  - all-versus-all approach 153
  - one-versus-all approach 153
  - one-versus-one approach 153
  - pre-processing 182–183
  - Receiver-Operating-Characteristic (ROC) 157–158
  - systematic uncertainties 183–184
- Classifier 154, 162–181
  - artificial neural network (ANN) *see* Artificial neural network (ANN)
  - bias 160
  - boosted decision trees (BDT) *see* Boosted decision trees (BDT)
  - Fisher linear discriminant *see* Fisher linear discriminant
  - generalisation properties 160
  - k-Nearest Neighbour classifier (kNN) *see* k-Nearest Neighbour classifier (kNN)
  - naive Bayes classifier 162–163
  - support vector machine (SVM) *see* Support vector machine (SVM)
  - variance 160
- Clopper–Pearson interval 121
- Closure test 271, 347
- Complementary error function 7
- Composite hypothesis 76
- Confidence belt 111
- Confidence level 108, 363
- Consistency
  - in hypothesis testing 81
  - of an estimator 28
- Constrained fits 227–261, 376
  - Lagrange multipliers *see* Lagrange multipliers
  - propagation of uncertainty 239–244
  - solution by elimination 230–232
  - statistical interpretation 231–232
  
- Convolution 193–195
- Correlation 4
  - correlation of estimators 351
  - of systematic uncertainties 285–288
- Cost function 82
- Counting experiment 358–361
  - optimisation 360
- Counting method (trigger) 331
- Covariance 4
- Covariance matrix 4, 228
- Coverage 108, 374
  - overcoverage 121
  - undercoverage 116
- Credibility 108
- Critical region 77–79
- Cross-validation 161
- Cumulative distribution function 5, 120
- Curse of dimensionality 163
- Curtosis 3–4
- Cut variations 279–285
  
- d**
- Data blinding *see* Blind analysis
- Data mining 153
- Data-to-Monte Carlo comparison 277–278
- Deciles 5
- Decision boundary 157
- Deconvolution 193–195
- Degree-of-belief 18
- Dennis–Moré theorem 255
- Deviance information criterion (DIC) 400
- Directional derivative 251
- Discovery 365
- Discrete cosine transformation 193
- Discrete Fourier transform 210
- Distributions 5–16
- Double-blind test 352
  
- e**
- Educated guess 273–274
- Efficiency
  - of an estimator 28
- Eigenvalue decomposition 198
- Ensemble tests *see* Bootstrapping
- Error function 7
- Errors of first and second kind 79, 155–156
- Estimator 28
  - bias 28
  - consistency 28
  - efficiency 28
- Evidence 386
- Evidence-based prior 386
- Exclusion 365

- Expectation value 2
- Expected significance (level) 359
- Expected-posterior prior construction 388
- Experimenter's bias *see* Blind analysis
- Extended maximum likelihood 55–59, 370
  - fitting rates of processes 56–57
  - improving fitted parameters 57–59
- f**
- Factorisation 298–308
  - factorisation scale 300, 305–307
  - scope and limitations 300–301
- Feature vector 154
- Feed-forward network 170
- Feldman–Cousins intervals 123
- Fisher information 23, 38
  - information matrix 28
- Fisher linear discriminant 165–168
  - between-class matrix 167
  - bias 166
  - weight vector 166
  - within-class matrix 167
- Fletcher's augmented Lagrangian 255
- Flip-flopping 143
- Fredholm integral equation 187, 391
- Frequentist inference 20–21
- Frequentist interval construction 110–133
- Full width at half maximum (FWHM) 11
- g**
- Garwood intervals 123
- Gaussian distribution 6–8
- Gibbs phenomenon 203, 205
- Gini index 179
- Goldstein conditions 254
- Goodness-of-fit (GoF) test 84, 92–102
  - Kolmogorov–Smirnov test 99–101
  - maximum-likelihood-based test 98–99
  - Pearson's  $\chi^2$  test 93–96
  - run test 96–98
  - Smirnov–Cramér–von Mises test 101
  - two-sample tests 101–102
  - unbinned  $\chi^2$  test 98
- Gradient boost 180
- h**
- Hessian matrix 35, 239, 319–321
- Hidden signal box method 352
- Hodges–Lehmann estimator 109
- Hypothesis testing 75–104
  - Bayesian approach 92
  - bias 81
  - combination of tests 87–88
  - consistency 81
  - discovery 365
  - exclusion 365
  - frequentist approach 76
  - goodness-of-fit (GoF) test *see* Goodness-of-fit (GoF) test
  - hybrid method 86
  - look-elsewhere effect (LEE) *see* Look-elsewhere effect (LEE)
  - marginalization approach 86
  - power of a test 78
  - profile likelihood *see* Profile likelihood
  - size of a test 78
  - systematic uncertainties 86
  - test inversion 89–90
  - test statistic *see* Test statistic
- i**
- Image reconstruction 390–394
- Improper prior 22, 386
- Inference 2, 20–24, 381
  - Bayesian inference 21–24
  - frequentist inference 20–21
- Information *see* Fisher information
- Interval construction 108–110
  - asymptotic likelihood-ratio test 130
  - automatic percentile 131
  - Bayesian elimination 130
  - Bayesian interval construction 133–140
  - equal-tailed intervals 133
  - frequentist interval construction 110–133
  - generality 110
  - highest posterior density 109, 133
  - interval length 108
  - likelihood regions 134
  - likelihood-ratio bootstrap 131
  - likelihood-ratio profile bootstrap 131
  - likelihood-ratio test inversion 130
  - lowest posterior loss regions 134
  - Mandelkern–Schulz intervals 141
  - naive method 130
  - Neyman's construction 110–116
  - ordering rule 113–115
  - parameter transformation 109
  - physical boundaries 109
  - relation to point estimate 109
  - simple percentile 131
  - systematic uncertainties 109
- Interval estimation 107–146
  - Intrinsic discrepancy loss 134
  - Intrinsic prior construction 388
  - Inverse problem 187–196
    - direct processes 187–189

- discretisation 189–192
- inverse processes 187–189
- Iterative unfolding 213–215
  - implicit regularisation 213
  - Landweber iteration 214
  - Lucy–Richardson convolution 213
- j**
- Jacobian determinant 232
- Jacobian matrix 239, 258
- Jeffreys prior 23–24, 135, 386
- k**
- k-Nearest Neighbour classifier (kNN) 163–165
  - metric 165
  - smoothing 165
- Kernel density estimation 340
- Kernel function 188
- Kolmogorov axioms 17
- Kolmogorov–Smirnov test 99–101
- Kullback–Leibler divergence 134, 399
- Kurtosis 3–4
- l**
- Lagrange multipliers 176, 232–237, 321
  - feasible direction 233
  - feasible points 233
  - feasible set 233
  - iterative solution 244–259
    - choice of direction 245–250
    - convergence 256–258
    - initial values 258–259
    - propagation of uncertainty 259
    - step length 250–256
  - Lagrange function 232
  - linear independence constraint qualification (LICQ) 234, 258
  - unmeasured parameters 236–237
- Landau distribution 11–12
- Landweber iteration 214
- Least squares 40–52
  - binned fit 60–61
  - design matrix 42
  - least-squares estimate 40
    - variance 41
  - linear least squares 42–48
    - averaging of correlated measurements 44–46
    - averaging of measurements 43–44
    - best linear unbiased estimator (BLUE) 42
    - normal equations 42
    - straight-line fit 46–48
- variance 42
- non-linear least squares 48–52
  - mass-peak fit (signal position) 50–52
- Leptokurtic 4
- Likelihood 20
- Likelihood principle 20
- Likelihood-free analysis 382
- Likelihood-ratio bootstrap 131
- Likelihood-ratio profile bootstrap 131
- Likelihood-ratio test inversion 130
- Line search 252–253
  - Armijo condition 253
  - cubic interpolation 254
  - Goldstein conditions 254
  - sufficient-decrease condition 253
  - Wolfe conditions 253
- Log-normal distribution 15
- Look-elsewhere effect (LEE) 52, 88–89, 361
- Lorentzian distribution 10–11
- Loss function 159
- Lucy–Richardson convolution 213
- m**
- Machine learning 159
- Mahalanobis distance 165
- Mandelkern–Schulz intervals 141
- Manhattan norm 251
- Maratos effect 255–256
- Marginal distribution 4, 339
- Markov chain Monte Carlo (MCMC) 401
- Mass-peak fit 50–52, 61–63
- Matrix
  - Cholesky decomposition 197, 238, 258
  - column rank 239
  - condition of a matrix 197
  - correlation matrix 4
  - covariance matrix 4, 228
  - design matrix 42
  - effective rank of a matrix 200
  - Fisher information matrix 28
  - Hessian matrix 35, 239, 319–321
  - Jacobian matrix 239, 258
  - Moore–Penrose pseudo-inverse 191, 248–249, 256
  - resolution matrix 203
  - response matrix 189, 217–218
  - unfolding matrix 215
- Matrix method 334–337
- Maximum entropy method 391, 393–394
- Maximum-likelihood method 29–40
  - averaging of measurements 32–33
  - binned maximum likelihood 59–60
  - extended maximum likelihood 55–59, 370

- maximum-likelihood estimate  
(MLE) 29
  - properties 31
  - transformation invariance 31
  - variance 33–35
- profile likelihood *see* Profile likelihood
- solution 30–31
- unbinned maximum likelihood 52
- M**
  - Mean 3
  - Median 5
- Minimum-variance bound (MVB) 28, 38–40
- Mode 5
- Model complexity 399–400
- Moments 3
- Monte Carlo generators 313–314
- Moore–Penrose pseudo-inverse 191, 248–249, 256
- Morozov discrepancy principle 207
- Multi-layer perceptron (MLP) 170
- MultiNest 404
- Multinomial distribution 10, 338
- Multivariate analysis 381
- Multivariate classification 153
- n**
  - Naive Bayes classifier 162–163
  - Negative binomial distribution 12
  - Nested sampling 401–404
  - Neuron 169
  - Newton–Raphson method 48, 246
  - Neyman–Pearson lemma 155, 158–159
  - Neyman’s  $\chi^2$  344
  - Neyman’s construction 110–116
  - Normal distribution 6
  - Nuisance parameters 86, 128–133, 372
  - Null hypothesis 75, 155
- o**
  - Objective function 228
  - Objective prior 24, 69, 108
  - Observed significance (level) 83, 359
  - On/off problem 383–390
  - Overtraining 160, 172
- p**
  - p*-value 7, 78, 83–89, 358
  - Parameter estimation 27–72
    - Bayesian approach 67–69
    - least squares *see* Least squares
    - maximum-likelihood method *see* Maximum-likelihood method
  - Parametrised unfolding 195–196
  - Parton distribution functions 299, 314–324
    - comparison of PDF sets 322–324
- DGLAP evolution equation 315
- NNPDF approach 321–322
- parametric uncertainties 318–322
- uncertainties 316–322
- Pattern recognition 153
- Pearson’s  $\chi^2$  344
- Pearson’s  $\chi^2$  test 93–96
- Penalty function  $\ell_1$  251
- Percentiles 5
- Perceptron 170
- Philipps regularisation 205
- Pivoting 118–123
- Pixon method 391
- Platykurtic 4
- Plug-in principle 124
- Point spread function (PSF) 188, 213, 390
- Poisson distribution 8–9
- Posterior probability 19, 338
  - marginal distribution 339
- Power of a test 78
- Power-constrained limits 90, 366
- Prediction 2
  - retrospective prediction 15
- Principal component analysis (PCA) 182
- Prior probability 19, 22–24, 69, 338
  - evidence-based prior 386
  - expected-posterior prior construction 388
  - improper prior 22, 386
  - intrinsic prior construction 388
  - Jeffreys prior 23–24, 135, 386
  - objective prior 24, 69, 108
  - proper prior 108, 386
  - reference prior 24
  - subjective prior 386
  - uniform prior 22
- Probability 2, 16–19
  - Bayesian definition 18
  - classical definition 17
  - frequentist definition 17–18
  - mathematical definition 17
- Probability density function 2–5
  - marginalisation 4
  - projection 4
- Profile likelihood 37–38, 362–369, 372
  - profile likelihood ratio 362
- Projection 4
- Projection methods
  - discrete cosine transformation 210–211
- Proper prior 108, 386
- Pruning 179

- Pseudo-experiment 345  
 Pull distribution 12, 347, 374
- q**  
 QR decomposition 248  
 Quantiles 5  
 Quartiles 5
- r**  
 Random process 2  
 Random variable 2, 5  
 Receiver-Operating-Characteristic (ROC) 157–158  
 Rectangular distribution 10  
 Reference prior 24  
 Regularisation 203–209
  - derivative regularisation 206–207
  - low-pass regularisation 212–213
  - norm regularisation 204–206
  - Philipps regularisation 205
  - presentation of results 220–221
  - regularisation methods 203
  - regularisation parameter 203, 207–208
  - regularisation schemes 203
  - Tikhonov regularisation 205
 Resampling 124, 346  
 Response function 188  
 Response matrix 189, 217–218  
 ROOT 222  
 Root mean square 257  
 Run test 96–98
- s**  
 Schwarz information criterion (SIC) 400  
 Semi-convergence 213  
 Sideband fit 360–361, 383  
 Sideband subtraction 274  
 Significance (level) 78, 84–85
  - expected significance (level) 359
  - observed significance (level) 83, 359
 Significance tests 84  
 Simple hypothesis 76  
 Single-blind test 352  
 Singular value decomposition (SVD) 196–198, 258
  - Fourier coefficients 199
  - pre-scaling 197
  - pre-whitening 197
  - singular values 197
  - thin SVD 197
  - truncated SVD 199–202
 Size of a test 78  
 Skew 3–4  
 Smirnov–Cramér–von Mises test 101
- Software
  - APLCON 260
  - BAT 69, 337
  - GURU 222
  - MINUIT 31, 38, 124, 195
  - R 7
  - ROOFIT 357, 378
  - ROOSTATS 69, 357
  - ROOT 7, 37
  - ROOUNFOLD 222
  - RUN 217
  - TUNFOLD 222
  - for constrained fits 260
  - for unfolding 221–222
 Standard deviation 3  
 Standard normal distribution 6  
 Statistical hypothesis 75–76
  - alternative hypothesis 76
  - composite hypothesis 76
  - null hypothesis 75
  - simple hypothesis 76
 Statistical inference 381  
 Straight-line fit 46–48  
 Student’s *t* distribution 12–14  
 Subjective prior 386  
 Subjective probability 18  
 Supervised machine learning 159  
 Support vector machine (SVM) 172–178
  - margin 173
  - soft-margin approach 174
  - support vector 173
 Systematic uncertainties 263–295, 375
  - binning 272
  - combination 285–288
  - correlation 285–288
  - cut variations 279–285
  - data-to-Monte Carlo comparison 277–278
  - detector acceptance 268–269
  - educated guesses 273–274
  - estimation 272–288
  - experimenter’s bias 292–293
  - in hypothesis testing 86
  - in interval construction 109
  - sources 265
  - theory uncertainties *see* Theory uncertainties
  - tolerance 273

**t**  
 Tag-and-probe method 331–332  
 Taxicab norm 251  
 Template method 337–345  
 Template morphing 342, 371

- Test inversion 89–90, 116–117
  - likelihood-ratio test inversion 130
- Test statistic 76–77, 155, 362
  - sufficiency 80
- Theory uncertainties 276–277, 297–325
  - colour reconnection 314
  - hadronisation 311–313
    - corrections 312
    - fragmentation functions 312
  - multi-parton interactions 310–311
  - parton distribution functions 316–322
  - perturbative expansion 301–308
    - combining different orders 307
    - factorisation scale 305–307
    - multi-scale problems 307–308
    - renormalisation scale 301–305
    - resummation methods 307–308
  - power corrections 300, 308–310
    - higher-twist corrections 308
    - operator product expansion 308–309
  - sources 297
  - underlying event 310–311
- Tikhonov regularisation 205
- Times series analysis 382
- Tolerance 273
- Top-hat distribution 10
- Transfer function 213
- Two-sample tests 101–102
- Type I and type II errors 79, 155–156
  
- u**
- Unbinned  $\chi^2$  test 98
- Unbinned maximum likelihood 52–59, 370–371
  - fitting fractions of processes 54–55
- Unfolding 187–223, 391
  - bin migration 188
  - bin-by-bin correction method 208, 214
  
- v**
- Variance 3
  - in classification 160
  
- w**
- Weibull distribution 15–16
- White noise 211
- Wilks' theorem 123
- Wolfe conditions 253
  
- z**
- Z-value 78, 84