Springer Series in Astrostatistics

Stefano Andreon
Brian Weaver

Bayesian Methods for the Physical Sciences

Learning from Examples in Astronomy and Physics

Springer

Springer Series in Astrostatistics

Springer Series in Astrostatistics,
More information about this series at http://www.springer.com/series/1432

Springer Series in Astrostatistics

Astrostatistical Challenges for the New Astronomy: *ed. Joseph M. Hilbe*

Astrostatistics and Data Mining: *ed. Luis Manuel Sarro, Laurent Eyer, William O'Mullane, Joris De Ridder*

Statistical Methods for Astronomical Data Analysis: *by Asis Kumar Chattopadhyay & Tanuka Chattopadhyay*

Stefano Andreon • Brian Weaver

Bayesian Methods for the Physical Sciences

Learning from Examples in Astronomy and Physics

Springer

Stefano Andreon
INAF
Osservatorio Astronomico di Brera
Milano, Italy

Brian Weaver
Statistical Sciences
Los Alamos National Laboratory
Los Alamos, NM, USA

ISSN 2199-1030 ISSN 2199-1049 (electronic)
Springer Series in Astrostatistics
ISBN 978-3-319-15286-8 ISBN 978-3-319-15287-5 (eBook)
DOI 10.1007/978-3-319-15287-5

Library of Congress Control Number: 2015932855

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This book is a consultant's guide for the researcher in astronomy or physics who is willing to analyze his (or her) own data by offering him (or her) a statistical background, some numerical advice, and a large number of examples of statistical analyses of real-world problems, many from the experiences of the first author. While writing this book, we placed ourselves in the role of the researcher willing to analyze his/her own data but lacking a practical way to do it (in fact one of us was this once). For this reason, we choose the JAGS symbolic language to perform the Bayesian fitting, allowing us (the researchers) to focus on the research applications, not on programming (coding) issues. By using JAGS, it is easy for the researcher to take one of the many applications presented in this book and morph them into a form that is relevant to his needs. More than 50 examples are listed and are intended to be the starting points for the researchers, so they can develop the confidence to solve their own problem. All examples are illustrated with (more than 100) figures showing the data, the fitted model, and its uncertainty in a way that researchers in astronomy and physics are used to seeing them. All models and datasets used in this book are made available at the site http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/.

The topics in this book include analyzing measurements subject to errors of different amplitudes and with a larger-than-expected scatter, dealing with upper limits and with selection effects, modeling of two populations (one we are interested in and a nuisance population), regression (fitting) models with and without the above "complications" and population gradients, and how to predict the value of an unavailable quantity by exploiting the existence of a trend with another, available, quantity. The book also provides some advice to answer the following questions: Is the considered model at odds with the fitted data? Furthermore, are the results constrained by the data or unduly affected by assumptions? These are the "bread and butter" activities of researchers in physics and astronomy and also to all those analyzing and interpreting datasets in other branches as far as they are confronted with similar problems (measurement errors, variety in the studied population, etc.).

The book also serves as a guide for (and has been shaped thanks to) Ph.D. students in the physical and space sciences: by dedicating 12 h of studying the content of this book, the average student is able to write *alone and unsupervised* the complex models in Chap. 8 (and indeed the example in Sect. 8.2 was written by a student).

Content
The book consists of ten chapters. After a short description on how to use the book (Chap. 1), Chaps. 2 and 3 are a gentle introduction to the theory behind Bayesian methods and how to practically compute the target numbers (e.g., parameter estimates of the model). In later chapters, we present problems of increasing difficulty with an emphasis on astronomy examples. Each application in the book has the same structure: it presents the problem, provides the data (listed in the text, or in the case of large datasets, a URL is provided) or shows how to generate the data, has a detailed analysis including quantitative summaries and figures (more than 100 figures), and provides the code used for the analysis (using again the open-source JAGS program). The idea is that the reader has the capability to fully reproduce our analysis. The provided JAGS code (of which the book contain over 50 examples) can be the starting point for a researcher applying Bayesian methods to different problems. The key subject of model checking and sensitivity analysis is addressed in Chap. 9. Lastly, Chap. 10 provides some comparisons of the Bayesian approach to older ways of doing statistics (i.e., frequentist methods) in astronomy and physics. Most chapters conclude with a set of exercises. An exercise solution book is available at the same URL given above.

Contact Details
Send comments and errors found to either stefano.andreon@brera.inaf.it or theguz@lanl.gov. Updated programs and an erratum are kept at http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/.

Acknowledgments
SA thanks Merrilee Hurn for her collaborations that inspired much of the content of this book, Steve Gull for enjoyable discussions on this book's content and for the use of his office in Cambridge (UK), Joe Hilbe for encouragements to continue the effort of writing this book, Andrew Gelman for his reactiveness in answering questions, Andrea Benaglia for the example in Sect. 8.2, and Giulio D'Agostini for computing the analytical expression of the posterior probability distribution for a few problems of astronomical relevance. The practical way to summarize our knowledge about the quantity under study, more than philosophical debates, is maximally useful to researchers.

BW thanks William Meeker and Max Morris for teaching him how to be a practical statistician. He also thanks Michael Hamada, Alyson Wilson, and Harry Martz for taking a young frequentist and opening his eyes to the beauty of Bayesian statistics, and David Higdon for teaching him how to have fun being a statistician.

Lastly, we thank our beautiful wives for their encouragement and support.

Finally, a warm thanks goes to Martin Plummer for JAGS, i.e., the software used to perform the numerical computations of this book. The ease of use of JAGS made our daily research work easier and allows us to easily explore other possible models and the sensitivity of our conclusions to (unavoidable) assumptions.

Contents

Chapter 1
Recipes for a Good Statistical Analysis

or "How to use this book"

Performing a sound analysis of a data set is never a simple endeavor. There are, however, a few well-known steps that are necessary for any statistical analysis. We present them below.

Learn the theory. A bit of background about statistical methods is useful when analyzing complex data sets, even when the data set looks simple at first sight. Some probability theory is described in Chap. 2. It is also useful to learn how the computer will compute (for you) the set of numbers (i.e., statistics) that you want (Chap. 3), especially if these have an influence on your career or reputation.

Start simple. Start with simple examples, in particular the one in Sect. 4.1, where a step-by-step guide provides the basics of building a Bayesian model. If you are unfamiliar with the concept of a prior, then read the first part of Chap. 5.

Practice. Next develop some experience by practicing with models of medium complexity: try to reproduce, by yourself, one of the many examples given in the book. During your study period, pay attention to possible coding mistakes: it may occur that the model you fit (i.e., the one you wrote) is not the one you intended to, much like when you write a program in a program language that you are not familiar with. If you are in doubt about your fitted model, simulate data from the model (as explained in Sect. 9.2) and try to recover the input parameters of your model. If you do not recover the input parameters, then most likely the fitted model is not the model intended to be fitted (the one used to generate the data). In fact, in a Bayesian framework, bias (a systematic difference between input and fitted values) does not exist.

Build the model and check that it adequately describes the data. Search this book for examples similar to the problem you are interested in, or that may be the building blocks of your (likely complex) model. Modify your initial (simple) model until it matches the model you want to fit. This is one of the main themes of the book, we almost always start simple and then add complexity (e.g., intrinsic scatter, heteroscedastic errors, a complex data structure, contamination by

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_1

something unwanted like a background source, a selection function, incompatible data, etc.). To check if the model adequately describes the data, generate simulated data from the model, and check that they look like the real data (how to do this is explained in Sect. 9.2). If there is a clear discrepancy between the real and the simulated data, the fitted model is inadequate for the data being fitted. You should then modify the fitted model, and iterate until adequacy is demonstrated.

Fit. You may now fit the model to the real data and get the (posterior) probability distribution of the parameters. As you (almost certainly) are using a computer to make the computations, it is recommended to check if the performed numerical computations are miscalculated (Chap. 3).

Sensitivity analysis & Model checking. Are the data informative about the quantity you want to measure or are the results sensibly dependent on details of the fitted model? Furthermore, is your prior uncertain, i.e., it is not too clear which exact form of a prior you should use? Regardless, if you have not already checked the various parts of the model, this is the right time to do so, because very soon you will publish your results. If you have not assessed your model, we urge you to look at Sect. 9.2.

Publish & Distribute. Do not forget to state (and possibly show) the adopted prior (perhaps on the same figure where you show the posterior), and, possibly, distribute the specific form of the model (i.e., the code), the data, and every other additional quantity you may have used, in such a way that your analysis is reproducible and can be used as a starting point for more complex analyses involving other data.

If you have used this book, you are taking the Bayesian approach. At some moment in your life, you may also want to be familiar with “the other half of the sky,” i.e., what non-Bayesian approaches use, what do their estimated quantities represent, how does one interpret these numbers, and what these numbers are not. Chapter 10 gives a very brief, and largely incomplete, view about these methods.

Chapter 2
A Bit of Theory

The purpose of this chapter is to provide the very basics of the probability theory needed for Bayesian statistics. If the reader is interested in a more extended and thorough description of probability foundations, we recommend Casella and Berger (2001) and Athreya and Lahiri (2006).

When we informally talk about probabilities (e.g., "there is a 10 % probability that it will rain tomorrow" or "Earth has a 3 % probability of having a catastrophic impact with an asteroid in the next 20 years") our own concept of probability satisfies three rules (formally known as Kolmogorov's axioms of probability) listed and described below.

2.1 Axiom 1: Probabilities Are in the Range Zero to One

Probabilities are bounded between 0 and 1. In formulae:

$$0 \leq p(E) \leq 1 \tag{2.1}$$

for some object (event, proposition, set, or whatever) E. Equation (2.1) is the first axiom of probability. If someone were to tell you that a physical parameter has 130 % probability (i.e., 1.3) or -30 % (i.e., -0.3) to be in a given range, then you must not trust this person, because you know probabilities are bounded to be between 0 and 1 (or 0 % and 100 %).

2.2 Axiom 2: When a Probability Is Either Zero or One

If a die is thrown, the probability of observing any of its faces is one: one of the die's faces must be viewed. If all possible outcomes are included in the set Ω (i.e., the set that includes everything, known as the sample space), then one is 100 % sure

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_2

Table 2.1 Likelihoods of drawing different objects from a bag (bag content)

	Ball	Die	Total
Blue	4/22	5/22	9/22
Red	10/22	3/22	13/22
Total	14/22	8/22	22/22

(i.e., with probability one) of observing the set Ω. In formulae:

$$p(\Omega) = 1 \quad . \tag{2.2}$$

Equation (2.2) is the second axiom of probability. Informally, this axiom ensures that the probability that we will observe anything (i.e., we will observe one of the faces of the die) is one.

For example, suppose we are talking about the physical distance of a source. The allowed values of this distance are between zero and infinity. So, intuitively, the second axiom of probability states that the probability that the distance to the source is between zero and infinity is one.

From the second axiom, it immediately follows that

$$p(\emptyset) = 0 \quad , \tag{2.3}$$

where $\emptyset$ is the symbol for "nothing" and is called the empty set. For example, by Eq. (2.3), the probability of throwing a die and not observing a face is zero.

This axiom states that for the whole (empty) set of possible outcomes, the probability is one (zero).

2.3 Axiom 3: The Sum, or Marginalization, Axiom

Suppose that in a bag I have 4 blue balls, 10 red balls, 5 blue dice, and 3 red dice. What is the probability of extracting something blue?

Quite naturally, it is given by the probability of getting a blue ball, $4/22$, plus the probability of getting a blue die, $5/22$, i.e., $4/22 + 5/22 = 9/22$. This simple example illustrates the next axiom of probability known as the sum axiom.

In mathematical terms, the sum axiom states

$$p(x) = \sum_y p(x,y). \tag{2.4}$$

In our example, y takes the values "ball" and "die" and x takes the values "red" or "blue," and the sum has to be performed on y. This rule is called marginalization because the sum was written, once upon a time, on the margin of the table on a piece of paper, as in Table 2.1.

Equation (2.4) above holds for categorical variables, i.e., when outcomes are a discrete set (like the letter of the alphabet or the integers between 1 and 100). Often, quantities are real numbers (sometimes bounded in a range), and the summation has to be replaced by an integral:

$$p(x) = \int p(x,y)dy \quad . \tag{2.5}$$

This is the third, and last, axiom of probabilities. The mathematical purist may disagree with this axiom and we refer him/her to Appendix B.1.

In addition to these basic axioms, there is another definition in probability that needs to be presented before we get to Bayes theorem which is the fundamental result needed for Bayesian statistics.

2.4 Product Rule

Let us return to another bag example. Suppose we emptied our bag and put in it 4 blue balls and 10 red balls. If I extract two balls without replacement (i.e., when I draw my first ball, I note the color and do not return it to the bag), what is the probability that the first ball is red and the other one is blue?

The probability of getting a red ball first is $p(x = \text{first ball is red}) = 10/14$. The probability of getting a blue ball second, after having drawn a red ball in the first extraction, $p(y = \text{blue ball second } | x = \text{first ball is red})$,[1] is $4/13$ (13 balls are left and 4 of them are blue). The probability of getting a red ball first and a blue ball second, $p(x,y)$, is the product of the two numbers above: $p(x) \times p(y|x) = 10/14 \times 4/13 \approx 0.22$.

Of course, the order of extraction does not matter, we may have started by considering the probability of getting a blue ball first, $p(y) = 4/14$, followed by the probability of getting a red ball after having acquired a blue one, $p(x|y) = 10/13$. The aimed probability is therefore $p(y) \times p(x|y) = 4/14 \times 10/13 \approx 0.22$, the same as before! This result is formalized in a single line:

$$p(x,y) = p(x|y) \times p(y) = p(y|x) \times p(x) \quad . \tag{2.6}$$

We now have enough probability foundations to present Bayes theorem.

2.5 Bayes Theorem[2]

A simple re-arrangement of Eq. (2.6) gives:

[1] The notation $p(y|x)$ reads "probability of y given x." See Appendix B.2 for the mathematical definition of conditional probability.

[2] Some of the material of this section has been drawn from Andreon and Hurn (2010).

$$p(x|y) = \frac{p(y|x) \times p(x)}{p(y)}. \tag{2.7}$$

This famous equation is known as Bayes theorem (or Bayes rule). We can freely change the name of the variables (since the names are arbitrary) and rewrite the above equation as:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \quad . \tag{2.8}$$

Each term in this expression is important enough to have a name:

$$Posterior = \frac{Likelihood \times prior}{evidence} \quad . \tag{2.9}$$

This rule is the central tool for parameter estimation in Bayesian statistics. Let us understand it with a physical example.

Suppose one is interested in estimating the mass of a particle, M, say the Higgs boson.[3] Before we collect any data, we may have certain beliefs and expectations about the values of M. In fact, these thoughts are often used in deciding which instrument will be used to gather data and how this instrument may be configured and operated. For example, one may want to construct an instrument with a large sensitivity (as the ATLAS and CMS experiments at the Large Hadron Collider) in the range where one expects to detect the Higgs: from 110 to 160 GeV/c^2. The same applies to astronomical objects, for example if we are measuring the mass, via the velocity dispersion, of a group of galaxies, it is preferable to select a spectroscopic resolution high enough to measure galaxy velocities with errors smaller than the scatter (velocity dispersion) we expect to measure. Crystallizing these thoughts in the form of a probability distribution for M provides the prior $p(M)$. Notice that we are describing our uncertainty in M, before observing any data, using a probability distribution, and we are not stating that the parameter (say, the mass of the Higgs boson) is randomly fluctuating.

For example, one may believe (i.e., from previous experiments) that the mass of the Higgs boson is most likely between 110 and 160 GeV/c^2 and this might be modeled by saying that the probability distribution of the mass is a Gaussian distribution centered on the midpoint, 135, and with σ, the standard deviation, equal to 12.5. A useful way to report that is by writing $M \sim \mathscr{N}(135, 12.5^2)$, where the symbol $\sim$ reads "is distributed as" and $\mathscr{N}$ denotes the Gaussian distribution.

Once the appropriate instrument has been selected (or perhaps built) and configured, data can be collected on the quantities of interest. In our example, this means we record a measurement of mass, say M^{obs}. The physics, or sometimes simulations, of the measurement process may allow us to estimate the reliability of such measurements. Repeated measurements are also extremely useful for assessing the quality of the measurements. Costly experiments, however, rarely can be repeated numerous times. The likelihood is the model which we adopt for how the noisy observation M^{obs} arises given a value of M. For example, we may find that the

[3] While the example is real, we have idealized it in several aspects.

measurement technique allows us to measure masses in an unbiased way but with a measurement error of 0.4 GeV/c^2 and that the error structure is Gaussian, i.e., $M^{obs} \sim \mathscr{N}(M, 0.4^2)$. If we observe $M^{obs} = 125.3$ GeV/c^2, we usually summarize the above by writing $M = 125.3 \pm 0.4$ GeV/c^2.

The likelihood ("instrument calibration," i.e., the probability distribution of the noisy observation we get when a given value is injected) is certainly interesting, but we may perhaps want to know which values of M are more probable after having observed M^{obs}, i.e., derive the probability of M given M^{obs}, $p(M|M^{obs})$. After all, even the most fanatic instrumental physicist should agree that spending billions of euros to build an instrument just to know how well we are able to calibrate it at ≈ 125 GeV/c^2 is not one of the most pressing research priorities.

Bayes theorem given in Eq. (2.8) allows us to update our beliefs about the unobserved mass M in light of the observed measurement, M^{obs}:

$$p(M \mid M^{obs}) = \frac{p(M^{obs} \mid M)p(M)}{p(M^{obs})} \ . \tag{2.10}$$

Simple algebra[4] shows that in our example the posterior distribution of $M \mid M^{obs}$ is Gaussian, with mean $\overline{M} = \frac{125.3/0.4^2 + 135/12.5^2}{1/0.4^2 + 1/12.5^2} \approx 125.3$ and $\sigma_{\overline{M}} = \sqrt{\frac{1}{1/0.4^2 + 1/12.5^2}}$ ≈ 0.4. $\overline{M}$ is just the usual weighted average of two "input" values, the prior and the observation, with weights given by the prior and the observation σ's. In this example, the posterior is numerically indistinguishable from the likelihood, because of the precise measurement and the quite broad prior (see Chap. 5 for more details on that), but this is not necessarily the rule.

In summary, the posterior quantifies what we know about a parameter after observing the data. If the posterior is sharply peaked on a value, then the parameter is well measured. If instead the posterior is a broad function, then the parameter is poorly measured, i.e., uncertain. Uncertainty can therefore be quantified by a measure of the posterior's width, such as standard deviation, 68 % interval, interquartile range, etc.

2.6 Error Propagation

There are many situations where the researcher is not interested in the probability distribution of a quantity, $p(x)$, but in a transformation of that quantity, for example $p(\log x)$. Similarly, we may have the error on flux and we want to convert it to error on mag, the latter is defined to be $-2.5 \log_{10} flux + c$.

If we have random draws x_i from the distribution $p(x)$, as will be always the case in this book (see next chapter), whose width (e.g., 68 % interval, interquartile range) gives the uncertainty of x, it is very simple to propagate this uncertainty on

[4] This is one of the very few analytic computation of this book. Later on, a computer will be charged with the burden of performing the computations. Why else call it computer?

derived quantities, say, $y = f(x)$: to obtain draws from the distribution of f_Y it is enough to compute $y_i = f(x_i)$ for each draw x_i (e.g., take the log of each draw x_i if we want $p(\log x)$). In this way, to compute the uncertainty of f_Y, it is enough to compute the spread of the y_i draws. The above holds true for any f, including non-linear functions (imagine the $\sin x$ function with $0 < x < 100$). The commonly used formula using the function's derivative only holds if it is linear in a region of size comparable to the spread in x.

2.7 Bringing It All Home

To perform Bayesian inference you only need to remember that inference passes through Bayes theorem (you do not need to remember this formula, other resources can remember it for you), and the marginalization rule.

An obvious consequence is that, in general, any other rule is allowed (which is good, there is nothing more to learn). A second consequence is that if you are using something different from one of the three probability rules, then the derived quantity will not behave as a probability, even if you baptize it with that name, as illustrated in the next section. Finally, if you use the Bayes theorem on things that are not probabilities, you will not obtain the probabilities you are looking for, precisely as if you use couscous in place of spaghetti, you are not preparing spaghetti alla "carbonara"!

2.8 Profiling Is Not Marginalization[5]

The sum rule (Eqs. 2.4 and 2.5) is at odds with the frequently adopted methods of computing errors on x when there is a further variable y. Let us illustrate this point with an example.

Consider X-ray data collected from one of the most distant galaxy clusters known (at the time of the writing of this book). Andreon et al. (2009) fit a model with two free parameters to the spectrum of the cluster. The two free parameters are the absorption n_H and the cluster temperature T. We calculate probabilities on a grid of 34×34 of (n_H, T) values. This grid, when tabulated, is similar in structure to Table 2.1, only differing in size: it has 34×34 elements instead of 2×2 elements. Since it is impractical to show this large table, both in this book and in scientific papers, it is illustrated using a figure. The top panel of Fig. 2.1 shows contours of constant probability at some standard levels, chosen here to be such that $-2\ln\mathscr{L}$ is lower than its maximal value by 2.3 and 6.17, where $\mathscr{L}$ is the likelihood. These levels are chosen in accordance with the current practice in astronomy and physics (at least), as recommended by Press et al. (2007). In order to get the probability distribution of

[5] This section can be skipped the first time through.

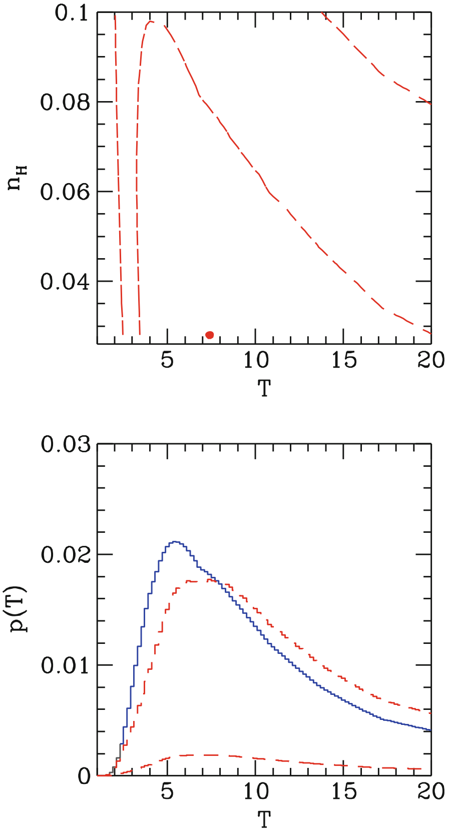

Fig. 2.1 *Top panel:* Iso-probability contours at the 68 % and 95 % confidence (*red, dashed*) levels for the (T, n_H) determination of the most distant cluster of galaxies known. The *dot* indicates the location where the likelihood takes its maximum. *Bottom panel:* Marginal probability (*blue*), profile likelihood (*lower, dashed, histogram*), and unit-normalized profile likelihood (*upper, dashed, histogram*). The figure is based on data from Andreon et al. (2009). The profile likelihood is broader and more (*right*) skewed compared to the marginal probability

T, we use the same rule used for the bag example, we marginalize over n_H, i.e., we sum up probabilities for different n_H values (when we wish to integrate out one of our random variables, in this case n_H, this variable is called a nuisance parameter). The result is shown in the bottom panel of Fig. 2.1 as a solid blue line.

Many researchers in physics and astronomy, including one of the authors early in their career, facing the problem of summarizing the probability distribution in one parameter when other nuisance parameters are present (as in our case), follow a different procedure, often mis-citing Press et al. (2007) and/or Avni (1976). For every value of T, the misinformed researchers scan the table/figure and select the value of n_H that yields the maximal probability, as shown as the lower dashed line in the bottom panel of Fig. 2.1. This procedure is often called profile likelihood. While the marginal probability sums along each column, profile likelihood records the largest values along each column. In frequentist statistics, profile likelihood methods are used for point and interval estimation of model parameters.

The upper dashed line of the bottom panel in Fig. 2.1 is again the profile likelihood after forcing its integral to be one. Normalizing the profile likelihood to one will not make it equal to the marginal probability, the former being broader and right skewed with respect to the latter.

Back to the bag example (i.e., Table 2.1) applying the profile likelihood approach returns a probability of extracting a ball of $10/22$ when there are 14 balls, out of 22 objects, in the bag. Marginalizing over the columns returns instead the obvious result ($14/22$). To summarize, if the word probability in your paper has the same sense as in your current life and if you are happy that the rules used in statistics return numbers in agreement with what good sense suggests, then remember to marginalize. This is one of the two rules you will mostly use in this book and the only one you need to remember, the other one (Bayes theorem) will be remembered by your computer.

The mathematical purist has certainly noted that our (n_H, T) variables are stepped on a grid, even if they take values on the real axes. In principle, we should have used the continuous version of the marginalization (Eq. 2.5). The latter may be (and actually is, except in purely mathematical contemplation) approximated by Eq. (2.4): as it is clear from physical and mathematical intuition, if the distribution being sampled is smooth, it will make little practical difference if one mathematically integrates the underlying function or one samples it with a tightly packed grid. In our examples, the function being sampled is quite smooth and our 34×34 grid is more than adequate for the computation.

2.9 Exercises

Exercise 1

Suppose you wish to bet on a roulette wheel. A roulette wheel is a game of chance where a number is randomly selected between 1 and 36 with the additional entries 0 and 00. The task of the gambler is to guess what the next number will be.

Exercise 1a

You place your chips on “00.” What is the probability that you will win on the next spin of the roulette wheel? (Answer: 1/38)

Exercise 1b

Using a table similar to Table 2.1, give the probabilities of getting an odd black number, an even black number, a red odd number, and a red even number on the next spin of the roulette wheel. (Answer: You should find that they are all 9/38)

Exercise 1c

Using your table from exercise 1b, what is the probability of getting a red number? (Answer: 9/19)

Exercise 1d

Do your probabilities from your table sum to one? Explain. (Answer: They should not because of the additional 0 and 00, which roulette does not define to be a color or even or odd).

Exercise 2

Suppose x and y are independent Gaussian distributions with means $\mathtt{m[1]}$ and $\mathtt{m[2]}$ and standard deviations equal to 1. Derive the joint distribution of x and y (i.e., use the fact that they are independent random variables and multiply their densities).

Exercise 3

Two random variables $\mathtt{x}[1]$ and $\mathtt{x}[2]$ have a bivariate Gaussian distribution if their joint distribution is given as:

$$f(\mathtt{x}[1],\mathtt{x}[2]) = \frac{1}{2\pi\mathtt{s}[1]\mathtt{s}[2]\sqrt{1-\rho^2}} \exp\left(\frac{z}{2(1-\rho^2)}\right)$$

where

$$z = \frac{(\mathtt{x}[1]-\mathtt{m}[1])^2}{\mathtt{s}[1]^2} + \frac{(\mathtt{x}[2]-\mathtt{m}[2])^2}{\mathtt{s}[2]^2} - \frac{2\rho(\mathtt{x}[1]-\mathtt{m}[1])(\mathtt{x[2]}-\mathtt{m}[2])}{\mathtt{s}[1]\mathtt{s}[2]}.$$

Show that the marginal distribution of $\mathtt{x}[1]$ is also Gaussian with mean $\mathtt{m}[1]$ and variance $\mathtt{s}[1]$.

Exercise 4

The Cauchy distribution can be derived by taking the ratio of two, independent Gaussian random variables both with mean 0 and variance 1. In other words, let `obsx` and `obsy` be Gaussians with mean 0 and variance 1. Then the random variable $\texttt{obsw} = \texttt{obsx}/\texttt{obsy}$ has a Cauchy distribution. Instead of proving this analytically (which the eager student can do), let us do this by simulation (a useful exercise when doing Bayesian statistics).

Exercise 4a

i) Simulate $n = 10{,}1000$, and 10,000 pairs of independent standard gaussian distribution $(\texttt{obsx}[i], \texttt{obsy}[i])$ and for each pair calculate $\texttt{obsw}[i] = \texttt{obsx}[i]/\texttt{obsy}[i]$.
ii) Plot a histogram of $\texttt{obsw}[i]$ and overlay on top of this histogram a Cauchy distribution. What do you observe as n increases?

Exercise 4b

An alternative way of obtaining random samples from a Cauchy distribution is to first simulate a uniform(0,1) random variable, say `obsz`, and calculate the new variable $\texttt{obsu} = \tan[\pi(\texttt{obsz} - 0.5)]$. Redo steps i) and ii) using this new definition of the Cauchy distribution. Compare your results from this exercise to those done in the other definition of a Cauchy distribution.

Exercise 5

This exercise will illustrate a well-known fact in probability theory (and the careful reader should remember the results of this exercise).

Exercise 5a

i) Simulate a large number of random variables (say 100,000) from your favorite distribution whose expected value exists (for example, a Gaussian or Gamma distribution, but not a Cauchy distribution). Call these simulations $\texttt{obsx}[i]$.
ii) For each $\texttt{obsx}[i]$, calculate $\texttt{obsy}[i] = \texttt{obsx}[i]^2$.
iii) We can use Monte Carlo integration to approximate an expected value using the sample average (i.e., $\bar{Z} = (\sum Z_i)/n$). Use the sample average to estimate $E(\texttt{obsx})$ and $E(\texttt{obsy})$ from your simulated random variables.
iv) Compare $[E(\texttt{obsx})]^2$ and $E(\texttt{obsy})$. What do you observe?

Exercise 5b

Repeat steps i)–iv) using the alternative transformations $\texttt{obsy[i]} = \sqrt{\texttt{obsx[i]}}$ and $\texttt{obsy[i]} = 1 + 2\texttt{obsx[i]}$. Compare $E[\sqrt{\texttt{obsx}}]$ to $\sqrt{E[\texttt{obsx}]}$ and $E[1 + 2\texttt{obsx}]$ to $1 + 2E[\texttt{obsx}]$. Make sure the distribution you are using has positive support because of the square root function. This results is actually quite general. The interested reader should investigate Jensen's inequality.

Exercise 6

One of the probability axioms is $p(\Omega) = 1$ (Sect. 2.2). Although this probability is quite obvious, we invite the reader to browse the literature of their favorite subject to find counterexamples, i.e., cases in which the integral over the whole set of possible values of the "probability" distribution differs from one. Lazy astronomers may find a suggestion inside Sect. 2.2.

References

S. Andreon and M.A. Hurn. The scaling relation between richness and mass of galaxy clusters: a Bayesian approach. *Monthly Notices of the Royal Astronomical Society*, 404(4):1922–1937, 2010.

S. Andreon, B. Maughan, G. Trinchieri, and J. Kurk. JKCS 041: a colour-detected galaxy cluster at $z_{phot} \sim 1.9$ with deep potential well as confirmed by X-ray data. *Astronomy & Astrophysics*, 507:147–157, 2009.

K.B. Athreya and S.N. Lahiri. *Measure theory and probability theory*. Springer-Verlag New York Inc, 2006.

Y. Avni. Energy spectra of X-ray clusters of galaxies. *The Astrophysical Journal*, 210:642–646, 1976.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, 2001.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, third edition, 2007.

Chapter 3
A Bit of Numerical Computation

The posterior probability distribution quantifies what we known about a quantity after observing the data, the narrower this distribution is the better the quantity is measured. But how to compute this distribution in practice? In problems of some complexity, the analytical computation of the posterior distribution is either cumbersome, exceeds the analytic skills (or time available) of the researcher, or, more simply, does not exist in closed form. It is of the opinion of the authors that CPU time costs less than brain time, so why should researchers spend their own time when a machine can do the work for them?

The development of powerful software for stochastic computation, for example WinBUGS, OpenBUGS, and JAGS,[1] replaces the researcher's analytical skills with CPU power, i.e., with numerical computation: the "desired" function, i.e., the posterior distribution, is computed by numerical approximations, quite similar to the way integrals are often computed. In particular, the target function is evaluated by some sort of sampling scheme, as in Monte Carlo computations. The value of $f(\theta)d\theta$ is given by the number of θ_is between $\theta - \Delta\theta/2$ and $\theta + \Delta\theta/2$: the larger the value of the function at θ, the more samples will fall in the considered range. In this book, we refer to the list of θ_is as a chain, or Monte Carlo sample.

To the end user, the precise knowledge about how the posterior is numerically computed has the very same relevance as how the computations are performed for the sine or logarithm functions: almost none.[2] There are however a few things, most of which fall in the category of "numerical tips," that need to be known:

- Internal to the mentioned software (WinBUGS, OpenBUGS, and JAGS) are the full expressions of most common functions and distributions. For example, to specify the Gaussian distribution, one does not need to type:

[1] JAGS (Plummer 2010) can be downloaded from http://mcmc-jags.sourceforge.net/.

[2] Skeptical readers are compelled to check that the numerically-sampled posterior is identical to the analytically-derived posterior listed in the exercises.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_3

$$p(x_0|x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right] \quad , \tag{3.1}$$

but simply

$$x_0 \sim \mathcal{N}(x,\sigma^2) \quad . \tag{3.2}$$

This not only saves time for the researcher, but also prevents coding mistakes, especially with distributions more complex than Gaussians.

- The posterior distribution is numerically sampled. These samples are called chains. Additional draws (a longer chain) typically deliver a more precisely determined posterior distribution unless the newly added samples are sampling the posterior in a poorer way than those we already have. From these draws one may directly estimate means, standard deviations, and credible regions of any parameter or any function of parameters. For example, for a point estimate one may compute the sample mean, median, or mode of the chain. Similarly, to compute a 90 % interval of the parameter σ in Eq. (3.1), it is sufficient to take the interval that contains 90 % of the σ samplings.
- Most software platforms, such as WinBUGS, OpenBUGS, and JAGS, sample the posterior using a special Monte Carlo computation technique called Markov Chain Monte Carlo (MCMC), in particular, the Gibbs sampler or Metropolis-Hastings algorithm. Using these methodologies, values in the chain are not fully independent of each other, or, said more simply, the precision of the posterior increases slowly with chain length. This is yet another reason to prefer a long chain to a shorter one.
- If independent samples are needed, it is useful to decimate (take one element every 10) or centimate (one in 100) the chain (known as thinning). Additionally, thinning also saves disk space. Note that the computational time for our examples or for science based models is often of the order of seconds or minutes, even on an obsolete laptop, i.e., computations of a long chain is totally inexpensive, and always shorter than the time needed to write a paper describing the found results. There is, therefore, always time for computing a long chain.
- The computations used in an MCMC depend on a set of initial values. At the beginning of the MCMC, the sampler is not necessarily sampling from the desired function (the posterior distribution). Given enough time, the MCMC will begin drawing samples from the desired function. Therefore, it is good practice to discard a generous number of draws from the beginning of the chain. This way you have a better chance of obtaining a chain that is from the posterior. This practice of discarding the first set of draws from the MCMC is known as "burn-in." Again, it is inexpensive to obtain long chains and so accounting for the additional "burn-in" samples is not a huge burden.

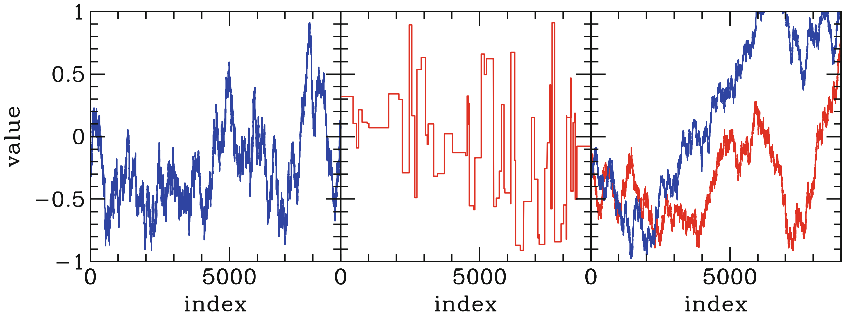

Fig. 3.1 Examples of trace plots that indicate that the chain's length needs to be longer: in the *left panel*, the chain is struggling to explore the posterior (shows a slow mixing), in the *central panel*, the chain resides on the same value for too long before changing, in the *right panel*, the two chains are still not converging

3.1 Some Technicalities[3]

As with all numerical computations, there is always the risk that things will not work correctly. A possibility is that the saved chain is only coarsely sampling the posterior because it is "not long enough" for the scientific application in mind. Fortunately, there exist some diagnostic tools (e.g., CODA, Plummer et al. 2006) that can be used to determine when this occurs (although the proposed methods do not always guarantee to catch chains that are too short):

- Refer to Fig. 3.1. Start by inspecting the chain's trace (i.e., plot draw i against its index i). If the chain slowly moves from one value to another (central panel), or if the chain is slowly wandering (left-hand panel), then longer chains or more efficient sampling schemes are needed. Sometimes, researcher's "cleverness" may replace "brute force" CPU, devising a way to output better behaved chains.
- Recall that all MCMCs must begin with a set of initial values. To assess the MCMC's dependency on the initial values, and to check convergency, it is advised to start multiple chains from different starting values (this is done in JAGS and Open/WinBUGS at no additional programming cost for the researchers). Then, if the multiple chains converge, one can have confidence that the MCMC is converging to the desired function. Otherwise, the chains have not converged (right-hand panel) and longer chains (or more thoughts from the researcher) are needed. Some programs, such as CODA (Plummer et al. 2006), automate the task of comparing chains produced by JAGS and Open/WinBUGS, and return a number of diagnostics. One of them is the R statistic, given by the scatter in the chain vs. the scatter across the chains. For converged chains, R should be near to 1.

To be fully honest, there is a more powerful method: your brain. For a hypothetical statistical model, it is difficult to imagine the shape of the posterior. In a

[3] This section can be skipped the first time through.

mathematical world, one can always imagine a problem for which there is no way to guess the posterior without computing it. For a real experiment that was carefully performed, we approximately know how well we will be measuring the target quantity, i.e., we have a qualitative summary of the posterior. We may have it by physical intuition, by similarity to a previously run experiment, or from forecasts which recently have become common given the large cost of experiments, telescopes, and instruments. In astronomy, where data is abundant and mostly public, careful researchers first do a qualitative (or quick and dirty) analysis aimed at determining which data is best for a given purpose (the one giving the sharper posterior, as qualitatively computed). Once a powerful data set is identified, the researchers start to download the data from the archives, clean the data, and then perform the analysis. Using software, the researchers quantitatively compute a point estimate and associated uncertainty of the target quantity, improving upon the initial qualitative analysis (e.g., that the target parameter is expected to be determined at, say, around 10 % accuracy with the chosen data set).

We conclude this section by stating that the numerical issues just raised are not a problem of Bayesian methods, but instead the computational issues related to the numerical implementation of the methods adopted in this book (MCMC). Analytic Bayesian computations do not have convergence issues, as well as posterior samples obtained using sampling methods such as importance sampling (as MultiNest, see Feroz et al. 2009). Unfortunately, analytical methods and importance sampling require the researcher's time: the researcher has to either program the analytical results or program the likelihood for the importance sampler, rather than using a simpler symbolic expression (as in this book). It is for this reason we use the symbolic expressions in this book.

3.2 How to Sample from a Generic Function[4]

While many distributions have been implemented in JAGS, there are still many that have not (or it is hard for a researcher with normal analytical skills to note that the wanted distribution is an "obvious" case of one available). Here we want to introduce the most straightforward implementation of a generic distribution in JAGS.

We begin by telling JAGS a little white lie, called the "zero trick," namely that we observe a datum point with a value of 0 from a Poisson distribution with mean $\log(g(y,\theta))$. For a Poisson(μ) distribution, the probability of observing a zero is $e^{-\mu}$. So, for a Poisson($\log(g(y,\theta))$) distribution, the probability is $g(y,\theta)$. Because the rate of a Poisson is strictly positive, one may need to add some suitably big C, $\log(g(y,\theta))+C$ to enforce this constraint.

Let us suppose, for example, that we are interested in obtaining a sample from a Cauchy distribution,

$$p(y|\theta) \propto \frac{1}{1+(y-\theta)^2} \tag{3.3}$$

[4] This section can be skipped the first time through.

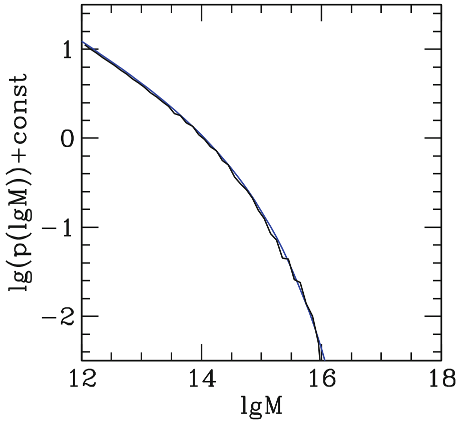

Fig. 3.2 Schechter (1976) luminosity function. The analytic expression is given by the *blue, smooth, solid line*. The histogram of a sample derived using the "zero trick" is given by the *black, jagged curve*

with location parameter θ. In JAGS, this is implemented as:

```
data {
zeros <- 0
C <- 1
}
model {
#Using the zero trick
y ~ dnorm(0.0,1.0E-6)
phi <- log(1+pow(y-theta,2))
zeros ~ dpois(phi+C)
}
```

In short, we only need to complete the line starting with "`phi <-`" with minus the natural logarithm of the mathematical expression of the distribution that we want to sample, and remember to add a suitable constant `C` to keep `phi+C` positive.

For example, if we want to sample from a function known as the Schechter (1976) function with parameters lgM^*, α:

$$p(lgM|\alpha, lgM^*) \propto 10^{0.4(\alpha+1)\times(lgM-lgM^*)} \times \exp\left[-10^{0.4\times(lgM-lgM^*)}\right] \quad (3.4)$$

$$\ln p(lgM|\alpha, lgM^*) \propto 2.3025\times 0.4(\alpha+1)\times(lgM-lgM^*) - 10^{0.4\times(lgM-lgM^*)}. \quad (3.5)$$

As seen from the analytic expression and Fig. 3.2, the Schechter function is a straight line bended at one end (damped) by an exponential.

In JAGS, the model reads:

```
data{
zeros <- 0
C <- 10
}
```

```
model{
lgM ~ dunif(12,18)
#Using the zero trick
phi <- -2.3025*0.4*(alpha+1)*(lgM-lgMstar)+10^(0.4*(lgM-lgMstar))
zeros ~ dpois(phi+C)
}
```

Readers with advanced analytical skills might find more elegant ways of drawing a sample from an arbitrary distribution than the zero trick. For example, one may recognize that the previously mentioned Cauchy distribution can be computed as the ratio between a standard Gaussian (mean 0 and standard deviation 1) and the square root of a χ^2 distribution with one degree of freedom. In turn, a χ^2 is a Gamma distribution with parameters (0.5,0.5). The Cauchy can then be coded in JAGS as:

```
model{
tau.eta ~ dgamma(.5,.5) # chi2
x~dnorm(0.,1.)
# cauchy = normal/sqrt(chi2)
theta <- x / sqrt(tau.eta)
}
```

Similarly, the Schechter function with $\alpha > -1$ can be written using a Gamma distribution and a power-law can be coded as a Pareto distribution. Nowhere in this book do we assume that you will be able to recognize a target distribution as a combination of other distributions. In this book, we will use the "zero trick" even when we are aware of a more elegant solution.

References

F. Feroz, M. P. Hobson, and M. Bridges. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398:1601–1614, 2009.

M. Plummer. *JAGS Version 2.2.0 user manual*, 2010.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11, 2006.

P. Schechter. An analytic expression for the luminosity function for galaxies. *The Astrophysical Journal*, 203:297–306, 1976.

Chapter 4
Single Parameter Models

In this chapter we introduce some of the basic concepts of Bayesian statistics. These concepts are illustrated with probability models containing a single scalar-valued parameter. In particular, a Gaussian distribution with a known standard deviation is used to illustrate the basics of a Bayesian analysis of an experiment that attempted to measure the mass of the neutrino. Following the neutrino example, we introduce other single parameter probability distributions that are useful to Physicists and Astronomers.

4.1 Step-by-Step Guide for Building a Basic Model

4.1.1 A Little Bit of (Science) Background

The first step, and probably the most important, of any statistical analysis is understanding the problem.

Consider an experiment[1] where the researchers are interested in measuring the mass of the (anti-electron) neutrino, denoted as `m`. For this simplified analysis, let us assume that the instrument resolution (the likelihood) is adequately modeled as a Gaussian distribution with known width (i.e., standard deviation σ) equal to 3.3 eV/c^2 (note from here on the units are not always quoted for this example).

The results of the experiment yielded an observed value of $\texttt{obsm} = -5.4$ where `obsm` denotes the observed mass. Notice that a negative number is observed for the positively defined mass. This type of result may occur in frontier measurements: the (electron) neutrino mass is computed (in our experiment) as the difference between two big and similar quantities, each one not perfectly known (the neutron and proton masses. Of course there is also the mass of the electron to be accounted for). For example, in 1994, seven out of the eight published values of the neutrino squared

[1] The experiment is real, but dated, for reasons that will be self-evident by the end of this section.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_4

mass were negative (Montanet et al. 1994), and the situation has not improved too much in the next 20 years. These negative observations of a positive quantity have prompted recent experimenters to develop different methods for measuring the mass of the neutrino.

4.1.2 Bayesian Model Specification

As with any statistical analysis (and Bayesian statistics is no exception), a model must be specified. A Bayesian model specification consists of specifying a data model (i.e., the likelihood) and a distribution for the parameter (i.e., the prior distribution).

For the neutrino mass example, the data model was already mentioned to be a Gaussian distribution with mean equal to the true value of the neutrino (m) and with standard deviation 3.3. JAGS, following BUGS (Spiegelhalter et al. 1996), uses precision $\texttt{prec} = 1/\sigma^2 = \texttt{pow}(\sigma, -2)$, in place of variance σ^2 (the pow function, as you might guess, stands for power). Therefore our wordy statement on the Gaussian error is mathematically expressed by:

```
obsm ~ dnorm(m,prec)
prec<-pow(3.3,-2) ,
```

where the symbol $\sim$ denotes "distributed as," and the arrow reads "takes the value of."

To finish our model specification, a prior distribution has to be assigned to m. Ideally, the prior distribution should come from experts who have a strong knowledge of the mass of the neutrinos. Here we take, for simplicity, a uniform prior between 0 (mass is positive) and 33 (a value certainly larger than all plausible values of mass):

```
m ~ dunif(0,33) .
```

In JAGS notation, our model reads:

```
model{
#Data Model
obsm ~ dnorm(m,prec)
prec<-pow(3.3,-2)
#Prior Distribution
m ~ dunif(0,33)
}.
```

4.1.3 Obtaining the Posterior Distribution

Now that we have specified our model, we need to obtain draws from our posterior distribution so that we may perform inference. Our model is simple enough that

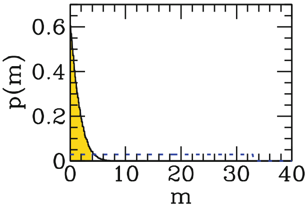

Fig. 4.1 Posterior probability distribution of the neutrino mass (*histogram*). The *shaded region* marks the 95 % highest posterior credible interval. The *dashed line* indicates the assumed prior

a posterior distribution can be obtained analytically (see, for example, D'Agostini 2003, from which this example is taken, and also the exercises at the end of this chapter). We prefer instead to compute the posterior numerically through Markov chain Monte Carlo (MCMC), i.e., with JAGS, because our time is more precious than CPU time.

Figure 4.1 displays the posterior distribution of m. The posterior has been computed from the MCMC draws (i.e., JAGS output) and plotted with a histogram, along with its prior distribution (dashed line).

Figure 4.1 illustrates how our current knowledge of the neutrino mass (m) has updated after observing obsm= -5.4. In particular, notice how before observing obsm, our prior knowledge stated that m was equally likely to be a value in the interval 0 and 33. After observing obsm, notice how the bulk of the posterior distribution's mass is now less than eight, indicating the data was informative in changing the prior to the posterior.

4.1.4 Bayesian Point and Interval Estimation

Once the posterior distribution is obtained, parameter estimates (m for the neutrino mass example) can be calculated by various summaries of the posterior distribution. For example, point estimates of the parameter can be computed using either the mean, median, or mode from the posterior distribution (i.e., of the Monte Carlo samples).

Table 4.1 provides various sample statistics for both the prior and posterior distributions. Notice again that before observing the data, the prior mean and median are quite large. After observing obsm, however, the likelihood shifted these quantities to smaller values.

As with any statistical analysis, we wish to not only estimate our model parameters, but also express our uncertainty in these estimates. We may, for example, compute the posterior width (scale), by quoting the standard deviation of the chain, or its interquartile range. We may also use the Bayesian analogue to the confidence interval called the credible interval. Unlike the confidence interval, these intervals have the interpretation "the probability that our parameters is in the interval is .95."

Table 4.1 Prior and posterior sample statistics for the neutrino mass $\mathtt{m}$

Sample statistic	Prior	Posterior
Mean	16.5	1.38
Median	16.5	1.04
Mode	–	1.01

Table 4.2 Prior and posterior widths for the neutrino mass $\mathtt{m}$

Distribution	Interquartile range	95 % HD interval
$\mathtt{m}$ prior	[8.25, 24.75]	–
$\mathtt{m}$ posterior	[0.45, 0.98]	[0, 3.84]
$\mathtt{m}^2$ posterior	[0.20, 0.39]	[0, 14.74]

Similar to point estimation, these intervals are obtained using the chain. For example, to calculate the 95 % highest density (HD) interval we take the interval that includes 95 % of the drawings with the constraint that all points in the interval have higher probability than points outside the interval. For scalar-valued parameters, 95 % highest density intervals are also the 95 % shortest intervals. For simplicity, we will sometimes refer to them as "95 %" intervals, or "95 % probability" intervals. We may also, if desired, compute 95 % upper limits, by taking the value larger than 95 % of the chain elements. Similarly, we may also compute the interquartile range, for example by ranking the elements in the chain and computing the interval between the 25th and 75th percentile.

Returning to the neutrino mass example, Table 4.2 provides the interquartile range, and the 95 % HD intervals for the neutrino mass $\mathtt{m}$ for both the prior and posterior distribution. Before observing our data, our uncertainty in $\mathtt{m}$ was quite large (relative to the posterior). As we make an observation, the interquartile range and the 95 % interval of the posterior illustrate how more certain we are about the neutrino mass $\mathtt{m}$ (its width is much smaller than the width of the prior interval).

The 95 % upper limit is 3.84, comparable with the instrument resolution (3.3), i.e., with what good sense expects it to be: if the neutrino mass turns out to be unmeasurable (we got a negative number!), then the upper limit to the neutrino mass should be of the order of the instrument resolution.

Intervals for other quantities can be easily derived. For example, we may want to know the highest 95 % posterior density interval for the squared mass, $\mathtt{m}^2$. It is just a matter of taking the square of each $\mathtt{m}_i$ value listed in the chain (Sect. 2.6) and computing the 95 % interval (as already done for $\mathtt{m}$). The 95 % upper limit of the square mass is 14.74 eV^2/c^4, the 96.5 % upper limit is 17.80 eV^2/c^4

4.1.5 Checking Chain Convergence

Figure 4.2 shows the first 950 draws (the remaining look qualitatively similar), after burn-in, of the three chains used. Qualitatively speaking, we do not see anything

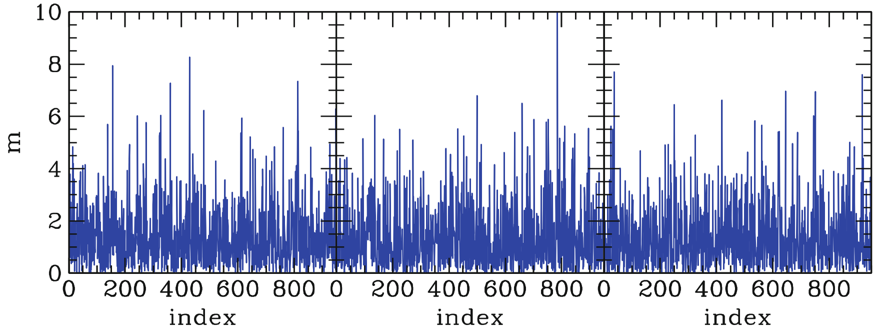

Fig. 4.2 Trace plots of the neutrino mass posterior where each of the three panels contains a different chain

unusual in these chains (see Sect. 3.1). Quantitatively, our inferences (mean values, intervals, etc.) are the same using any of the three chains and, finally, the computation of the R statistic (see Sect. 3.1) is already equal to 1.00 with the first 950 samplings (note that for Table 4.2 we used the full chains). This strongly suggests our MCMC has converged to the correct distribution.

4.1.6 Model Checking and Sensitivity Analysis

Before relaxing and being proud of our obtained result (the upper limit above), it is wise to check if the data are consistent with the fitted model and assessing the role of the prior on the result. These key ingredients of every statistical analysis are approached in Chap. 9. In particular, the sensitivity analysis of this experiment is detailed in Sect. 9.1.1.

4.1.7 Comparison with Older Analyses

Chapter 10 provides some detailed comparisons of the Bayesian approach to older (and simpler) ways of doing statistics in astronomy and physics. Let us focus on a couple of previous analyses of the neutrino mass.

Our analysis returns [0,17.80] eV^2/c^4 as the 96.5 % interval of the square mass. Some other (and older) analyses state that "$\mathtt{m}^2$ is positive with only a 3.5 % probability." We prefer our (and indeed every positive) $\mathtt{m}^2$ 96.5 % range, only including possible physical values for the quantity being measured, to stating that the quantity being measured has 96.5 % probability to have an impossible or unphysical value (the square of a real number is non-negative!).

Our analysis returns 3.84 eV/c^2 as the 95 % upper limit, comparable with the instrument resolution (3.3 eV/c^2). Let us consider what may occur in a sloppy analysis of the neutrino data if one forgets that mass is positively defined and computes the 84 % $(= 50 + 68/2\,\%)$ upper confidence interval as the observed value plus the likelihood σ (effectively only using the likelihood and ignoring the prior). These analyses find $-2.1(= -5.4 + 3.3)$ eV/c^2 as the 84 % upper confidence interval because the observed value is -5.4 and the likelihood σ is 3.3. If we are to completely trust this result, then this 84 % upper confidence interval is more stringent (it includes a very short range of the physically acceptable range, precisely one of zero length) than every positive value a future precise experiment will return for the neutrino mass (e.g., 0.4 eV/c^2). This is against logic (precisely as a negative square of a real number): a clearly untrustable upper limit (such as a negative quantity for a positive defined quantity) cannot be more trustworthy than the upper limit derived in a precise experiment.

4.2 Other Useful Distributions with One Parameter

Now that we have introduced some of the concepts related to Bayesian analysis, we introduce other distributions with one parameter that are of interest to astronomers and physicists. In particular, we introduce the Poisson and binomial distribution.

For the remaining examples in this book, we will not be as detailed in our analysis as we were in Sect. 4.1. Instead, we skip these common steps of the analysis and focus more on the steps that change from one analysis to another (e.g., inference). We urge the reader to reference back to Sect. 4.1 if more of the details are needed.

4.2.1 Measuring a Rate: Poisson

Many natural phenomena can be modeled as a Poisson process: the number of photons being emitted from a source, the number of galaxies, the number of nuclear particles in a nuclear reaction, etc. Consider, for example, the number of photons from a source. Because of the quantum nature of photons and possibly other errors in the observation process, the observed and true photon counts are not the same. In this situation, the observed number of collected photons, denoted as `obsn`, may be adequately modeled using a Poisson distribution with parameter `s` > 0. Here, `s` is interpreted as the true number of photons being emitted from the source. In JAGS, this model is denoted by

```
obsn ~ dpois(s) .
```

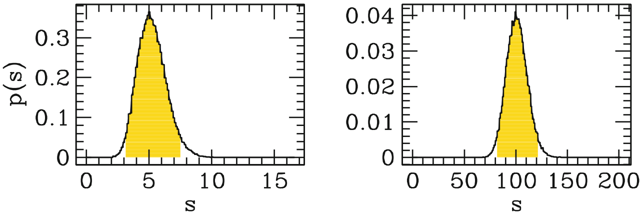

Fig. 4.3 Posterior probability distribution for the flux *s* of a source when `obsn` = 4 (*left panel*) and `obsn` = 100 (*right panel*). The *shaded region* marks the 95 % highest posterior credible interval

Table 4.3 Sample statistics and 95 % credible intervals for `s` in the point source example for the different observation cases

`obsn`	Mean	Median	Mode	Standard deviation	95 % HD Interval
4	5	4.7	3.8	2.2	[1.21, 9.44]
100	101	100.7	100.1	10.0	[81.8, 121.1]

For illustrative purposes, we assign a uniform prior on `s`, where `s` lies between zero and some large value, say 1×10^7 (for this example the results do not depend on which large value is taken, though in general this may not be true). After observing `obsn` and specifying a Bayesian model (likelihood and prior), we are now able to make inference on `s` through its posterior distribution. Our model (in JAGS) reads:

```
model{
obsn ~ dpois(s)
s~ dunif(0,1.0E+7)
}.
```

Consider two possible data observations:

1. We observed 4 counts from the source (which is typical for sources detected by the Chandra Space X-ray Telescope).
2. We observed 100 counts from the source.

As in the previous section, interest is in obtaining the posterior probability distribution for the source signal `s`.

Figure 4.3 shows the posterior distribution for these two cases. In the second case (right panel), the shape of the posterior looks symmetric about its mode, similar to a Gaussian distribution. In the former case (left panel), however, the distribution is slightly asymmetric (or skewed), similar in shape as a Gamma distribution. Additionally, we provide various point estimates of `s` and uncertainty estimates for this parameter in Table 4.3.

4.2.2 Combining Two or More (Poisson) Measurements

Suppose now that we have k independent measurements of the signal s from the point source, i.e., k values of obsn, say $obsn_1$ and $obsn_2$. JAGS uses "[i]" in place of the subscript i, i.e., obsn[i] in place of $obsn_i$, which will serve as our notation for the rest of the book. Clearly, we want to use all these k observations to estimate the source flux s. If we assume all measurements are independent, then the probability of observing this set of measurements is just the product of the probabilities of observing each measurement (i.e., through the probability product rule, see Sect. 2.4).

Fortunately, JAGS is general enough that we only need to add terms with an index i in a for loop structure. The model reads:

```
model{
for (i in 1:length(obsn)){
obsn[i] ~ dpois(s)
}
# uniform prior on s
s ~ dunif(0,1.0e+7)
},
```

where length(obsn) is the number of observations in obsn.

Suppose we have performed various experiments on the same point source yielding different numbers of observations for each run of the experiment, say one, four, ten, and forty observations. Figure 4.4 shows the posterior distributions obtained from these different data scenarios. At the start, when only a single measurement is available, the posterior distribution is quite broad: there is little information in this single data point. As the amount of data increases, the posterior is dominated by the information contained in the data. As the amount of data increases, the posterior becomes a sharper distribution, becoming more centered on the true value of s.

The simulated observations were generated from the following probability model (JAGS code):

```
model{
obsn ~ dpois(4)
},
```

i.e., we observed N observations from a Poisson distribution with parameter s$= 4$ where the number of observations (saved elements of the chain) is $N = 1, 4, 10, 40$. The first four drawn values are $8, 4, 3, 5$ (these differ at the next running of the JAGS code, of course).

4.2.3 Measuring a Fraction: Binomial

Consider processes or experiments where there are only two possible responses, for example recording the result of flipping a coin, or if a particle is decayed, or if a

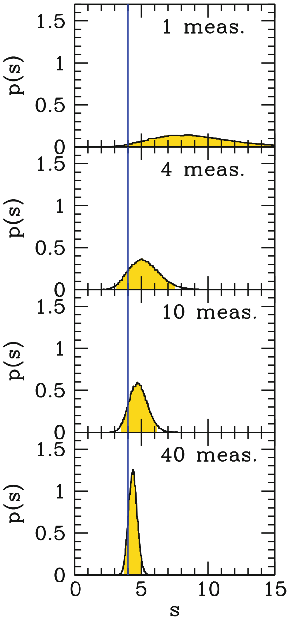

Fig. 4.4 Posterior probability distribution of the Poisson signal *s*. The *shading* marks the 95 % highest posterior interval. As the number of measurements increases the posterior get narrower and narrower, i.e., the signal will be more precisely measured. The *vertical line* represents the true value of s

galaxy can be characterized as being a red galaxy or not. Such processes are called Bernoulli trials. A Bernoulli trial is characterized by a random draw with only two outcomes, typically labeled as success or failure, with probability f of observing a success (and likewise, 1-f of observing a failure).

Suppose we are interested in characterizing the probability that a galaxy is either a red galaxy (success) or not (failure) or, equivalently, the fraction f of red galaxies. Then, if we observe n galaxies, the number of observed red galaxies obsn will follow a binomial distribution with probability f (here we are assuming that we can perfectly characterize a galaxy as being a red galaxy or not and that the number of galaxies from which the n are drawn is sufficiently large). Other physical examples that can be adequately modeled with a binomial distribution are: the fraction of absorbed active galactic nuclei in the Universe, the branching ratio of a nuclear reaction, the fraction of particles of a given type, etc.

The binomial model states that the observed value obsn is binomially distributed with parameters f and n:

```
obsn ~ dbin(f,n) .
```

For the red galaxy example, on average, we expect the number of red galaxies to be near f × n, and rarely far from this number. In other words, for fixed values of the parameters f and n, we expect to observe a distribution of obsn values, through repeated experimentation, centered on f × n.

To help illustrate the use of the binomial model, consider the following example from Astronomy. According to some theories, gas-poor galaxies are formed by the merging of other gas-poor galaxies. Just before merging, the galaxies should gradually be approaching each other, i.e., we should observe gas-poor galaxies with a nearby gas-poor companion. Therefore, using the fraction f of gas-poor galaxies having a nearby gas-poor companion, one may derive the gas-poor (dry, in the astronomical parlance) merging rate: a low value means that gas-poor galaxies did not experience many gas-poor mergers in the past and a large value means that present-day gas-poor galaxies were separate gas-poor galaxies previously.

The problem can be formulated as follows, after observing n passive (gas-poor) galaxies, obsn of these galaxies have a passive nearby companion. We then wish to estimate the fraction f of passive galaxies having a nearby passive companion. As just mentioned, the parameter f is relevant in astrophysics after its value is transformed into the appropriate astronomical units, yielding the dry merging rate.

To begin our Bayesian inference, we must first specify a prior for f (recall that n is known and so no prior is needed). We take a uniform prior between 0 and 1. This may be coded in JAGS as f $\sim$ dunif(0,1). After noting that a $beta(1,1)$ function is uniform between 0 and 1, the posterior can be easily analytically computed: $p(\mathtt{f}|\mathtt{obsn}) = beta(\mathtt{obsn}+1, \mathtt{n}-\mathtt{obsn}+1)$ (Laplace 1812, this result is known as conjugacy). For now, suppose we ignore this analytical result and instead proceed as usual using JAGS. The Bayesian model reads:

```
model{
obsn ~ dbin(f,n)
f~dbeta(1,1)
}.
```

The data for our example is taken from the remarkable work by De Propris et al. (2010): the authors observed n= 2127 galaxies and characterized obsn= 2 of these galaxies as having a nearby gas-poor companion galaxy.[2] The posterior distribution of the fraction f, computed using JAGS, is plotted in Fig. 4.5 as the black histogram, whereas the analytical derived *beta* distribution, $beta(3,2126)$, is superimposed in blue. The analytically and numerically computed posteriors are, of course, identical (apart from sampling errors in the MCMC). The posterior has a peak at $\mathtt{f} = 2/2127 = 0.094\,\%$ and a standard deviation of 0.081 %.

De Propris et al. (2010) are interested in quoting a 5σ upper limit for f. We are interested in estimating the same quantity here, meaning, more precisely, a $1-2\times 10^{-7}$ (or 99.99998 % confidence) upper limit, which corresponds to the 5σ value for a Gaussian distributed quantity.

[2] The author's starting galaxy sample is larger, but not all galaxies from their sample can be recognized as having a nearby companion, even if they truly have one. This is due to observational limitations. Therefore, using the same approach as the authors, we adopted an effective sample of 2127 galaxies.

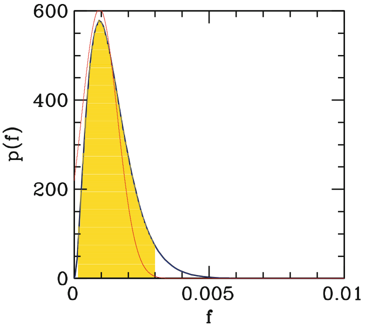

Fig. 4.5 Posterior probability distribution for the fraction f of galaxies in pairs computed using both JAGS (*the black histogram*) or by (analytic) calculus (*blue line*, hardly distinguishable from the histogram), based on data in De Propris et al. (2010). The *shaded region* marks the 95 % highest posterior credible interval. The *red, symmetric, curve* is the Gaussian approximation adopted in De Propris et al. (2010)

To compute the upper limit with a numerical sampling of the posterior (e.g., with JAGS) we need:

a) to save a chain from the MCMC of length l;
b) sort the chain of `f` in increasing order;
c) take the value that leaves to its left $(1-2\times 10^{-7})\times l$ elements of the chain.

For a chain of length $l = 2\times 10^{7}$ elements, this means take the largest or the second-largest point. This operation is computationally expensive, especially if we want to compute this number with high accuracy. In fact, in order to reduce the Monte Carlo error, one may want a chain containing 10^8 elements! We therefore turn to the analytically derived posterior distribution. To calculate this quantity, we integrate the $beta(3,2126)$ density function from `f` $=0$ until its integral is $1-2\times 10^{-7}$. Again for computational precision, it is instead preferable to start from `f` $=1$ and stop when the integral of $1-$`f` is 2×10^{-7}, using the well-known fact that all probabilities sum up to one. We find a value of 0.017, i.e., that $f < 0.017$ at 5σ.

What would happen if De Propris et al. (2010) observed five pairs of galaxies instead of two? Integrating the posterior distribution, $beta(6,2123)$, we obtain an upper limit of $0.020 = 0.2\,\%$. For such a case, De Propris et al. (2010) quote an upper limit half this size because they approximate the binomial likelihood with a Gaussian (the red curve shown in Fig. 4.5 for the case of 2 pairs). This factor of two applies also to the dry merging rate, which is the very same number expressed in astronomical units. Because of the adopted approximation, they obtained an upper limit to the dry merger rate that is double than what it should be, i.e., they were more optimistic than their data by a factor of two.

4.3 Exercises

Exercise 1

Many processes having to do with times between events may be modeled with an exponential distribution, such as decay time. Assume the likelihood of our data is an exponential distribution with rate parameter $\texttt{s}$, i.e., $\texttt{obsy} \sim \texttt{dexp(s)}$. Assume a gamma prior distribution for $\texttt{s}$ with parameters $\texttt{a}$ and $\texttt{b}$, $\texttt{s} \sim \texttt{dgamma(a,b)}$. Using Bayes theorem, show that the posterior distribution is $\texttt{s} \sim \texttt{dgamma(a}+1,\texttt{obsy}+\texttt{b)}$, a gamma distribution with parameters $\texttt{a}+1$ and $\texttt{b}+\texttt{obsy}$. Note this is one of those special cases where the posterior can be derived. Use JAGS to verify this result with $\texttt{a}=\texttt{b}=1$ and $\texttt{obsy}=0.819$.

Exercise 2

Many positively defined quantities in astronomy and physics are distributed as log normal, for example, the mass of galaxy clusters. Suppose that $\texttt{obsm} \sim \texttt{dlnorm (m,prec)}$ where $\texttt{prec}$, the precision, is known. Assume the prior $\texttt{m} \sim$ dnorm $\texttt{(a,b)}$. Using Bayes theorem, show that the posterior distribution is also normal with posterior mean

$$\mu = \frac{\texttt{a}*\texttt{b}+\texttt{prec}*\log(\texttt{obsm})}{\texttt{b}+\texttt{prec}}$$

and posterior precision $\texttt{b}+\texttt{prec}$. Check this result with JAGS when $\texttt{obsm}=2$, $\texttt{prec}=1$ and the prior parameters are $\texttt{a}=0$ and $\texttt{b}=1$.

Exercise 3

Suppose $\texttt{obsx} \sim \texttt{dunif}(0,\texttt{b})$ and assume a power law (Pareto) prior for $\texttt{b}$ with parameters $\texttt{s}$ and $\texttt{f}$, $\texttt{b} \sim \texttt{dpar(s,f)}$. Using Bayes theorem, show that the posterior distribution is $\texttt{b} \sim \texttt{dpar(s}+1,\max\{\texttt{obsx},\texttt{f}\})$ which is a Pareto with parameters $\max\{\texttt{obsx},\texttt{f}\}$ and $\texttt{s}+1$. Use JAGS to verify this result with $\texttt{s}=2$, $\texttt{f}=1$, and $\texttt{obsx}=1.138$.

Exercise 4

Suppose $\texttt{obsn} \sim \texttt{dpois(s)}$ and $\texttt{s} \sim \texttt{dunif}(0,\texttt{smax})$. Show that the posterior distribution is:

$$p(\texttt{s}|\texttt{obsn}) = \frac{c}{\Gamma(\texttt{obsn}+1)}\texttt{s}^{\texttt{obsn}}e^{-\texttt{s}}I[\texttt{s} \in (0, \texttt{smax}))]$$

where

$$c = \frac{1}{\int_0^{\texttt{smax}} \frac{1}{\Gamma(\texttt{obsn}+1)}\texttt{s}^{\texttt{obsn}}e^{-\texttt{s}}d\texttt{s}}.$$

where I is the indicator function with $I(x \in A) = 1$ if $x \in A$ and 0 otherwise. Use JAGS to verify this result with $\texttt{smax} = 5$ and $\texttt{obsn} = 3$.

Exercise 5

Again consider $\texttt{obsn} \sim \texttt{dpois}(\texttt{s})$ and $\texttt{s} \sim \texttt{dgamma}(k, r)$ where $r = 1/\texttt{theta}$ for $\texttt{s} \in (0, \texttt{smax})$. Show that the posterior distribution is:

$$p(\texttt{s}|\texttt{obsn}) = \frac{c(\theta+1)^k}{\theta^k}\frac{1}{\Gamma(k+\texttt{obsn})}\texttt{s}^{k+\texttt{obsn}-1}e^{-\texttt{s}(\theta+1)/\theta}I[\texttt{s} \in (0, \text{smax}))]$$

where

$$c = \frac{1}{\int_0^{\text{smax}} \frac{(\theta+1)^k}{\theta^k}\frac{1}{\Gamma(k+\texttt{obsn})}\texttt{s}^{k+\texttt{obsn}-1}e^{-\texttt{s}*(\theta+1)/\theta}d\texttt{s}},$$

and I is the indicator function with $I(x \in A) = 1$ if $x \in A$ and 0 otherwise. Use JAGS to verify this result with $\texttt{k} = 5$, $\texttt{theta} = 1$, $\texttt{smax} = 8$, and $\texttt{obsn} = 3$. A note of caution: in JAGS, the parameterization of the gamma distribution uses $r = 1/\theta$.

Exercise 6

Analytically derive the posterior distribution for the binomial model with a uniform prior, following the suggestion in Sec 4.2.3. For the analytical derivation, recall that the posterior has an integral equal to one (as all probability distributions should). Thus, you may leave aside all constants (i.e., quantities independent of $\texttt{f}$) and derive just the functional dependence with $\texttt{obsn}$ and $\texttt{n}$. You should find that the posterior is:

$$p(\texttt{f}|\texttt{obsn},\texttt{n}) = \frac{(\texttt{n}+1)!}{\texttt{obsn}!(\texttt{n}-\texttt{obsn})!}\texttt{f}^{\texttt{obsn}}(1-\texttt{f})^{\texttt{n}-\texttt{obsn}}.$$

Use JAGS to verify this result with $\texttt{a} = 1$, $\texttt{b} = 1$, $\texttt{obsn} = 2$, and $\texttt{n} = 10$.

Exercise 7

Neutrino mass, from D'Agostini (2003), see Sect. 4.1. Analytically derive the posterior for the neutrino mass example and check that with a uniform prior $\mathtt{m} \sim \mathtt{dunif}(0,30)$, the result is

$$p(\mathtt{m}|\mathtt{obsm}) = \frac{19.8}{\sqrt{2\pi}\sigma} \exp - \frac{(\mathtt{m} - \mathtt{obsm})^2}{2\sigma^2}.$$

References

G. D'Agostini. *Bayesian Reasoning In Data Analysis: A Critical Introduction*. World Scientific Pub Co Inc, 2003.

R. De Propris, S. P. Driver, M. Colless, M. J. Drinkwater, J. Loveday, N. P. Ross, and et al. An upper limit to the dry merger rate at $< z > \sim 0.55$. *The Astronomical Journal*, 139:794–802, 2010.

L. Montanet, K. Gieselmann, R. M. Barnett, D. E. Groom, T. G. Trippe, C. G. Wohl, and et al. Review of particle properties. *Physical Review D*, 50:1173–1826, 1994.

Spiegelhalter, D., Thomas A., Best N., Gilks W., BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5, 1996 http://www.mrc-bsu.cam.ac.uk/bugs/documentation/Download/manual05.pdf

Chapter 5
The Prior

In Bayesian statistics (and in human life), our conclusions depend both on the data (the event itself) and on what we already know before the event occurred (prior information). For example, when we draw with replacement ten cards from a full deck, getting ten kings is considered good luck if cards are randomly extracted from a regular deck of cards. This is not a surprising outcome if, however, the pack of cards only contain kings, or is considered a card trick if the draws were done by a conjurer.

For the three different card-deck scenarios just given, the data ("ten kings in ten draws") are unique. On the other hand, our conclusions (what we claim is truly going on) depend on the other (prior) information too (i.e., what we believed about who is drawing the cards and the state of the deck of cards before we observed the ten kings). Bayes theorem quantifies the above scenarios and will return different (posterior) inferences for the same data because of the different prior information. Therefore, the prior matters in inference.

In this chapter, we first show an example where ignoring the prior can lead to erroneous conclusions. Then, we illustrate with examples the role of the prior. We also provide some advice on where to look for information useful for sculpting the prior. Finally, we conclude this chapter by illustrating two more real examples where the prior is essential.

5.1 Conclusions Depend on the Prior …

5.1.1 … Sometimes a Lot: The Malmquist-Eddington Bias

As in the conjurer example, conclusions may heavily depend on the prior so much that ignoring the prior may lead one to be wrong most of the time, as we now illustrate with an example common in astronomy. In this example, the prior drives the posterior mean away from the data (the observed value).

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_5

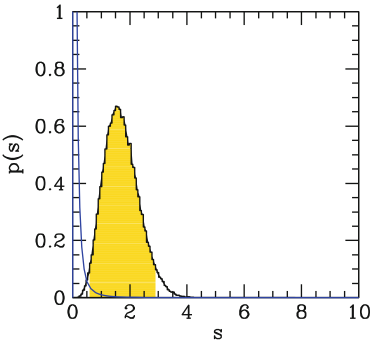

Fig. 5.1 Prior (*blue curve* diverging toward zero) and posterior (*solid curve*) probability distribution of the source flux, having observed four photons and knowing that the source number counts have a euclidean distribution. The highest posterior 95 % interval is *shaded*

Astronomers have known for about 100 years that when observations are noisy the true value `s` of the source's intensity is systematically smaller than the observed value `obsn`, regardless of the type of source they are considering. This conclusion comes after having re-observed many sources in much better conditions and having found that in general `s` $<$ `obsn` most of the time. Why does this happen? In the Universe there are many faint sources for each bright source. Because of this uneven distribution of fluxes and of the presence of noise, there are more sources whose flux is scattered up by noise than scattered down. Therefore, a source with `obsn` is likely a fainter source with an overestimated flux. This knowledge is the astronomer's prior: before measuring the source's flux, the probability that the source has a flux `s` is not a uniform distribution of `s` but a steep function of `s`.

Let us consider a quantitative example, tailored around the X-Bootes survey (Kenter et al. 2005), which is a survey of X-ray sources performed with the Chandra Space Telescope. As in Sect. 4.2.1, we continue to assume that counts are Poisson distributed,

```
obss ~ dpois(s) .
```

For past and deep (precise) observations, however, we know that the probability distribution of log `s` (called number counts by astronomers and prior by statisticians) is a power-law with slope -2.5, which is also adopted by the authors. Figure 5.1 shows this prior, apart from other numerical constants (irrelevant to the statistical computation, but of astronomical importance). A possible implementation of this prior is

```
s <- pow(tmps, -0.66666666)
tmps ~ dunif(0,100) .
```

Here, we used the fact that if a quantity `tmps` is uniformly distributed, $\texttt{tmps}^{-1/1.5}$ is distributed as a power-law with slope -2.5 (in case the reader is not aware, this is also known as a Pareto distribution). Our model now reads:

```
model {
obsn ~ dpois(s)
s <- pow(tmps, -0.666666666)
tmps ~ dunif(0,100)
}.
```

We let JAGS do the computational work of numerically sampling the posterior, leaving the mathematically skilled readers the exercise of computing the (known) analytic formula.

Consider observing a source of `obsn` = 4 photons, a typical value for sources observed in the X-Bootes survey. The posterior distribution has its mode at `s`= 1.5 and its mean at `s`= 2.5 (Fig. 5.1), in agreement with the astronomer's experience that the true flux, `s`, of a source is likely lower than the observed value `obsn`.

The conclusion is that by taking the observed value `obsn` = 4 as an estimate of the source flux `s`, the pragmatist will be wrong most of the time by systematically overestimating `s`: a better re-observation of a 4 photon source will discover it will be, on average, a 2.5 photon source. The pragmatist, refusing to use any information other than the data being analyzed, is not able to profit from past observations showing the power-law nature of the distribution of source fluxes. The Bayesian benefits from the knowledge of the steep number counts (by encoding this information in the prior) and by using both the prior and the data is able to draw conclusions in agreement with better re-observations. In this example, the prior shows such a large gradient that the posterior mean significantly differs from the observed value, i.e., plays a major role in estimating the source's flux.

The posterior's sensitivity to the prior is routinely used in cosmology, where no experiment measures a (cosmological) parameter per se, but only some combinations of these parameters. In order to break this multiple dependency among the parameters, researchers can combine results from different experiments, i.e., assume the posterior of a past experiment as a prior for the current experiment (this is known as sequential Bayes, the accumulation of knowledge through experimentation).

It should be noted that for those times that the prior probability displays a large variation in the range where the likelihood is different from zero, one should expect a Malmquist- or Eddington-like effect (after the names of the astronomers who explained why such an effect occurs), i.e., that the prior matters. This effect is the manifestation of the phenomenon that by ignoring the prior one may be wrong most of the time.

5.1.2 ... by Lower Amounts with Increasing Data Quality

The impact of the prior decreases with increasing data quality, as we now illustrate by revisiting the four counts example of Sect. 4.2.1. In particular we compare the posterior using two different priors and by increasing the amount of data.

We assume now that the researcher, before running the experiment, thought that the source `s` is 100 counts plus or minus 33 %. We mathematically quantify the

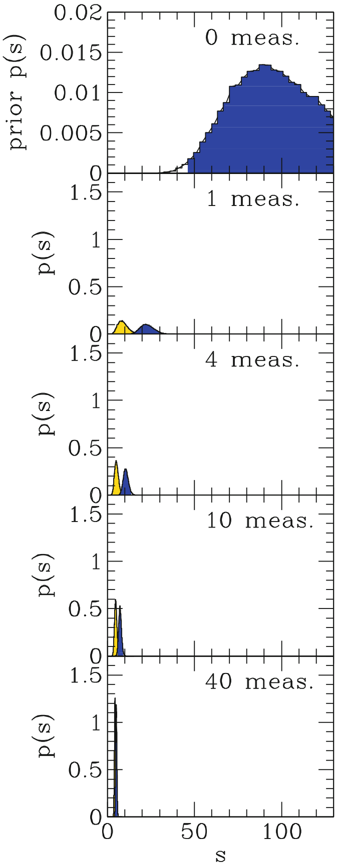

Fig. 5.2 Prior (*top panel*) and posterior probability distributions for the Poisson signal example for a log-normal prior (*dark/blue shading*) or a uniform prior (*light/yellow shading*)

stated prior by taking a log-normal distribution. The first argument of a log-normal is the log of the median ($4.605 = \ln(100)$), and its second is the precision ($1/0.33^2 = 10$) of the log of `s`. The top panel of Fig. 5.2 shows this prior. The model with the log-normal prior reads:

```
model {
for (i in 1:length(obsn)){
obsn[i] ~ dpois(s)
}
# lognormal prior on s
s ~ dlnorm(4.605,10)
}.
```

Figure 5.2 plots the posterior thus obtained (dark/blue shaded), and the posterior obtained in the previous model when a uniform prior was adopted (light/yellow shaded). Obviously, when one observes just a single measurement (second panel from the top), i.e., when the data are poor, the two posteriors widely differ because the two priors widely differ. When the quality of the data increases (lower panels), however, the two posteriors become more and more similar. This occurs because what matters in parameter estimation is the shape of the prior in the range where the likelihood is significantly non-zero and this range becomes narrower and narrower as the quality of the data increases. Every prior, without singularities, can be approximated by a constant in a sufficiently small range. So it does not matter if we started with a constant as a prior or if we end up with such a narrow posterior that the prior may be approximated by a constant, the role of the prior is decreasing with increasing data quality.

This example also offers us the possibility of reassuring some of our colleagues with very little experience with Bayesian methods. Some of them worry about the possibility of specifying a "wrong" prior, a prior that is in tension with the data or the posterior, as in the above example: a prior centered on 100 counts when the observed values are of the order of 4 counts. As far as the prior correctly matches the researcher's idea before performing the experiment, the prior is "right," even if the experiment will return a fairly different number. Actually, these are the most exciting results, when we found something different from our prior expectations. This discussion should clarify that a researcher should not refrain from using the Bayesian way because he/she worries about the risks of specifying a "wrong" prior. A prior's correctness is not established by how it matches the true value (or the outcome of the experiment), but how it matches the available information before the experiment. The only "wrong" prior is one that is sloppily chosen, say a prior that allows positively defined quantities (as mass, or dispersion) to take negative values, i.e., those formulated so sloppily that they do not match the intent of who is formulating them.

5.1.3 … but Eventually Becomes Negligible

When the data constrain conclusions very tightly, priors do not play a major role in drawing conclusions, as we now illustrate using the dry merging rate (the fraction of passive galaxies) example of Sect. 4.2.3. In that example, we assumed a uniform prior for the fraction `f`. Which role is played by the prior?

Recall that Bayes theorem tells us that the posterior is given by the product of the likelihood and the prior (Sect. 2.5). Figure 4.5 shows that the posterior is significantly different from zero only in the small range $0 < \mathtt{f} < 0.005$. Since the prior has been selected so that the density values are constant in the range [0,1], the likelihood is equal to the posterior, except for a normalizing constant. Hence, the likelihood is non-zero only in the small range $0 < \mathtt{f} < 0.005$. Therefore, the behavior of the prior outside this range matters very little, given that it is multiplied by values (likelihood) very close to zero. Furthermore, in order to find a noticeably different posterior we need a prior with large variation across the range $0 < \mathtt{f} < 0.005$, one with an almost singularity in this range. If the presence of a singularity can be excluded (and it can in this example) then the behavior of the prior in the $0 < \texttt{f} < 0.005$ range matters only a little to the overall conclusion of the experiment: the dry merging rate (i.e., `f`) is extremely low. The lesson to be learned from this example is that if the likelihood is sharp and is non-zero in only a small range of the parameter space, then the prior plays a little role, provided it is not allowed to take widely different values in this small parameter range. Of course, our quite general statement should be verified for the given data (e.g., by adopting a different prior) and cannot be taken for granted in every possible situation one may imagine.

5.1.4 ... and the Precise Shape of the Prior Often Does Not Matter

In the previous section, we introduced the concept of how the shape of the prior influences the posterior. Here we describe this a bit more. Note that what really matters is the general shape of the prior and not its precise shape. This can be analytically understood by noting that the posterior mean is given, through the use of Bayes theorem, by the average of the likelihood, weighted by the prior. In formulae:

$$E(\theta|data) \propto \int \theta p(data|\theta)p(\theta)d\theta \ . \tag{5.1}$$

Those that instead prefer numerical examples may consider the following illustration.

We are still interested in measuring the source flux of a faint source with 4 observed photons. In order to understand the effect of the precise shape of the prior, we consider two mathematically different priors, but at the same time grossly similar in shape: a uniform distribution versus a Gaussian distribution with a large sigma. These priors are depicted in the top panel of Fig. 5.3. Since flux is a positively defined quantity, negative values are unphysical. Therefore, we set a zero prior for negative values to keep a connection with reality. Priors are zero-ed for negative values of `s` by truncating them with the JAGS function `T(0,)`. Alternatively, we may use the mathematical `abs()` function.

The model now reads (priors are listed and should be uncommented in turn when the program is run) :

```
model {
# likelihood
obsn ~ dpois(s)
# Gauss prior with large sigma
#s ~ dnorm(0,0.01) T(0,)
# uniform
#s ~ dunif(0,1.0e+7)
}.
```

The bottom panel of Fig. 5.3 shows the posterior distribution computed for the two priors. Clearly, the exact shape of the prior plays a negligible role: differences between posteriors are so tiny that we are able to note that there are two lines plotted

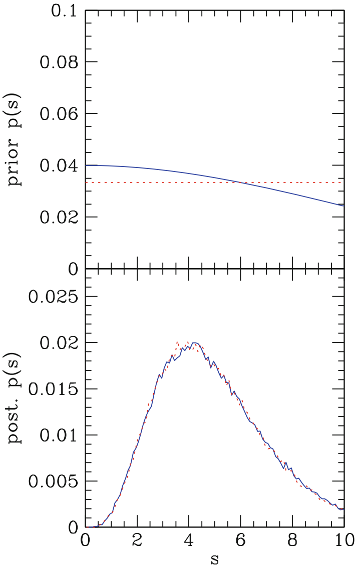

Fig. 5.3 *Top panel:* Gaussian (*solid blue line*) and uniform (*dotted red line*) priors. *Bottom panel:* Posterior probability distribution for the Gaussian (*solid blue line*) and uniform (*dotted red line*) priors. Very narrow bins are adopted for the histograms in the bottom panel, to emphasize numerical noise and hence shows that two histograms, otherwise indistinguishable, are plotted

in the bottom panels only because of numerical noise associated with our finite sampling of the posterior. Therefore, what matters is the gross shape of the prior. This is a reassuring fact because we typically do not possess a precise knowledge of the exact shape of the prior unless the prior is the posterior of a previous experiment. What really matters is to specify the prior's gross behavior.

5.2 Where to Find Priors

At this point, the reader should be convinced that conclusions depend on the prior by an amount that varies from case to case.

The next logical question to ask is how does one determine the prior? Said simply, it should match the researcher's ideas before performing the experiment, i.e., what we know about the quantity being measured other than the information coming from the considered experiment.

Our experience as scientists tells us that researchers in physics or astronomy have some sort of prior knowledge and that we rarely know nothing about the "rough size" of what is about to be measured. If at all, some of us just do not know that prior is the name for what is already known about the quantity being measured.

Researchers get access to experimental facilities by competition. This usually entails writing a facility request and a portion of this request is dedicated to what is called a "feasibility section." The feasibility section requires that researchers quantify their expectation about the size of the quantity they want to measure. This is the prior. This is also essential for evaluating if a result will likely occur: if the measurand is likely outside the range probed by the instrument, the experiment may result in a waste of money. Therefore, it is usually requested that researchers quote a plausible range of where they expect the quantity to reside. This is the prior. The prior is also needed to select the most appropriate instrument for the experiment, its configuration, and how long the experiment should run (exposure time).

Even if no facility request has been written, conscious scientists do not waste their time trying to measure an effect at a facility that does not offer the sensitivity/resolution requested for the measurement. Even these scientists have a prior, they know that the measurement is, likely, in a given range, and, less likely, outside it. This is their prior.

Therefore, rarely do researchers not have some sort of prior. On the contrary, when they decide to use their own time to perform a study, they already know that the quantity being measured is more likely to be in some range rather than outside it.

As detailed in Chap. 9, irrespective of the adopted prior, it is always opportune to measure the impact of the prior on conclusions.

5.3 Why There Are So Many Uniform Priors in this Book?

In the following chapters we will not provide much discussion about the prior as we have done here, and we will often use uniform priors. This should not be taken as a suggestion to use uniform priors in real problems. We (will) use uniform priors only because the prior depends on the scientific problem under study, and we cannot imagine for which application a model will be used!

5.4 Other Examples on the Influence of Priors on Conclusions

5.4.1 The Important Role of the Prior in the Determination of the Mass of the Most Distant Known Galaxy Cluster[1]

Andreon et al. (2009) calculated the mass of the most distant cluster of galaxies, JKCS041 at $z = 1.8$, and found an observed (log) mass of $obslgM = 14.6 \pm 0.3$ $M_\odot$ (units are not longer quoted). In our notation, `obslgM` $\sim$ `dnorm(lgM, pow(0.3,-2))`, and `obslgM <- 14.6`.

[1] This section can be skipped the first time through.

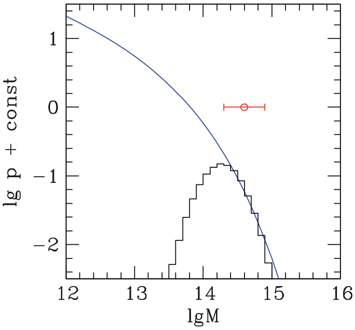

Fig. 5.4 JKCS041 mass estimate. The *blue continuous line* depicts the prior (named mass function by astronomers), the *red point* indicates the observed value, and the histogram represents the posterior probability distribution. Note the log scale on the ordinate and the effect of the steep prior, that moves the most probable value of the cluster's mass to values smaller than observed

What is the JKCS041 mass? This question seems odd because we have just mentioned its value! Why not simply use the observed value? Because, as the reader by now suspects, the prior (named the mass function by astronomers) changes significantly in the range where the likelihood is not negligible, as shown in Fig. 5.4. Note the adoption of a log scale on the ordinate because of the very large change in the prior's value (more than four orders of magnitude in the range of interest). Using the observed value leads to incorrect inferences about the likelihood of living in a ΛCDM Universe (Mortonson et al. 2011; Andreon et al. 2009), because wrong probabilities of observing clusters like the ones we observe would be inferred.

The theoretically predicted mass function, $p(lgM)$, at the cluster's redshift, $z = 1.8$, has been computed several times with several approximations in various astronomical papers. For the sake of clarity, we take the Jenkins et al. (2001) mass function, that we approximate with a Schechter function (Schechter 1976) with slope `alpha=-1.8` and `lgMstar=13`. This is our prior and we generate it (quite verbosly) using JAGS as described in Sect. 3.2. Our model (quite verbose, all but one single line is used to implement the Schechter prior in JAGS) reads:

```
data {
zeros <- 0
C <- 10
}
model {
# likelihood
obslgM ~ dnorm(lgM,pow(0.3,-2))
# prior, using the zero trick
lgM ~ dunif(12,18)
phi <- -2.3025*0.4*(alpha+1)*(lgM-lgMstar)+10^(0.4*(lgM-lgMstar))
zeros ~ dpois(phi+C)
}.
```

The posterior probability of the cluster's mass is depicted in Fig. 5.4. Its mean and standard deviation are 14.3 ± 0.3, lower than the observed value 14.6 (by a factor 2 in mass, or 0.3 dex). This shift is due to the important variation of the prior in the range where the likelihood is significant (more than a factor of 10 in the range [14.3,14.9], i.e., in the 1σ range centered on the observed value). Astronomers explain this shift by saying that since there are a lot of clusters of small mass, there are more clusters with overestimated mass than massive clusters with underestimated masses, and therefore the error (scatter) pushes more systems up than down. This is the very same Malmquist bias discussed in Sect. 5.1.1 for the source flux. While stated in this way, the "bias" is obvious (at least to astronomers), yet the "bias" was forgotten in at least four papers in 2010 dealing with high redshift massive clusters, in spite of the warning in Andreon et al. (2009), making a wrong inference about the likelihood of the current cosmological model (Mortonson et al. 2011).

5.4.2 The Importance of Population Gradients for Photometric Redshifts[2]

We have already emphasized in previous sections the importance of accounting for the prior when it appreciably changes in the range where the likelihood is non-null. In Sect. 5.1.1, we have illustrated the case of the flux and in Sect. 5.4.1 the case of mass; we now illustrate the case of redshift, echoing what was been previously presented. Recall that $p(s)$ was the specified prior for both the object's intensity `s` and the distribution (abundance) of objects of intensity `s`. Similarly, $p(\texttt{zspec})$ is both the prior on the redshift `zspec` and the distribution (abundance) of objects of redshift `zspec`.

Determining a galaxy's redshift (distance) by spectroscopy is an expensive task on a telescope and faint galaxies are often too dim to get their redshift by spectroscopy. Instead, one can use the galaxy's color (or even better, the galaxy's spectral energy distribution) because it exhibits a tight correlation with the redshift and can be collected in a reasonable amount of telescope time, even for faint galaxies. In such a way, we can get an estimate of the cluster's redshift: its photometric redshift `zphot`.

Figure 5.5 shows a quite common plot of `zphot` against the spectroscopic redshift `zspec`. The latter is taken from Le Fèvre et al. (2005) while the former, `zphot`, has been determined with EAZY (Brammer et al. 2008). The bottom panel shows residuals, given by $(\texttt{zspec} - \texttt{zphot})/(1 + \texttt{zphot})$. A trend in the residuals is clearly evident, and astronomers tend to work hard to eliminate this trend, either by fitting the residuals with a smooth function and using it to correct `zphot` or by modifying the `zphot` estimation until the effects disappear. The rationale is that some sort of systematic problem is present and should be corrected.

The right panel shows the distribution in spectroscopic redshift, in small bins, $p(\texttt{zspec})$, i.e., the prior $p(\texttt{zspec})$. It is clearly not uniformly distributed.

[2] This section can be skipped the first time through.

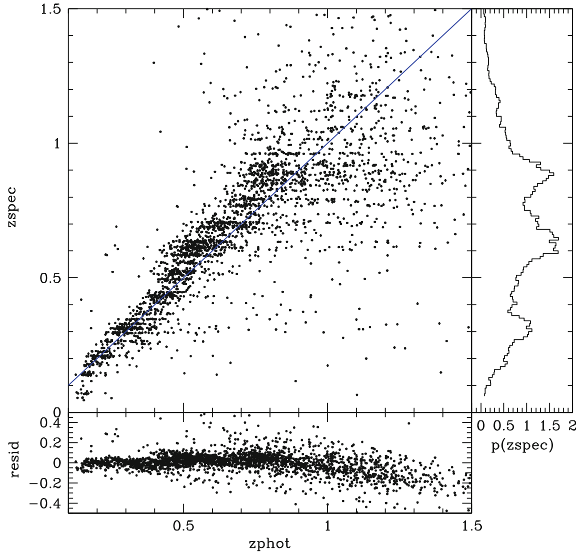

Fig. 5.5 zphot vs zspec for the Le Fèvre et al. (2005) true data (*top panel*), residuals (*bottom panel*), and redshift distribution (*right panel*)

We want to now show that the systematic trend is a manifestation of the Malmquist bias: the observed value tends to overestimate the true value because of the (neglected) population gradient (i.e., non-uniform prior, plotted in the right panel) of the spectroscopic sample. This is an important difference compared to rectifying the estimate of zphot, because if zphot is used for a fainter galaxy sample, which is usually the case, the latter will have a different redshift distribution and requires a different correction or none at all.

To explain how the systematic error arises, we consider simulated data. We first draw a large sample of zspec and zphot pairs from a Gaussian distribution centered around zphot with dispersion (σ) equal to $0.03 \times (1+\mathtt{zphot})^3$. The adopted coefficients are set to reproduce real Le Fèvre et al. (2005) data, as will be clear in a moment. Simulated data have been generated by the JAGS code:

```
model {
zphot ~ dunif(0,2.0)
zspec ~ dnorm(zphot,pow(0.03*(1+zphot)^3,-2))
}.
```

In the simulation there is no systematic effects: the mean zspec at a given zphot is equal to zphot, and the mean zphot at a given zspec is equal to zspec (except at zspec$\approx$ 0 and 2).

The simulated data above are not distributed as $p(\mathtt{zspec})$ in the Le Fèvre et al. (2005) sample. Therefore we now match the $p(\mathtt{zspec})$ distribution of this simulated sample to the Le Fèvre et al. (2005) distribution in Fig. 5.5 (right panel) by extracting from the large sample of simulated data a (random) sub-sample having the same $p(\mathtt{zspec})$ as the real data. The upper-left panel of Fig. 5.6 shows the

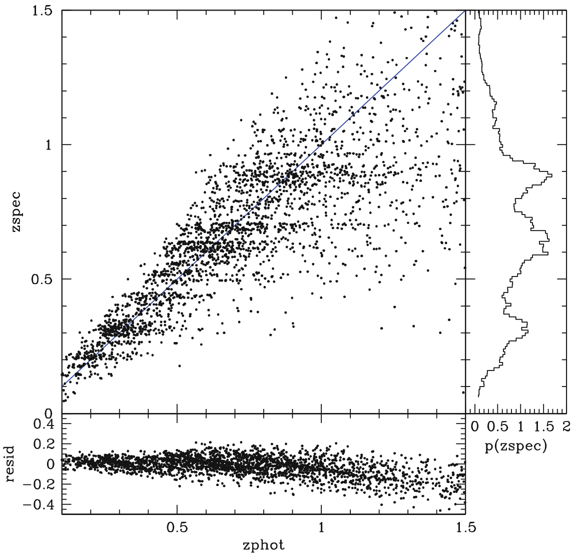

Fig. 5.6 `zphot` vs `zspec` for simulated data. Contrast it with the previous Figure

selected (`zphot`, `zspec`) pairs. If we now plot the residuals (bottom panel), a trend appears, with increasingly negative residuals as `zphot` increases. Since our `zphot` and `zspec` are correct and not biased at all, the trend visible in the residuals is due to the population gradient of `zspec` values entering in the spectroscopic sample. In fact, the population gradient in the spectroscopic redshift makes it more probable that a, say, `zphot` $= 1.4$ galaxy is a lower redshift galaxy scattered up by `zphot` errors than a `zspec` $\gtrsim 1.4$ galaxy scattered down, almost none of which is in the spectroscopic sample. This sentence should sound familiar, if we were talking about fluxes instead of redshift.

Back to the Le Fèvre et al. (2005) data set, $p($`zspec`$)$ decreases at $z > 0.9$, not because the Universe ends there, but because of the VVDS spectroscopic selection function (see Chap. 7 about how to deal with selection functions) makes the inclusion of galaxies in the spectroscopic sample less probable as the redshift increases. To be included in the Le Fèvre et al. (2005) sample, a `zspec` > 1 galaxy should be a special galaxy, either a star forming galaxy or an AGN: passive galaxies at $z > 1$ are under represented in the sample because of the Le Fèvre et al. (2005) spectroscopic selection function. Forcing an agreement at `zspec` > 1 is biasing the estimate of `zspec` of galaxies with a different a priori redshift distribution, for example faint red galaxies or galaxies fainter than the spectroscopic redshift (in both cases the median value of the redshift probability distribution is larger).

In the presence of population gradients, `zphot` and `zspec` should systematically disagree where the population has a redshift gradient unless the scatter between them is negligible (i.e., `zphot` errors are negligible).

A forced agreement between `zphot` and `zspec` is also the standard approach of empirical photometric redshift estimates (i.e., those that look for a trend between zspec and combinations of galaxy colors) and therefore these have to be considered with skepticism when applied to galaxy populations with a different a priori redshift distribution.

5.5 Exercises

Exercise 1

Suppose we observed the following data, drawn from a Gaussian with mean $\mathtt{m} = 7$ and variance $\mathtt{s}^2 = 4$: $\{16.25, 11.82, 6.59, 2.44, 11.8, 12.62, 6.18, 4.17\}$. Assume $\mathtt{m} \sim \mathtt{dnorm}(-50, .01)$ (recalling that JAGS uses the precision in place of the variance) and $\mathtt{s} \sim \mathtt{dunif}(0, 100)$.

Exercise 1a

Using JAGS, obtain draws from the posterior distribution of m and s. Plot the priors and the posteriors on the same graph.

Exercise 1b

simulate 100 more data points from the dnorm(7, 0.25) and use this new data set to obtain draws from the posterior. Plot the priors and the posteriors on the same graph.

Exercise 1c

simulate 10000 more data points from the dnorm(7, 0.25) and use this new data set to obtain draws from the posterior. Plot the priors and the posteriors on the same graph.

Exercise 1d

What do you observe about the different posteriors obtained in a)–c)?

Exercise 2

Is your flat prior truly noninformative? Consider the simple example. Assume a uniform prior to f, i.e., $\mathtt{f} \sim \mathtt{dunif}(0, 1)$, where f is the success probability in a binomial likelihood. Now suppose that for whatever reason, you are not interested in estimating `f` but rather $\mathtt{f}^2$ and $\log(\mathtt{f})$. Using simulation, plot an estimate of the density of $\mathtt{f}^2$ and $\log(\mathtt{f})$ based on the flat prior assigned to `f`. What to do you observe?

Exercise 3

Continuing with Exercise 2, the Jeffrey's prior (a class of noninformative priors meant to be transformation invariant) for `f` is given as a $\mathtt{f} \sim \mathtt{dbeta}(0.5, 0.5)$. Simulate from the beta(0.5,0.5) distribution and use those simulations to view the induced prior for $\mathtt{f}^2$ and $\log(\mathtt{f})$.

Exercise 4

Assume that $\mathtt{obsy} \sim \mathtt{dnorm}(\mathtt{m}, \mathtt{prec})$, where `m` is unknown but the precision `prec` is known, and the prior $\mathtt{m} \sim \mathtt{dnorm}(\mathtt{a}, \mathtt{b})$. Derive the posterior distribution of `m`. Compare the form of the prior to the posterior (and again gives you an idea of how conjugacy is defined).

Exercise 5

Recall that when we introduced the Malmquist-Eddington bias (Sec 5.1.1), we used a Poisson distribution as the likelihood. Some authors claim that this bias appears because of the asymmetry of the Poisson distribution. Again assume that we observed 4 counts from the source, refit the model using a Gaussian distribution (a symmetric distribution) with standard deviation of 2. Comment on your results.

Exercise 6

Returning to the Malmquist-Eddington bias section:

Exercise 6a

Verify that with the adopted uniform prior the posterior of the Poisson signal determination has mean `obss`+1 and standard deviation $\sqrt{(\mathtt{obss}+1)}$.

Exercise 6b

Change the uniform prior in the Poisson signal determination from the adopted uniform to a Jeffreys prior $(1/\texttt{s})$. Verify that the posterior is Poisson centered on `obss`.

References

S. Andreon, B. Maughan, G. Trinchieri, and J. Kurk. JKCS 041: a colour-detected galaxy cluster at $z_{phot} \sim 1.9$ with deep potential well as confirmed by X-ray data. *Astronomy & Astrophysics*, 507:147–157, 2009.

G. B. Brammer, P. G. van Dokkum, and P. Coppi. EAZY: A fast, public photometric redshift code. *The Astrophysical Journal*, 686:1503–1513, 2008.

A. Jenkins, C. S. Frenk, S. D. M. White, J. M. Colberg, S. Cole, A. E. Evrard, and et al. The mass function of dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 321:372–384, 2001.

A. Kenter, S.S. Murray, W.R. Forman, C. Jones, P. Green, C.S. Kochanek, and et al. XBootes: An X-Ray survey of the NDWFS Bootes Field. II. The X-ray source catalog. *The Astrophysical Journal Supplement Series*, 161:9, 2005.

O. Le Fèvre, G. Vettolani, B. Garilli, L. Tresse, D. Bottini, V. Le Brun, and et al. The VIMOS VLT deep survey. First epoch VVDS-deep survey: 11564 spectra with $17.5 < I_{AB} < 24$, and the redshift distribution over $0 < z < 5$. *Astronomy & Astrophysics*, 439:845–862, 2005.

M. J. Mortonson, W. Hu, and D. Huterer. Simultaneous falsification of ΛCDM and quintessence with massive, distant clusters. *Physical Review D*, 83:023015, 2011.

P. Schechter. An analytic expression for the luminosity function for galaxies. *The Astrophysical Journal*, 203:297–306, 1976.

Chapter 6
Multi-parameters Models

In this chapter we consider multi-parameter estimation problems. This chapter is split into three parts. In the first part, we address common problems such as the determination of location and spread and simple measurements of the spectral shape (slope and curvature). In the second part we consider measurements that require the modeling of two populations, for example the interesting one and a nuisance population. We then list, in the section titled "Advanced Analysis," quite complex models precisely tailored to specific experiments. The latter section can be skipped the first time through.

6.1 Common Simple Problems

6.1.1 Location and Spread

Suppose we have some measurements with Gaussian errors, and that the values scatter more than the measurement error allows. We want to measure the center of the distribution from which the data are drawn and the size of the spread attributed to the variability in the population.

To focus this idea, let us refer to a common measurement of spread in astrophysics: the cluster velocity dispersion. Galaxies, as gas particles, move with velocity vector `v` distributed as a nearly-Maxwell distribution. Astronomers are only able to measure the line of sight component of the velocity, `v`, which is (with a good approximation) normally distributed around the true value `vcent` with dispersion `sigma.clus`:

```
v[i] ~ dnorm(vcent,pow(sigma.clus,-2)).
```

As usual, every measurement is affected by noise, which we assume to be Gaussian, whose amplitude `sigma.v[i]` varies from measurement to measurement,

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_6

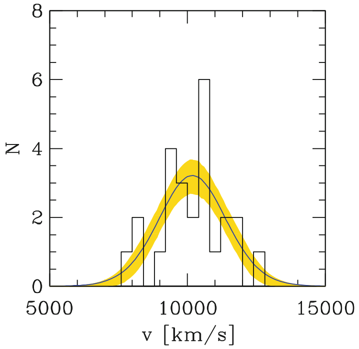

Fig. 6.1 Histogram of the (simulated) data. We measured 25 velocities of (fake) galaxies. The mean model is marked with the *continuous line*. The *shaded region* marks the 68 % interval

```
obsv[i] ~ dnorm(v[i],pow(sigma.v[i],-2)).
```

We want to know the velocity barycenter, `vcent`, and the intrinsic spread, `sigma`, called cluster velocity dispersion by astronomers.

To complete the model description we need to assign priors for `vcent` and `sigma.clus`. The prior on `vcent` is largely irrelevant in many situations because a few measurements are sufficient to determine this parameter. Instead, the prior on sigma may matter if the sample size is small (do you remember the Malmquist-like bias?).

Let us assume a uniform prior for the velocity dispersion `sigma.clus` (remember the prior chapter, do not adopt this prior as default). The model reads:

```
model{
for (i in 1:length(obsv)){
v[i] ~ dnorm(vcent,pow(sigma.clus,-2))
obsv[i] ~ dnorm(v[i],pow(sigma.v[i],-2))
}
vcent ~ dnorm(10000,1e-7)
sigma.clus ~ dunif(0,5000)
} .
```

Figure 6.1 shows (simulated) data for 25 galaxies.

After fitting the model to these 25 data points, we estimate a location `vcent`= 10180 ± 255 km s^{-1} and a dispersion `sigma.clus` of 1240 ± 195 km s^{-1} (the data were generated from a Gaussian centered in 10000 km s^{-1} and with sigma 1000 km s^{-1} as detailed at the end of this section).

Figure 6.1 shows the mean model (continuous line) and its 68 % interval, superposed to the data. Figure 6.2 shows the marginal posterior distribution of the two key parameters, location and spread. Figure 6.3 shows the joint probability distribution of location and spread. The contours are drawn at 68 % and 95 % probability, i.e., enclose 68 % and 95 % of the draws. More specifically, we consider higher posterior density contours, indicating that the region enclosed by the contour has, locally,

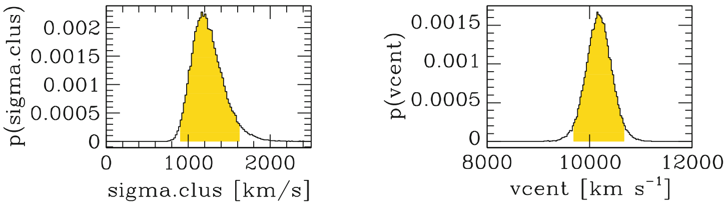

Fig. 6.2 Posterior probability for the (velocity) dispersion `sigma.clus` and central location `vcent` of the velocity dispersion problem. The *shaded region* marks the 95 % interval

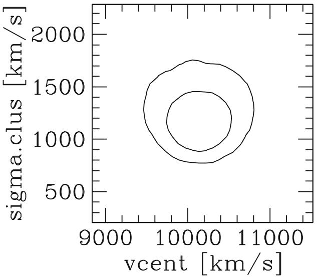

Fig. 6.3 Joint probability distribution of location and spread. Contours are drawn at 68 % and 95 % probability

a higher probability density than non-selected regions. The circular shape of the contours indicates that the two parameters are approximately not correlated.

Derived velocity dispersions have the properties of non-negativity, and their uncertainties do not include unphysical (negative or complex) values of the velocity dispersion. While the above properties seem useless to state, it should be noted that they are non-trivial properties, because unphysical velocity dispersions, derived using different methods, appear, from time to time, in astronomical papers.

The model discussed in this section is depicted in Fig. 6.4, named a Bayesian graph: arrows indicate logical links between quantities (tilde symbols in our coding), the observed values are in the bottom line, the parameters that have a prior are in the top line, whereas "intermediate" quantities (latent variables), true velocities `v[i]` in this example, are in the middle line. The graph makes clear that both parameters, `vcent` and `sigma.clus` affect the `v[i]` values, constrained by the observed values `obsv[i]`. Bayesian graphs are useful for more complex models, allowing one to visualize the complex link between the many intervening quantities.

Let us open a parenthesis for astronomers. First, all our computations derive results in the system of the provided numbers, i.e., in the observer's frame in the case of the velocity dispersion. Therefore, do not forget to remove the cosmological $(1+z)$ stretching factor if you are quoting a measurement in the cluster's rest-frame. Second, if you are dealing with noisy velocity dispersions, remember to adopt the prior appropriate for your problem, to avoid what we called the Malmquist bias (see Sect. 5.1.1), i.e., a systematic overestimate of the true value of the velocity dispersion.

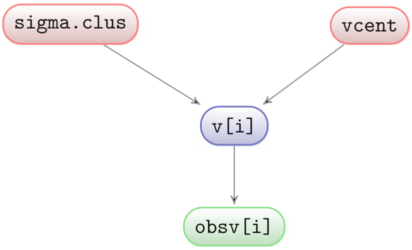

Fig. 6.4 Graph showing the stochastic relationship between quantities involved in estimating location and spread

The simulated data used in this example were generated with the model:

```
model{
prec.v ~ dunif(pow(300,-2),pow(80,-2))
v ~ dnorm(vcent,prec.clusdisp)
obsv ~ dnorm(v,prec.v)
vcent <- 10000
prec.clusdisp <- pow(1000,-2)
}.
```

As the model indicates, we generated heteroscedastic errors (velocity measurements have different uncertainties, from 80 to 300 km/s). Because draws within a chain are not independent (Chap. 3), remember to decimate, or centimate them, if you use this model to generate fake data.

The model discussed in this section deals with Gaussian-distributed quantities with Gaussian errors, but these assumptions can be easily modified if the data behave differently. If, for example, the modelled quantities are, say, Student-t distributed, then a Student-t distribution, `dt` should be used in place of the normal `dnorm` in the model.

A numerical comment is needed: when performing numerical computations, it is good practice to feed programs with reasonable numbers (for example to avoid numerical overflows). Velocities in astronomy can be huge numbers. It may sometimes be useful to remove a round number near the data average before passing data to JAGS (and add this number back after stochastic computation).

6.1.2 The Source Intensity in the Presence of a Background

Often we are confronted with measuring the flux of a source superposed on a background. This very same problem may take other forms, for example, it may consist of measuring the richness (number of galaxies) of a cluster of galaxies projected on the top of background galaxies. Or, in a nuclear reaction, the number of particles satisfying some properties (e.g., compatible with being Weakly Interacting Massive Particles) or, in particle physics, the number of neutrinos produced by the Earth's

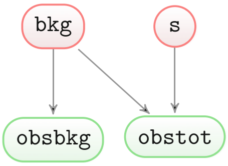

Fig. 6.5 Graph showing the stochastic relationship between quantities involved in estimating the signal in the presence of a background

radioactivity. In all these cases, and in many similar others, we have to infer the value of an interesting signal from noisy measurements contaminated by a background or by other sources. From now on, we call background everything except the signal of interest.

Because of various reasons, observed and true values are not identically equal. The variables `s` and `bkg` represent the true signal and true background in the studied solid angles. In these solid angles, assumed to be perfectly known, we measured `obstot` and `obsbkg`, respectively. The solid angle ratio (or the ratio between the time the background and the source are observed) is `C`. We consider the case in which observed values are Poisson distributed.

In formulae, our wordy statements can be summarized by:

```
obstot ~ dpois(s+bkg/C)
obsbkg ~ dpois(bkg).
```

For the time being, we take for the signal `s` and for background `bkg` independent uniform priors in the physically acceptable range (i.e., positive):

```
s ~ dunif(0,any_large_value)
bkg ~ dunif(0,any_large_value).
```

In summary, the statistical model reads:

```
model {
obstot ~ dpois(s+bkg/C)
obsbkg ~ dpois(bkg)
s ~ dunif(0,1.0e+7)
bkg ~ dunif(0,1.0e+7)
},
```

where we took 1.0×10^7 as `any_large_value`, because the numbers considered in our examples are much smaller.

Note that the model is more general than stated: the quantity named `C`, which we called the solid angle can be other things as well, depending on the experiment we are modeling. If we were running a nuclear experiment, `C` would be the ratio between the times used for signal and background determination (e.g., a large `C` means that the experiment is run for a long time with the source switched off).

Figure 6.5 summarizes the stochastic links between the involved quantities of our model. The model is quite simple, `obsbkg` is only affected by `bkg`, while `obstot`

Table 6.1 Data for the signal on background example: posterior mean and standard deviations are also reported

Case	obstot	obsbkg	C	bkg	s
a)	31	1	1	2.0± 1.4	30.0± 5.8
b)	60	30	1	31.0± 5.5	30.0± 9.5
c)	60	1	1/30	33.3±16.1	28.6±16.5
d)	60	3000	100	30.0± 0.5	31.0± 7.8
e)	170	172	1	165.0±10	14.0±10

depends on both parameters bkg and s. The lack of an intricate network (arrows) makes this example quite simple.

Let us consider a number of possible observations, summarized in Table 6.1:

a) We observed 31 counts in the source direction vs. one count observed in the background area (obstot= 31, obsbkg= 1) of the same solid angle as the source solid angle (C= 1).
b) We observed 60 counts toward the source, and 30 from the background. Solid angles are equal (C= 1). Net counts, computed as in the butcher shop, is total minus tare, obstot-obsbkg/C, are equal to those in the first example, 30 net counts, but the background level is higher than in previous example.
c) A small C. We observed 60 counts in the source direction and 1 in the background direction, measured, however, on a solid angle 30 times smaller than the one on which the source is measured (C= 1/30). This example differs from the previous example only by the fact that the background is measured on a smaller area, although the net counts computed as in the butcher shop, obstot-obsbkg/C is the same, 30 net counts. This example resembles the situation where there is a noisy background determination, either because the solid angle available for the background measurement is small, or because it is expensive (or difficult to realize) to have a measurement with the source switched off. This is the typical case when we cannot switch off the signal source, as for weakly interacting massive particles, or for neutrinos coming from the reactions in the Earth's core.
d) A large C. We observed 60 counts in the source direction and 3000 in the background direction, measured however on a solid angle 100 times larger than the one where the source is measured (C= 100). This example differs from example b) only by the fact that the background is measured on a larger area, although the net counts, again computed as in the butcher shop, obstot-obsbkg/C are the same, 30 net counts. This example resembles situations where the measurement of the background is less costly than the measurement of the signal and for this reason, a large C can be taken (or the experiment is run a lot without the source signal).
e) Let us consider the cluster of galaxies XLSSC 12 (Andreon et al. 2006). In this example, solid angles are equal (C= 1) and the total galaxy counts in the source direction is lower than the background value. This may occur as a result of back-

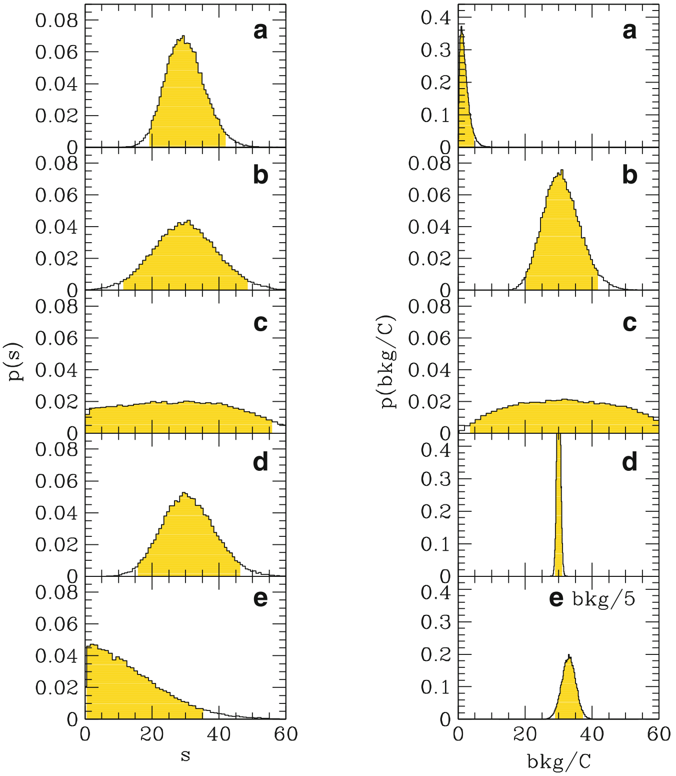

Fig. 6.6 Probability distribution of the signal s and background per unit solid angle bkg/C for the signal plus background example. The *shaded region* marks the 95 % probability interval

ground fluctuations. The net number of galaxies, computed as in the butcher shop, is obstot-obsbkg, which is negative ($170-172=-2$) although there is no doubt about the fact that the number of galaxies comes in positive units (have you ever seen -1 galaxy?), and there is no doubt either about the existence of this cluster, because there are two independent, and unambiguous astronomical observations: the detection of X-ray emission from the intracluster medium, and spectroscopic observations of 12 galaxies belonging to XLSSC 12 (Andreon et al. 2004). Actually, if one just counts red galaxies (for which background counts are lower and hence lower Poisson fluctuations), the cluster has a positive number of net galaxies, as all clusters do. In summary, the butcher shop rule for computing the net weight (number of galaxies) seems to not work in this example.

Figure 6.6 shows the posterior distribution of the signal s and of the average background in the source solid angle, bkg/C for these five examples. In the panels, the highest posterior 95 % interval is shaded.

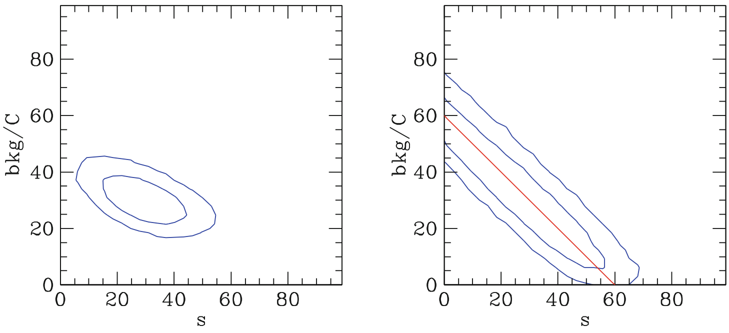

Fig. 6.7 Joint 68 % and 95 % probability contours for the signal plus background examples in the case b) (*left panel*) and c) (*right panel*). In the right panel contours stop at s= 0 and bkg/C= 0, even if not obvious from the plot

Let us start by commenting on the background posterior distributions. In case a) the true background is low, a few counts at most, because the observed value of the background is one count only. Cases b), c), and d) have similar values of bkg/C, about 30 counts, but in the last case the true value of the background is sharply determined: it could be only about half unit away from 30 (bkg/C= 30 ± 0.5), thanks to the large solid angle used to determine it. Case c) is the opposite of case d), the true value of the background is poorly determined, because only a small solid angle (C) is available. Case b) is intermediate between these two extremes (bkg/C= 31 ± 5.5).

The signal's posterior of the first example is bell-shaped, with an average of 30 and $\sigma = 5.8$. In the second example, the higher background makes the width of the posterior wider, $\sigma \sim 9.5$. In case c), the background is measured on a tiny solid angle, making the average value of the background very uncertain. For this reason, the posterior is much wider than in case b), $\sigma = 16.5$ (and with a non-Gaussian shape). If instead the background is measured on a wide solid angle, as in case d), the signal's posterior is narrower, $\sigma = 7.8$, because the average background is well known, and this source of error becomes negligible.

Finally, panel e) shows the case of the cluster XLSSC 012: the posterior has no maximum (in the mathematical sense, i.e., a point of zero derivative inside the range of acceptable values, with a negative second derivative). It has an extremum at the boundary of the acceptable range. The net number of galaxies, i.e., the central location of the s posterior, is poorly characterized by any single number (e.g., the median or mean), but it is positive no matter which summary is used (the mean and median are 14 and 11, respectively), fairly different from the butcher rule, which returns a negative number of galaxies. The (highest posterior) 95 % interval is found to be [0,35].

Figure 6.7 shows 68 % and 95 % probability contours for the cases b) and c). In case c), the data actually constrain bkg/C+s to be about 60, because obsbkg is very noisy. Therefore contours of equiprobability are aligned along the line bkg/C+s= 60, plotted in red. In case b), the correlation between bkg/C and s arises from the fact that a larger value of s has to be somewhat compensated by a lower value of bkg/C to keep s+bkg/C near to obstot. The correlation is mitigated, in this case, by the other datum, bkg cannot be too different from obsbkg/C.

The uncertainties computed with this model are preferable to those derived assuming Gaussian errors on counts propagated on the source signal: at the very least they do not extend outside the range of physically acceptable values, and can be derived even in absence of a clean signal detection.

In all of the examples, we keep the interval including 95 % of the posterior probability to compute the 95 % interval. In particular, we do not use different recipes for computing the interval depending on whether the source is detected. In fact, we have one single computer code to compute the 95 % interval, also used for the signal of case e), where the signal is manifestly not detected. In the Bayesian approach, in order to compute a probability interval, we do not need to state whether the signal s has been detected. This property is welcome, because there are cases where the signal detection is ambiguous (imagine a source marginally detected if any at all), and the Bayesian way does not oblige the researcher to state what he/she ignores, i.e., whether a signal is there. This is an advantage compared to other methods, sometimes used in physics, that determine the confidence interval using two different recipes, one in the case the signal is (claimed to be) detected, and one in the case it is not, forcing the researcher to assume, for sure, what is intrinsically uncertain. In passing, the upper limit computed under the two assumptions turns out often to be different.

6.1.3 Estimating a Fraction in the Presence of a Background

We now want to estimate a fraction, but with a slight twist from Sect. 4.2.3. Here we allow our data to be contaminated by a background, i.e., by some unwanted events. This is the typical case in astronomy and physics where we want to infer a property of a population contaminated by another population not of interest (called here background, as in the signal plus background example). The aimed fraction takes different names, e.g., blue fraction or branching ratio. Background properties are often measured on a sample with the signal switched off, or along a (reference) line of sight uncontaminated by the objects under study. For example, one may want to infer the fraction of blue galaxies in clusters from an observed sample formed by galaxies both in the cluster and outside it, estimating the latter from a control field sample. An identical situation occurs in nuclear physics where the determination of a branching ratio is contaminated by a particle background. Let us focus on the former case, for clarity.

We observed obsnbkg galaxies in the reference line of sight, and obsntot in the cluster's line of sight. These two quantities are Poisson distributed. In the two directions, we observed obsbluebkg and obsbluetot blue galaxies. These two quantities are binomial distributed. Perhaps the most interesting quantity is the blue fraction fclus. To derive it, we use (but this is not compelling) an auxiliary

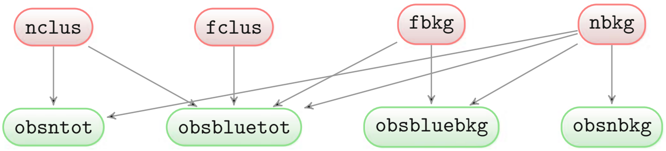

Fig. 6.8 Graph showing the stochastic relationship between quantities involved in estimating a fraction in the presence of a source of contamination

variable `f`, the blue fraction in the cluster direction, that can be analytically derived from the two true blue fractions `fbkg` and `fclus` as:

```
f <- (fbkg*nbkg/C+fclus*nclus)/(nbkg/C+nclus) ,
```

where `C` is the ratio between the solid angles on which the number of galaxies have been counted.

The problem is summarized in Fig. 6.8. The interesting quantity `fclus` only affects `obsbluetot`, but the latter is affected by all other parameters.

We take uniform priors for all variables: a `beta(1,1)` for fractions, and a uniform for values above 1 (and up to a large number approximating infinity in our computer) for the number of galaxies in the cluster and reference lines of sight. The model reads:

```
model {
obsnbkg~dpois(nbkg)
obsbluebkg~dbin(fbkg,obsnbkg)
obsntot~dpois(nbkg/C+nclus)
obsbluetot~dbin(f,obsntot)
f <- (fbkg*nbkg/C+fclus*nclus)/(nbkg/C+nclus)
nbkg ~ dunif(1,1e+7)
nclus~ dunif(1,1e+7)
fbkg ~ dbeta(1,1)
fclus ~ dbeta(1,1)
}.
```

In short, we “merged” the model of the signal+background determination and the model of the fraction determination. This high modularity is one of the advantages of the Bayesian approach: once the full problem is broken down into smaller parts and these parts are modeled, the full problem is largely solved too.

Let us start from a simple case: we observed 10 blue galaxies out of 10 in the reference line of sight, and 90 blue galaxies out of 170 in the cluster’s line of sight. Using the butcher rule (6.1), we obtain a fraction of $(90-10)/(170-10) = 50\,\%$. The model above returns 0.50 ± 0.04.

Let us now consider the (difficult) case of XLSSC 012, the cluster mentioned in Sect. 6.1.2 with more galaxies in the reference line of sight than in the cluster’s direction. The observed values are: `obsntot`= 170, `obsbluetot`= 78, `obsnbkg`= 2,861, and `obsbluebkg`= 1,625 with `C`= 16.61 (we consider here a larger solid angle for the reference line of sight than in Sect. 6.1.2). Note that we observed almost identical number of galaxies in the cluster and reference line of sight

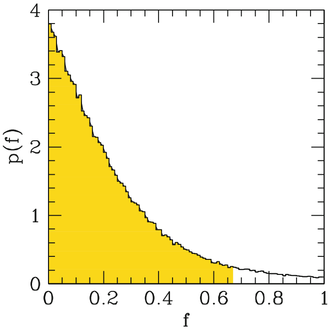

Fig. 6.9 Posterior probability of the fraction `f` of blue galaxies in a cluster in the difficult case of XLSSC 012 described in the text. The *shaded region* marks the 95 % (highest posterior) interval

per unit solid angle (i.e., `obsntot`≈ `obsnbkg/C`), but largely different numbers of blue galaxies (78 vs ≈ 98 expected =`obsbluetotbkg/C`). Blue galaxies are in deficit toward XLSSC 012 and therefore non-blue galaxies are in excess. The butcher rule (Eq. 6.1 detailed below) returns as an estimate of the fraction the value $(78-98)/(170-172) = 10$ which is non-sense (and even more so noticing that it is derived from the ratio of two negative numbers, when instead both are positively defined). The model above returns instead the posterior plotted in Fig. 6.9, a result in agreement with good sense: the fraction is small because there is a deficit in the observed number of blue galaxies in the cluster's line of sight, whereas a large blue fraction implies larger observed values of blue galaxies.

Note that a naive analysis may have concluded that because the number of cluster galaxies is low (so low to be negative!), the denominator should be very uncertain, and therefore the blue fraction should be very uncertain as well, i.e., that large values of the blue fraction are possible. Instead, the correct approach is to note that we observed a deficit of blue galaxies in the cluster's line of sight whereas a large blue fraction implies a large excess. To summarize, a reasonable constraint on the blue fraction `fclus` is achieved even for XLSSC 012, which, according to the butcher rule, has a vanishing (small), and thus very uncertain, number of galaxies.

The butcher rule to compute a fraction is to compute the ratio of the net (cluster+background minus background) number of blue galaxies over the net number of galaxies of all colors:

$$f_b(clus) = \frac{n(blue, cluster + field) - n(blue, field)}{n(total, cluster + field) - n(total, field)} \tag{6.1}$$

using the observed values. This formula certainly holds in the everyday practice, but rarely for research, and thus one should not be too surprised that the use of the formula above in research produces values outside the zero to one range. In fact, in scientific applications, the population of interest is often numerically minor, and the background is subject to Poisson fluctuations. In such conditions, Poissonian fluctuations may make background counts larger than counts in the cluster's line of sight, or give more blue galaxies in the background than in the cluster's line of sight, situations that return meaningless numbers, because Eq. (6.1) returns negative values for the blue fraction (i.e., we would claim that there are more red galaxies than galaxies of all colors) or blue fractions larger than one (i.e., we would claim that there are more blue galaxies than galaxies of all colors), statements that are hard to defend. The mistake lays in using the *observed* number of galaxies, instead of the (unknown) true one. The use of this formula obliged researchers to discard precious data, e.g., Postman et al. (2005) discarded two out of three of their $z > 1$ clusters in the determination of the spiral fraction.

The model illustrated in this section has, under simple conditions, a known analytical posterior distribution (e.g., D'Agostini 2004), see exercise 3.

6.1.4 Spectral Slope: Hardness Ratio

We, astronomers, are often faced with measuring the spectral shape of sources. When only a few photons are available, we are happy with coarse measurements of the spectral shape, such as the hardness ratio (or the color, or the optical to X ray ratio, or the spectral slope, etc.), because these allow one to classify the source in broad classes (as hard/soft, or red/blue, etc). The model introduced to determine the source intensity (6.1.2) can easily be modified to measure the probability distribution of these ratios. For clarity, we consider the hardness ratio, defined by

$$HR = \frac{H-S}{H+S} \quad , \tag{6.2}$$

where H and S are the object's flux in the Hard and Soft bands. The hardness ratio is a coarse measurement of the object's spectrum: hard sources have higher fluxes in the hard band at a given soft-band flux, i.e., large (near to 1) HR.

In each of the two bands, we are faced with the usual signal+background determination, i.e., with the model of Sect. 6.1.2 for the soft band:

```
obstotS ~ dpois(S+bkgS/CS)
obsbkgS ~ dpois(bkgS)
bkgS ~ dunif(0,1.0e+7)
```

and the same for the hard band

```
obstotH ~ dpois(H+bkgH/CH)
obsbkgH ~ dpois(bkgH)
bkgH ~ dunif(0,1.0e+7) ,
```

where `CS` and `CH` account for the difference (ratio, mathematically) of solid angles (or exposure time, or efficiency, or effective area, depending on the considered experiment), between the background and source, in the soft and hard bands. Now we need to adopt a prior for the hardness ratio, taken to be uniform over the full range of possible values, -1 to 1:

```
HR~dunif(-1,1).
```

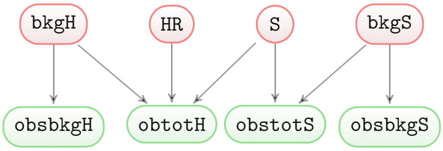

Fig. 6.10 Graph showing the stochastic relationship between quantities involved in estimating the hardness ratio

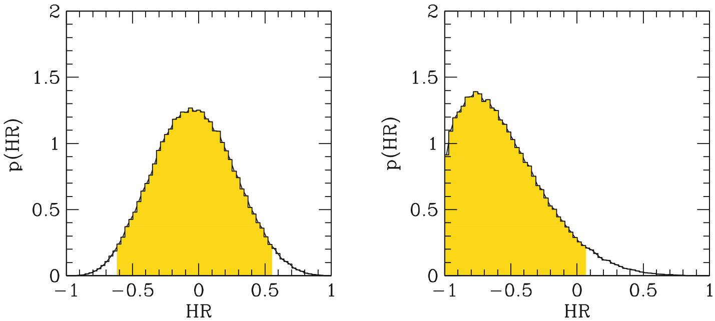

Fig. 6.11 Posterior probability distribution of the hardness ratio `HR` when `obstotH`= 5 (*left panel*) or `obstotH`= 1 (*right panel*). The *shaded region* marks the 95 % (highest posterior) interval

To end the model description, we need a prior on the intensity of the source in the soft band `s`, taken to be uniform over (approximately) the whole real positive axis:

```
S ~ dunif(0,1.0e+7)
```

and a simple re-arrangement of the equation defining the hardness ratio:

```
H<-2*S/(1-HR)-S.
```

To summarize, the model, graphically illustrated in Fig. 6.10, reads:

```
model{
obstotS ~ dpois(S+bkgS/CS)
obsbkgS ~ dpois(bkgS)
bkgS ~ dunif(0,1.0e+7)
obstotH ~ dpois(H+bkgH/CH)
obsbkgH ~ dpois(bkgH)
bkgH ~ dunif(0,1.0e+7)
S ~ dunif(0,1.0e+7)
HR~dunif(-1,1)
H<-2*S/(1-HR)-S
}.
```

Let us consider two cases, only differing in the number of observed counts in the hard band, `obstotH`, equal to 5 and 1. In both cases we observe `obstotS`= 5, `obsbkgS`= 10, `obsbkgH`= 30, and we have `CH`= 100 and `CS`= 100. The two posterior distributions of the hardness ratio are plotted in Fig. 6.11.

The uncertainties computed with this model are preferable to those derived assuming Gaussian errors on counts propagated on the hardness ratio because they do not extend outside the range of physically acceptable values and because they can be derived even in absence of a clean detection in one of the bands. Furthermore, as mentioned by Park et al. (2006), errors are shorter (and as consequence do not waste the information content of the data) and one selects more often less contaminated sub-samples of the sources with extreme properties, e.g., hard sources.

The hardness ratio in this section is defined as the ratio of intensities in counts units. If instead one would prefer to define it as ratios of intensities in other units (e.g., in flux units), one needs to edit the code and introduce two conversion factors to account for the change of units (from physical to counts). This may be important when considering observations taken from instruments with widely different relative sensitivities in the two considered bands, to bring the two hardness ratios on the same scale.

A technical detail: the MCMC sampling of this model requires a few thousands samples to achieve convergence (see Chap. 3) in the low signal to noise condition of our examples.

6.1.5 Spectral Shape

A crude measurement of the shape of the object's spectrum can be achieved by measuring the object's photons in three bands, S,M, and H and considering the two hardness ratios:

$$HR1 = \frac{M-S}{H+M+S}, \tag{6.3}$$

$$HR2 = \frac{H-M}{H+M+S}. \tag{6.4}$$

Such a crude measurement of the object's spectrum may be useful when a more detailed investigation is precluded by the low S/N of the source: the position of a source in the $HR1-HR2$ plane allows one to classify it in one of the several classes. For example, Binder et al. (2013) propose classifying sources in: X-ray binaries, background sources, absorbed sources, supernovae remnants, and indeterminated hard or soft sources, according to the location in the $HR1-HR2$ plane. To compute these two hardness ratios, we only need to replicate the modeling of the previous section, which was applied to two bands, to the three bands:

```
obstotS ~ dpois(S+bkgS/CS)
obsbkgS ~ dpois(bkgS)
bkgS ~ dunif(0,1.0e+7)
obstotH ~ dpois(H+bkgH/CH)
obsbkgH ~ dpois(bkgH)
bkgH ~ dunif(0,1.0e+7)
obstotM ~ dpois(M+bkgM/CM)
```

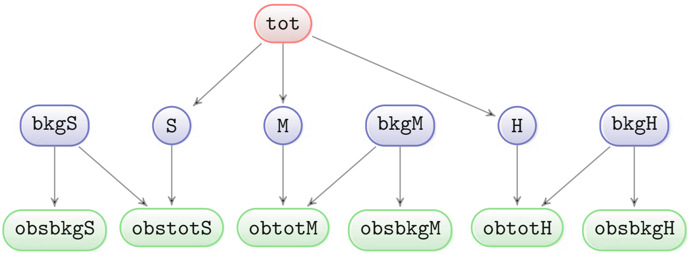

Fig. 6.12 Graph showing the stochastic relationship between quantities involved in estimating the spectral curvature. The interesting quantities HR1 HR2 are non-stochastic (deterministic) combinations of S, M, and H, and are therefore not plotted in the graph

```
obsbkgM ~ dpois(bkgM)
bkgM ~ dunif(0,1.0e+7).
```

We take a uniform prior for the total net counts tot=S+M+H, and a prior for the proportions S/tot , M/tot, and H/tot such that the joint prior of HR1 HR2 is uniform:

```
tot ~ dunif(0,1.0e+7)
p[1:3]~ddirch(alpha[1:3])
alpha[1]<-1
alpha[2]<-1
alpha[3]<-1
S <-p[1]*tot
M <-p[2]*tot
H <-p[3]*tot .
```

We compute in JAGS the two hardness ratios (but there is no real need of a statistical software package, as JAGS, to compute a ratio!):

```
HR1 <- (M-S)/(H+M+S)
HR2 <- (H-M)/(H+M+S) .
```

To summarize, the fitted model, graphically illustrated in Fig. 6.12, reads:

```
model {
obstotS ~ dpois(S+bkgS/CS)
obsbkgS ~ dpois(bkgS)
bkgS ~ dunif(0,1.0e+7)
obstotH ~ dpois(H+bkgH/CH)
obsbkgH ~ dpois(bkgH)
bkgH ~ dunif(0,1.0e+7)
obstotM ~ dpois(M+bkgM/CM)
obsbkgM ~ dpois(bkgM)
bkgM ~ dunif(0,1.0e+7)
tot ~ dunif(0,1.0e+7)
p[1:3]~ddirch(alpha[1:3])
alpha[1]<-1
alpha[2]<-1
```

Table 6.2 Simulated data sets for the spectral determinations

	obstotS	obstotM	obstotH	obsbkgS	obsbkgM	obsbkgH	CS	CM	CH
Source a)	50	30	7	10	50	30	10	10	10
Source b)	50	7	30	10	50	30	10	10	10
Source c)	7	30	50	10	50	30	10	10	10

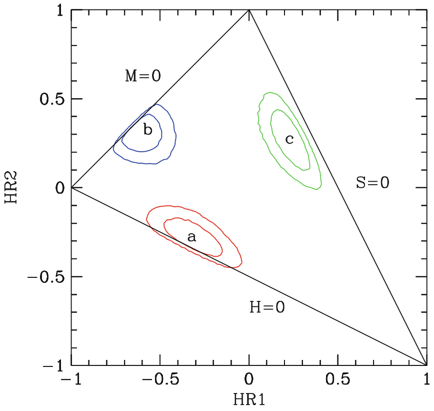

Fig. 6.13 Joint 68 % and 95 % probability contours of the two hardness ratios for three sources. The *triangle* includes the physical region. Note that the contours of sources a) and b) slightly overflow in the non-physical region because of an unavoidable approximation in the graphical plotting routines

```
alpha[3]<-1
S <-p[1]*tot
M <-p[2]*tot
H <-p[3]*tot
HR1 <- (M-S)/(H+M+S)
HR2 <- (H-M)/(H+M+S)
}.
```

Let us consider observations of three sources whose observations are listed in Table 6.2. The joint posterior distribution of the two hardness ratios of the three sources is plotted in Fig. 6.13. The triangle also drawn in the figure includes the physical region (no source can have a negative flux in any of the three bands). As the figure illustrates, there is a covariance between the two hardness ratios because four out of five of the terms entering in the two hardness ratio definitions are common between the two ratio definitions. Note that because of an unavoidable approximation in the graphical plotting routines (we need to estimate a density, this obliges us to sample a finite "pixel" region, part of which may be in the physical region even for pixels centered outside it), contours slightly overflow outside the physical region. However, as one may imagine, there is no value in the posterior chain falling in the non-physical region.

Table 6.3 Data for the globular cluster mixture model, from Brodie et al. (2012)

8.237 ± 0.740	7.604 ± 0.535	8.612 ± 0.300	5.500 ± 0.410	7.638 ± 0.786	5.648 ± 0.922
8.234 ± 0.309	8.038 ± 0.288	7.914 ± 0.314	6.108 ± 0.365	5.356 ± 0.653	4.979 ± 0.646
5.503 ± 0.629	7.720 ± 0.615	8.117 ± 0.598	4.268 ± 0.688	7.708 ± 0.337	7.559 ± 0.577
8.512 ± 0.379	7.677 ± 0.451	8.701 ± 0.521	4.890 ± 0.394	8.079 ± 0.198	6.970 ± 0.768
5.868 ± 0.349	5.021 ± 0.561	8.090 ± 0.459	7.153 ± 0.557	7.597 ± 0.511	5.930 ± 0.291
5.820 ± 0.313	8.750 ± 0.632	5.560 ± 0.487	8.831 ± 0.381	7.023 ± 0.388	7.084 ± 0.572
7.066 ± 0.402	8.556 ± 0.442	8.029 ± 0.556	7.227 ± 0.567	5.153 ± 0.610	7.922 ± 0.364
7.363 ± 0.358	8.663 ± 0.362	9.334 ± 0.505	4.664 ± 0.507	7.452 ± 0.891	8.738 ± 0.760
6.340 ± 0.905	6.907 ± 0.588	3.343 ± 0.566	5.221 ± 0.650	8.162 ± 0.445	5.022 ± 0.324
4.060 ± 0.419	4.842 ± 0.488	6.054 ± 0.342	6.273 ± 0.347	8.065 ± 0.343	8.244 ± 0.402
5.339 ± 0.577	6.282 ± 0.272	6.964 ± 0.441	5.608 ± 0.399	5.193 ± 0.474	4.695 ± 0.635
7.254 ± 0.360	3.316 ± 0.611	7.336 ± 0.477	4.444 ± 0.541	5.694 ± 0.402	

Sources can now be classified: the probability that a source is of a given type is given by the fraction of the samplings that satisfy the definition of the type (e.g., $HR2 < -0.4$ and $HR1 < -0.4$ for "supernova remnant").

Hardness ratios are defined in this section as ratios of intensities in counts units. Those who prefer instead to define them as ratios of intensities in other units (e.g., in physical units) should edit the code and introduce the three conversion factors (one per band).

6.2 Mixtures

The purpose of this section is to introduce a model to fit multimodal distributions, in particular distributions that can be composed as the sum of Gaussians. Many complex distributions can be modelled as sums of Gaussians. While we only consider Gaussians in this chapter, fitting mixtures of other distributions is similarly simple, e.g., to fit mixtures of Student-t one just needs to replace Normal with Student-t, (i.e., `dnorm` with `dt`). Similarly, although we consider two Gaussians only, our models are written in such a way that an arbitrary number of Gaussians can be fitted with little editing.

6.2.1 Modeling a Bimodal Distribution: The Case of Globular Cluster Metallicity

We begin with a simple application, we are interested in modeling the bimodal metallicity distribution of globular clusters in the nearby galaxy NGC 3115 (Brodie et al. 2012). The authors measured the metallicity-sensitive spectroscopic index CaT, named in our code `obsval[i]`, and its error `err[i]`, for 71 globular

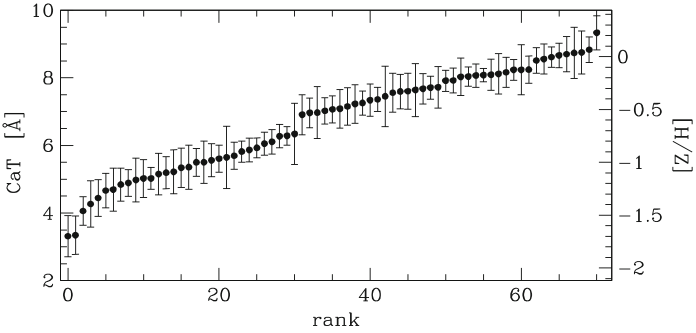

Fig. 6.14 Data and errors of the NGC 3115 bimodal metallicity distribution example

clusters (the data is listed in Table 6.3). The authors show, using population synthesis models, that the CaT index is a direct measurement of metallicity [Z/H]. Therefore, in our figures, we draw both scales, the raw one (CaT) and the derived one ([Z/H]). The observed values and errors are shown in Fig. 6.14 (the data has been first ranked by value). The average error is 0.5 Å. The histogram of the observed values is shown in Fig. 6.15.

We want to determine the parameters that describe the metallicity distribution of globular clusters in NGC 3115, namely the two Gaussian centers, `cent[1]` and `cent[2]`, the two sigmas `sig[1]` and `sig[2]`, and their relative importance (heights). We emphasize that we are interested in the more challenging distribution of the true metallicities, not to model the distribution of observed values which is increasingly broader as the errors increase.

As usual, observed values of metallicity are noisy. We assume, as the authors do, Gaussian errors:

```
obsval[i] ~ dnorm(val[i],pow(err[i],-2)).
```

As the authors, we model the distribution of the true values `val` as a mixture (sum) of two Gaussians:

```
val[i] ~ dnorm(cent[I[i]],pow(sig[I[i]],-2)).
```

`val[i]` is drawn from one of the Gaussians whose parameters depend on an index `I[i]` that takes value 1 or 2, depending on whether the point belongs to the first or second Gaussian. The first Gaussian is centered on `cent[1]` with sigma `sig[1]`, the second is centered on `cent[2]` with sigma `sig[2]`.

We emphasize that modeling the distribution as a sum of two distributions with the same analytic expression (for example, the sum of two Gaussians) induces an identifiability problem: the model with swapped $(1,2)$ indices is indistinguishable from the original one because, thus far, there is no information in the model about which of the two Gaussians should be called "first" and which "second." This lack of identifiability raises both numerical challenges (how to optimally sample the posterior) and annoying issues for the researcher (e.g., it becomes harder to know if the

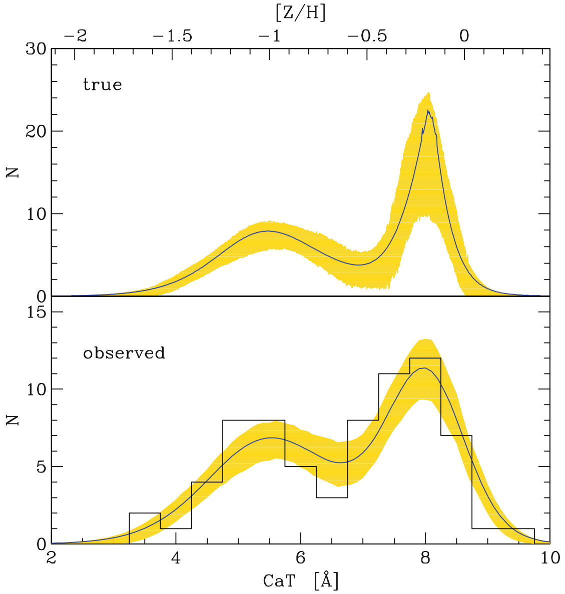

Fig. 6.15 Fitted distribution, in the true data space (*top*) and observed data space (*bottom*) of the NGC 3115 bimodal metallicity distribution example. The *blue line* indicates the mean model, the *shading* marks the 68 % error of the model. The *histogram* shows the distribution of the observed values

chain has converged). The problem may be fixed in several ways, in this example we assign "second" to the component with the larger center, i.e., we ask that the center of the second distribution be larger than the center of the first:

```
cent[1] ~ dunif(4,10)
cent[2] ~ dunif(cent[1],10).
```

As indicated, we take uniform priors for the centers (their location is quite obvious for the current data and therefore we do not need to bother ourselves anymore with the center's prior choice). We take 4 as lower value for the center of the first component for numerical reasons: this helps the chain converge in a few iterations. We have already emphasized in Chap. 3 the importance of inspecting trace plots, when dealing with mixtures this is even more important.

We take uniform distributions over a large enough range that certainly includes the true value for the two sigmas:

```
sig[1] ~ dunif(0.,10)
sig[2] ~ dunif(0.,10).
```

We also assume a priori that `I`, the variable that indicates which Gaussian the point belongs to, is equal to 1 or 2 with the same probability. The simplest implementation of this idea is to use the distribution `dcat`, which returns the value 1 or 2 with probabilities set by its argument `p[]`:

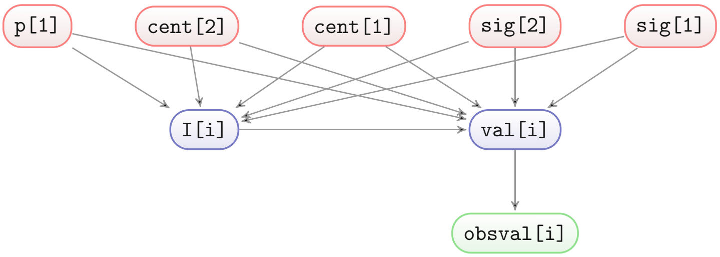

Fig. 6.16 Graph showing the stochastic relationship between quantities used for mixtures. This figure uses variable names appropriate for the bimodal example. With minor changes, this graph is easily generalized

```
I[i] ˜ dcat(p[]),
```

where p[1] and p[2] are the probability that the data belong to the first and second components. These are also the amplitudes of the two Gaussians, a large p[1] indicates that most of the points belong to the first population.

We take a uniform priors for the amplitudes p[1] and p[2]:

```
p[1] ˜ dunif(0,1)
p[2] <-1-p[1].
```

The whole model, illustrated in Fig. 6.16, reads:

```
model {
for (i in 1:length(obsval)){
obsval[i] ˜ dnorm(val[i],pow(err[i],-2))
val[i] ˜ dnorm(cent[I[i]],pow(sig[I[i]],-2))
I[i] ˜ dcat(p[])
}
sig[1] ˜ dunif(0,10)
sig[2] ˜ dunif(0,10)
cent[1] ˜ dunif(4,10)
cent[2] ˜ dunif(cent[1],10)
p[1] ˜ dunif(0,1)
p[2] <-1-p[1]
}.
```

Figure 6.15 shows the fitted distribution, in the true data space (top) and observed data space (bottom). The latter has been obtained, for simplicity, by convolving the true model by the mean error (0.5), in practice adding in quadrature 0.5 to the fitted sigmas. This approximation is not necessary, and in Sect. 9.2 we will describe (with less efforts than writing this sentence!) how to generate simulated data to be used to plot the predicted distribution. The histogram in the bottom panel shows the distribution of the observed values.

Figure 6.17 shows the posterior marginals (histograms) and the adopted priors (dashed lines). In each case the data (likelihood) dominates the posterior because the prior is quite flat in the range where the posterior is non-null. Two comments

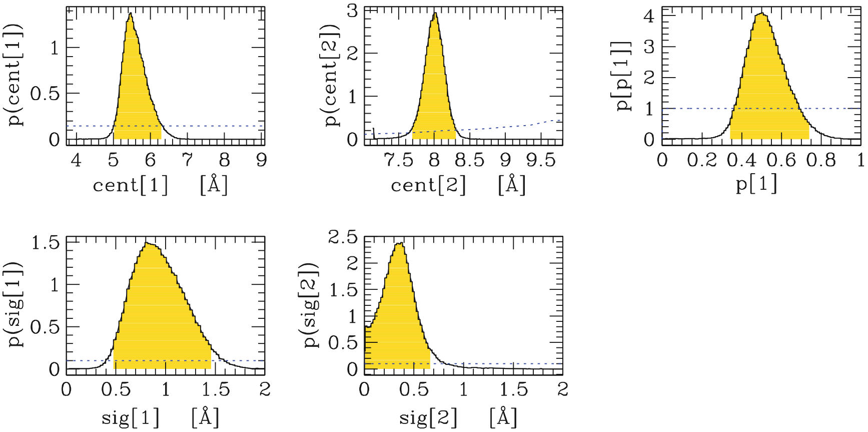

Fig. 6.17 Posterior (*solid lines*) and prior (*dashed lines*) distributions for the parameters of the NGC 3115 bimodal metallicity distribution example. 95 % intervals are *shaded*

are in order. First, panel c plots the posterior probability distribution of `p[1]`. Its value is ≈ 0.5, indicating that there are two Gaussians with similar amplitude, i.e., that approximately 50 % of the points belong to the first Gaussian. Second, the posterior distribution of `sig[2]` (see panel e) allows small values, i.e., the second component may have a small spread. This situation occurs because the data scatter marginally more than their error (hence large `sig[2]` are excluded) but the data uncertainty is comparable to the width of the Gaussian `sig[2]`, and hence insufficient to resolve the variance components.

Because we have modeled, probabilistically, the membership of each globular cluster to one of the two Gaussians, we have for each globular cluster the probability that it belongs to the first component. This number can be obtained by counting the number of the times the value 1 occurs in the posterior chain of the i^{th} datum `I[i]`. As expected, the probability depends on both the data value (extreme values are usually taken at the extremes of the Z/H scale) and its error (points with large error have less clear membership and thus usually avoids probabilities close to 0 or 1). For example, the 45th point has a high metallicity values ($Z/H \sim 0$), `I[45]` is 2 most (90 %) of the time. The 51th point, with a very low values ($Z/H \sim -1.5$), has `I[51]` equal to 1 almost always (99.9 % of the time). The 50th point, near the middle of the distribution ($Z/H \sim -0.6$), has `I[50]` equal to 1 50 % of the time.

Figure 6.18 shows the joint probability distribution. Most parameters show covariance, because when a parameter takes a lower-than-average value, the other parameters are adjusted to fit the data. For example, when `cent[1]` is higher, the width `sig[1]` scales to fit the data. In general, the data set the amplitude of the covariance between parameters, with the more constraining data producing parameter with lower covariance.

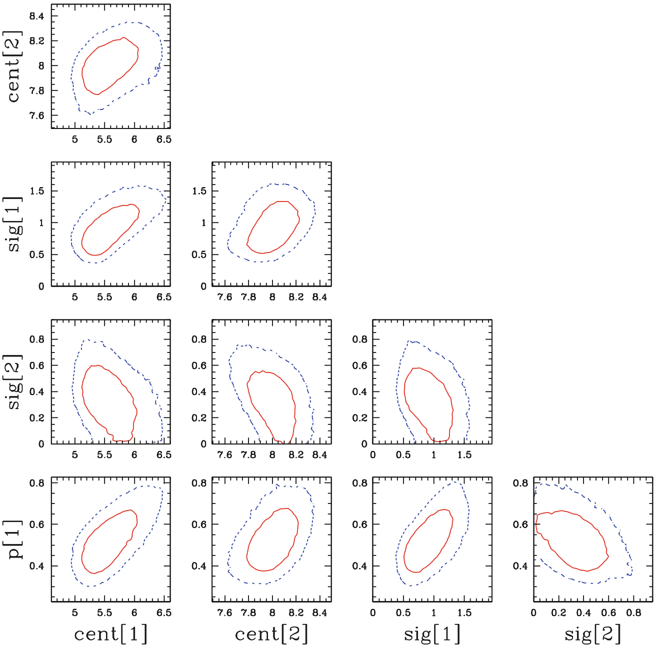

Fig. 6.18 Joint 68 % and 95 % probability contours for the metallicity example. Contours are fully enclosed in the range of physically acceptable values. In particular, they do not cross the `sig[2]` = 0 value

We emphasize that if this model is applied to other data sets, it is necessary, as usual, to pay attention to the prior. In particular, care should be exercised when specifying the minimal value adopted for the two sigma priors for two reasons: first, these sigmas are raised to the power of -2 (and very low values may cause numerical troubles), and second, a possible subtle difference may be present between the model being fitted and the model believed to be fitted: a model with one of the two sigmas going to zero is modeling a distribution with a Dirac-delta function fitting one point plus a Gaussian fitting all other points. Unless this is the aim of the researcher's modeling, this occurrence is not welcome, and thus should be checked (inspecting the trace plots is useful in this respect). If this occurs and is not welcome by the researchers, it can be fixed by restricting the sigma prior to strictly positive values.

Finally, in the case where one may want more flexibility in assigning a prior to `p[]`, one may adopt the (somewhat involved) construct:

```
I[i] ~ dcat(p[])
p[1:2] ~ ddirch(alpha[1:2])
alpha[1] <-1
alpha[2] <-1
```

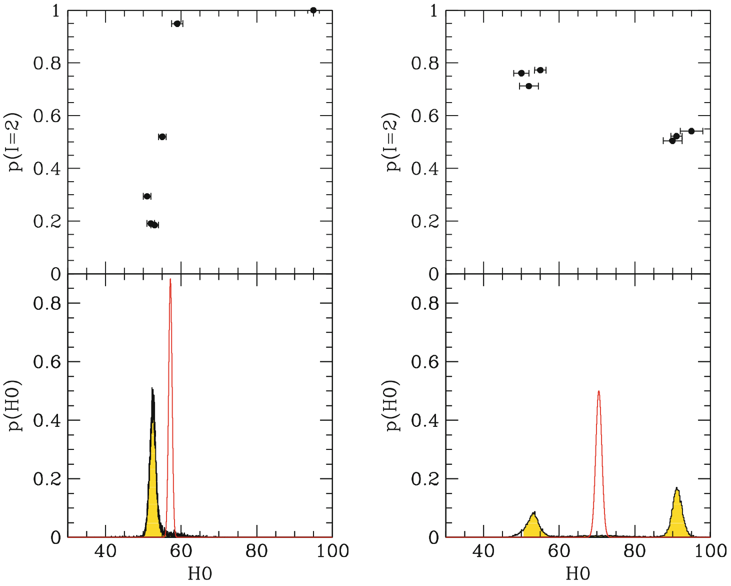

Fig. 6.19 Posterior probability distribution of the parameter H0 (*bottom panel*) for the data set shown in the *top panel*. The *bottom panel* also reports the strawman average and uncertainty, depicted as a Gaussian

and tune the `alpha[]` values to get the prior one is looking for. The dirichet distribution `ddirch[]` returns values, in the 0 to 1 range, distributed as power-laws with slopes given by the distribution's arguments minus 1.

6.2.2 Average of Incompatible Measurements

As soon as one has a few measurements of some quantity of scientific interest, it may occur that the spread of the observed values seems too wide compared to the data errors. Or, at the very least, there is one suspicious point, i.e., a point that seems too different from the other ones to be correct (known as an outlier).

Consider the following example from Press (1997). One has a number of estimates of a quantity, called `H0`, and they vary in the range 50 to 100 km/s with errors of the order of 1 km/s. How does one average[1] these incompatible measurements?

One possibility is to misuse the usual formula for averaging Gaussian quantities. The resulting Gaussian is depicted in the bottom panel of Fig. 6.19 for the (fake) data listed in Table 6.4: the Gaussian's sigma (i.e., the returned error on the average) is manifestly too small given the wide dispersion of the data, and the Gaussian's

[1] If this example sounds familiar to you, then you are an experienced astronomer. If this sounds unfamiliar, for decades astronomers discussed how to reconcile incompatible values of the Hubble constant.

Table 6.4 Simulated data sets for the two H0 determinations

Left panel:	51 ± 1.0	52 ± 1.0	53 ± 1.0	55 ± 1.0	59 ± 1.5	95 ± 1.5
Right panel:	50 ± 2.0	52 ± 2.5	55 ± 1.5	90 ± 2.5	91 ± 1.5	95 ± 3.0

location is suspect because it falls in the region where no data are observed. To a statistician, the oddity of this result comes as no surprise: the adopted formula assumes some regularity conditions manifestly not satisfied: the data are scattered more than the error. The model clearly does not fit the data.

Instead, let us perform a Bayesian analysis. Because the model does not fit the data, we need to change something about it. Either there is an unspecified systematic effect between the measurements of the various experiments, or researchers have an incurable optimism about the constraint given by their data. In this simplified analysis, we may suppose that either the measurements have correctly measured errors (coded with I[i]=1), or optimistically estimated (coded with I[i]=2) by some amount up to a factor of 50, and, of course, we do not know which experiment quotes correct or underestimated errors.

The model, very similar to the one in the previous section, reads:

```
model {
for (i in 1:length(obsH0)){
 obsH0[i] ~ dnorm(H0,pow(corr.err[i],-2))
 I[i] ~ dcat(p[])
 corr.err[i] <- obserr[i]*phi[I[i]]
 }
H0 ~ dunif(30,200)
# First Group errors are right
phi[1] <- 1
# Second Group errors are optimistically estimated
# by a common factor phi[2]
phi[2] ~ dunif(1,50)
p[1]~dunif(0,1)
p[2]<-1-p[1]
}.
```

The membership to the two groups of experiments (with correct or optimistic errors) is updated via the observed values of H0, that constrain both the value of H0 and err.tot, which in turn is the quoted error obserr multiplied by the error correction factor phi[i].

Figure 6.19 shows the result of the analysis for two hypothetical (simulated) data sets, detailed in Table 6.4. In the left panel an outlier, at H0= 90 km/s, is quite obvious and the analysis clearly identify it as such ($p(\mathtt{I}=2)\sim 1$). A second point, at H0= 60 km/s, has likely an underestimated error ($p(\mathtt{I}=2)\sim 1$), although it is not straightforward to identify it as such by visual inspection of the data. These two points drive the strawman estimate away from all points (all points disagree with the strawman average at 4 sigma or more). These could, in principle, be identified by some methods detecting outliers, and removed from the average. But what should

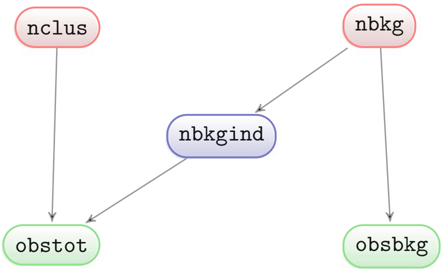

Fig. 6.20 Graph showing the stochastic relationship between quantities used for source intensity with over-Poisson background fluctuations determination

be done with the third point, the one with about 50 % probability to have the right error, and 50 % to have an underestimated error? In the Bayesian approach the right thing is done since the start of the analysis, each datum contributes to the posterior of `H0` with a weight which is larger when the error is smaller and the probability that the error is correct is large.

The right panel of Fig. 6.19 shows a case more typical of the historical determination of `H0`, two sets of points clustering at two different values. In such a case, the probability distribution of `H0` is bimodal, as the logic asks, fairly different from the (forced Gaussian) strawman average (commonly used for a number of years) centered on values not represented by the data in this example. None of the data is discarded in the `H0` determination (i.e., all the data are used and nothing is wasted), and outlier-detecting methods would have trouble deciding which points should be removed in this case.

6.3 Advanced Analysis[2]

6.3.1 Source Intensity with Over-Poisson Background Fluctuations

The X-ray background shows a non-zero variance: by performing repeated measurements of the background's value in different directions, values scatter more than Poisson fluctuations allow. On the arcmin angular scales, Andreon and Moretti (2011) measure intrinsic fluctuations of 20 %. These fluctuations limit the maximal precision with which a source flux may be measured because the background will never be known better than with 20 % error. We now account for these fluctuations and determine the probability distribution of the X-ray cluster's luminosity L_X, following Andreon and Moretti (2011). The model, illustrated in the graph in Fig. 6.20, is very similar to the one used in measuring the source intensity in the presence of

[2] This section can be skipped the first time through.

Table 6.5 Data for the source intensity with over-Poisson background fluctuations

Id	obstot	obsbkg	nbox	C
2	35	269	22	41.49
6	156	325	52	41.77
9	415	3827	10	39.67

a background (Sect. 6.1.2), accounting for the Poisson nature of counts, uncertainty on the mean value of the background, and the existence of boundaries in the data and parameter space. However, it also accounts for over-Poisson background fluctuations. Let us recap: because of errors, observed and true values are not identical. We call `nclus[i]` and `nbkg[i]` the true cluster and the true background counts in the studied solid angles. We measured the number of photons in both clusters and background regions, `obstot[i]` and `obsbkg[i]`, respectively, for each cluster that compose the sample. The background's solid angle is `nbox[i]` times larger than the cluster's solid angle. We assume a Poisson likelihood for both the cluster and background and that all measurements are conditionally independent given all necessary parameters:

```
obstot[i] ~ dpois(nclus[i]+nbkgind[i]/nbox[i])
obsbkg[i] ~ dpois(nbkg[i]).
```

`nbkgind[i]` has 20 % fluctuations around the global background value:

```
nbkgind[i] ~ dlnorm(log(nbkg[i]),pow(0.2,-2)) ,
```

where `dlnorm` stands for the lognormal distribution.

We assume uniform priors on cluster and background counts, that assign zero probability to non-physical values:

```
nbkg[i] ~ dunif(1,1.0E+7)
nclus[i] ~ dunif(0,1.0E+7).
```

The model returns the X-ray luminosity in counts units, that can be converted to the usual astronomical units (erg s^{-1}). To a non-astronomer, it is enough to know that we convert numbers into the standard astronomical units by taking the (decimal) log, and adding a suitable computed constant `C[i]` (that accounts for the exposure map, this assumes a thermal spectrum of a given temperature, metallicity, and redshift and accounts for the Galactic absorption):

```
lgLx[i] <- log(nclus[i])/2.30258 +C[i] ,
```

where the 2.30258 coefficient is needed to convert Neperian logs into decimal logs. The latter operation does not need to be done inside JAGS.

Therefore, the full model reads:

```
model {
for (i in 1:length(obstot)) {
obstot[i] ~ dpois(nclus[i]+nbkgind[i]/nbox[i])
```

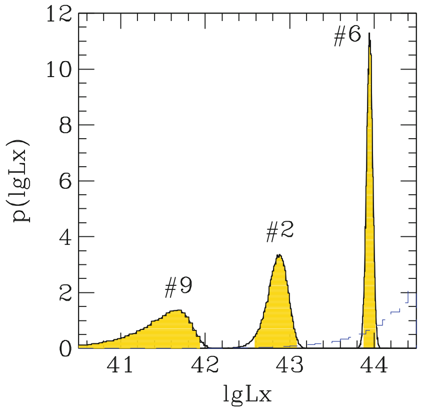

Fig. 6.21 L_X probability distribution for three sources. The posterior is marked with a *solid line*, and the prior is marked with a *dashed-blue line*. The 95 % probability interval is *shaded*. Note the non-Gaussian shape (e.g., asymmetry) of the faintest source (#9). Numbers indicate the object ID

```
nbkgind[i] ~ dlnorm(log(nbkg[i]),pow(0.2,-2))
obsbkg[i] ~ dpois(nbkg[i])
nbkg[i] ~ dunif(1,1.0E+7)
nclus[i] ~ dunif(0,1.0E+7)
# optional, JAGS is not needed to do it
lgLx[i] <- log(nclus[i])/2.30258 +C[i]
}
}.
```

The data for three sources, taken from Andreon and Moretti (2011), are listed in Table 6.5.

Figure 6.21 shows the (posterior) probability distribution of L_X for three clusters. Sharp distributions indicate precisely determined L_X, such as for cluster #6. Sources with noisier determinations, such as cluster #9, and to a lesser extent cluster #2, have asymmetric posterior probability distributions. For cluster #9, the data excludes bright L_X values, whereas at low intensities, the prior on the object flux dominates.

6.3.2 The Cosmological Mass Fraction Derived from the Cluster's Baryon Fraction

If clusters of galaxies are representative samples of the Universe, then the baryon fraction (i.e., mass in baryons over total mass) in clusters should be equal to the baryon fraction in the Universe Ω_b/Ω_m (denoted here by omegab and omegam, respectively). Clusters of galaxies offer a way to measure this combination of cosmological parameters. Because omegab is determined with precision of a few percent, the measurement of the baryon fraction in clusters is, by large, offering a way to determine omegam.

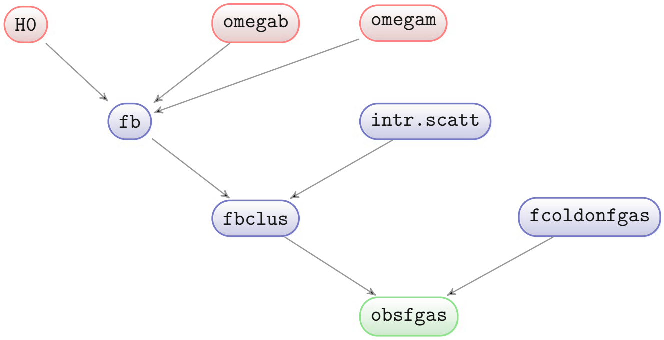

Fig. 6.22 Graph showing the stochastic relationship between quantities involved in the cosmological mass fraction derived from the cluster baryon fraction

Table 6.6 Data set for the cosmological mass fraction determination, from Ettori et al. (2009)

0.131 ± 0.027	0.146 ± 0.009	0.091 ± 0.007	0.133 ± 0.012	0.136 ± 0.014	0.121 ± 0.042
0.165 ± 0.012	0.120 ± 0.012	0.105 ± 0.017	0.160 ± 0.013	0.159 ± 0.010	

The actual path to omegam is a bit more complicated than just described, as detailed in the graph in Fig. 6.22: the constraint on omegam is indirect as it requires two intermediate steps, each one characterized by an intrinsic stochasticity and an additional parameter. First, the data, taken from Ettori et al. (2009) and listed in Table 6.6, only probe the baryon mass in gas. Clusters also host baryons in stars, and, following the authors, we assume that the ratio of the star and gas baryon fractions is given by

```
fcoldonfgas ~ dnorm(0.18,pow(0.05,-2)) .
```

This assumption has been improved in Andreon (2010a), but we keep it to make our analysis as close as possible to Ettori et al. (2009). Second, numerical simulations show that clusters are slightly depleted in baryons, compared to the Universe's value, by a factor 0.874 ± 0.023, i.e., that

```
fbcorr <- fb*corr
corr ~dnorm(0.874,pow(0.023,-2)) .
```

The observed value of the gas fraction of each individual cluster, obsfgas[i], scatters around the true value, fbclus[i], by an amount given by the error, errfgas[i], i.e.,

```
obsfgas[i] ~ dnorm(fbclus[i]*(1-fcoldonfgas),pow(errfgas[i],-2)).
```

Inspection of the data in Fig. 6.23, previous claims about the existence of an intrinsic scatter (Sun et al. 2009; Andreon 2010a), and prudence, all imply that it is sensible to allow for an intrinsic scatter, possibly being very small, in gas fraction. We assume Gaussian scatter for simplicity

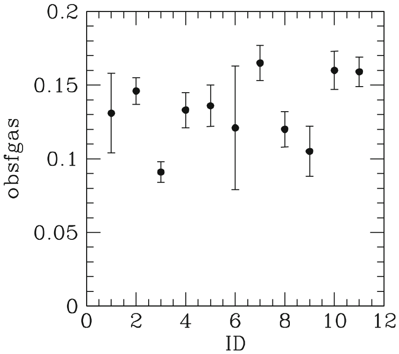

Fig. 6.23 Values of the gas fraction observed in the 11 LEC clusters with high quality data studied in Ettori et al. (2009)

```
fbclus[i] ~ dnorm(fbcorr,pow(intr.scatt,-2)),
```

although we note that being the fraction bounded in the zero to one range, the scatter cannot be, formally speaking, Gaussian because every Gaussian is (mathematically) non-zero everywhere, i.e., also for unphysical values for a fraction (i.e., fractions > 1 or < 0). This assumption has been improved in, e.g., Andreon (2010a), but we keep it to make our analysis as close as possible to Ettori et al. (2009). As shortly mentioned, independent observations allow one to put a quite stringent constraint on the Universe baryon fraction:

```
fb <- omegab/omegam
omegab <-omegabh2 * pow(H0/70,-2)
omegabh2 ~dnorm(0.0462,pow(0.0012,-2))
H0 ~ dnorm(70.1,pow(1.3,-2)).
```

To conclude our model description, we need to specify the prior on omegam, that we take to be uniform between 0 and 1, and on the intrinsic scatter, that we take uniform between 0 and 0.1

```
omegam ~ dunif(0,1)
intr.scatt ~ dunif(0.,0.1).
```

To sum up, our model reads:

```
model {
for (i in 1:length(obsfb)) {
obsfgas[i] ~ dnorm(fbclus[i]*(1-fcoldonfgas),pow(errfb[i],-2))
fbclus[i] ~ dnorm(fbcorr,pow(intr.scatt,-2))
}
intr.scatt ~ dunif(0.,0.1)
fbcorr <- fb*corr
corr ~dnorm(0.874,pow(0.023,-2))
fcoldonfgas ~ dnorm(0.18,pow(0.05,-2))
fb <- omegab/omegam
omegab <-omegabh2* pow(H0/70,-2)
```

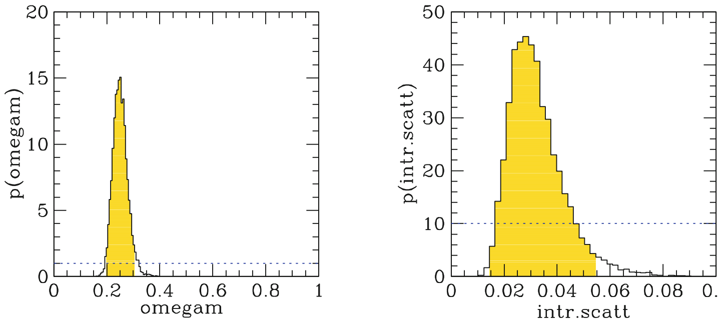

Fig. 6.24 Posterior probability distribution of `omegam` and `intr.scatt`. The *shaded region* marks the 95 % probability interval. The *horizontal dotted lines* indicate the assumed priors

```
omegabh2 ~dnorm(0.0462,pow(0.0012,-2))
H0 ~ dnorm(70.1,pow(1.3,-2))
omegam ~ dunif(0,1)
}.
```

Our assumptions are identical to those adopted by Ettori et al. (2009), except for our allowance of an intrinsic scatter, and the use of `omegam` and `intr.scatt` priors, while Ettori et al. (2009) use a χ^2 approach and thus “no prior” on these parameters. For the remaining parameters, Ettori et al. (2009) take the very same priors adopted here.

Figure 6.24 shows the posterior distribution computed using the model above on the 11 high quality LEC galaxy clusters in Ettori et al. (2009) listed in Table 6.6. We found $\Omega_m = 0.25 \pm 0.03$ and a posterior distribution on the intrinsic scatter that strongly avoids 0 (i.e., we definitively find an intrinsic scatter).

In their analysis, Ettori et al. (2009) add other data, which, however, do not constrain Ω_m, and obtain for the very same sample 0.26 ± 0.015. The authors' apparently more precise Ω_m determination is entirely a result of a simplified statistical analysis (errors are just summed in quadrature) and having assumed no intrinsic scatter. In fact, under this (manifestly incorrect) assumption, we found $\Omega_m = 0.25 \pm 0.02$.

We emphasize that the mean intrinsic scatter, 0.03, is larger, for 10 out of 11 clusters, than the measurement error, and this is the limiting factor in the precision of gas fraction measurements, not the measurement error.

6.3.3 Light Concentration in the Presence of a Background

Let us suppose one needs to measure how concentrated the light of a source is from measurements `obstot1[i]` and `obstot2[i]` performed in two apertures (the larger having a 5 times larger radius), of a set of N sources, in the presence of a background. The background rate has no reason to be the same from source to

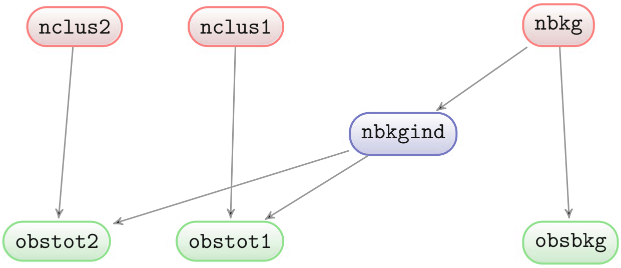

Fig. 6.25 Graph showing the stochastic relationship between quantities involved in the light concentration in the presence of a background determination

source, for example because the considered solid changes (from source to source), but it is the same inside the two source apertures. The model, quite similar to the determination of the source intensity in the presence of over-Poisson background fluctuations (Sect. 6.3.1), is depicted in the graph in Fig. 6.25, and given by:

```
model {
for (i in 1:length(obstot1)){
# aperture 1 (small)
obstot1[i] ~ dpois(nclus1[i]+nbkgind[i]/nbox[i]/25)
# aperture 2 (large minus small)
obstot2[i] ~ dpois(nclus2[i]+nbkgind[i]/nbox[i]*24/25)
nbkgind[i] ~ dlnorm(log(nbkg[i]),1/0.2/0.2)
obsbkg[i] ~ dpois(nbkg[i])
nbkg[i]~dunif(0,1e+7)
nclus1[i]~dunif(nclus2[i]/24,1e+7)
nclus2[i]~dunif(0,1e+7) # taken indep for the time being
}
}.
```

Note that we assumed that sources have higher brightness in the central aperture by specifying the prior on `nclus1[i]` to have a lower boundary given by `nclus2[i]/24`. This requirement is almost always satisfied for astronomical sources.

Light concentration is defined as the ratio of the true flux in the smaller aperture, `nclus1[i]`, over the true flux in the larger aperture `nclus2[i]`. Because of the adopted prior, no $\log_{10}$-concentration can be smaller than -1.38.

Data for three sources, taken from Andreon et al. (in prep), are listed in Table 6.7. Figure 6.26 shows the (posterior) probability distribution of the $\log_{10}$-concentration for the three sources. Source # 20 has a well-determined concentration, whereas the data only offer an upper limit value for source # 35. Source # 23 has a determination of the concentration of quality intermediate between the two.

Table 6.7 Data for the light concentration determination

Id	obstot1	obstot2	obsbkg	nbox
20	125	442	181	10
23	4	97	539	58
35	4	140	2438	58

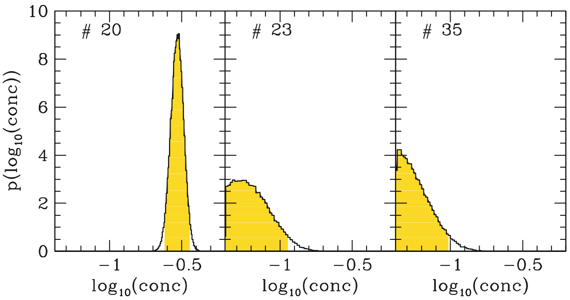

Fig. 6.26 Probability distribution of the three sources (log) concentrations. 95 % intervals are *shaded*

6.3.4 A Complex Background Modeling for Geo-Neutrinos

In this section we illustrate the flexibility of Bayesian modeling by re-analyzing the detection of geo-neutrinos by the Borexino collaboration (BOREXINO Collaboration et al. 2010), but with a different spirit: we are interested in illustrating the simplicity of a non-approximate statistical analysis, these authors are interested instead to quickly reach the finish line: obtaining physics results. To further emphasize the different spirit, the data used in this example are those in the first version of their paper, slightly different from the data used in the published version of their paper. Mainly, the overall problem addressed in BOREXINO Collaboration et al. (2010) is the usual signal+background problem discussed many times in this chapter.

6.3.4.1 An Initial Modeling of the Background

After an initial cleaning of the data, applied to enhance the signal and to decrease the background, the authors detected some events (21 events, for the sake of precision), some of which are background (not anti-neutrinos). Information on the background rate is partially available in the paper: the authors list 7 different sources of background with rates given in the form $x \pm y$ plus 5 more sources with 90 % upper limits to their rates, as detailed in Table 6.8, taken from their paper.

In their analysis, the authors ignore upper limits, sum the 7 point estimates (x's) to obtain the total background rate of 0.15, and sum in quadrature the error estimates (y's) to get the uncertainty on the total background rate, 0.02 (see last line of the Table 6.8). The result is depicted in Fig. 6.27 as the solid red line.

Table 6.8 Estimated backgrounds for the $\bar{\nu}_e$'s of the Borexino experiment

Adopted symbol	Source	Background rate
`bkg.lihe`	^{9}Li–^{8}He	0.03±0.02
`bkg.fastn`	Fast n's (μ's in WT)	<0.01
`bkg.fastn2`	Fast n's (μ's in rock)	<0.04
`bkg.untagmu`	Untagged muons	0.011±0.001
`bkg.acccoin`	Accidental coincidences	0.080±0.001
`bkg.timcor`	Time corr. background	<0.026
`bkg.fot`	(γ,n)	<0.003
`bkg.spontfi`	Spontaneous fission in PMTs	0.0030±0.0003
`bkg.scin`	(α,n) in scintillator	0.021±0.002
`bkg.buff`	(α,n) in the buffer	<0.061
	Total	0.15±0.02

The table is taken from BOREXINO Collaboration et al. (2010), except for the first column which is added by us

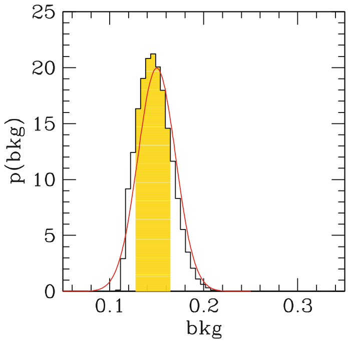

Fig. 6.27 Background rate probability distribution for the Borexino experiment, ignoring some background terms. The histogram accounts for the non-negative nature of the considered rates, the curve indicates the Borexino approximation. The *shading* indicates the 68 % interval

Let us start by checking the approximation involved in the adopted procedure. Ignoring (for a moment), as they do, background sources with upper limits, and considering, as they do, the other 7 sources of background to have Gaussian distributions, centered in x with standard deviation y, their background model reads:

```
bkg.lihe ~ dnorm(0.03,pow(0.02,-2)) T(0,)
bkg.untagmu ~ dnorm(0.011,pow(0.001,-2)) T(0,)
bkg.acccoin ~ dnorm(0.080,pow(0.001,-2)) T(0,)
bkg.spontfi ~ dnorm(0.003,pow(0.0003,-2)) T(0,)
bkg.scin    ~ dnorm(0.021,pow(0.002,2)) T(0,)
bkg <- bkg.lihe +bkg.untagmu +bkg.acccoin +bkg.spontfi
       +bkg.scin ,
```

where we took the x and y numbers from their paper and we add the positive constraint, T(0,), because negative values for the background are not possible physically.

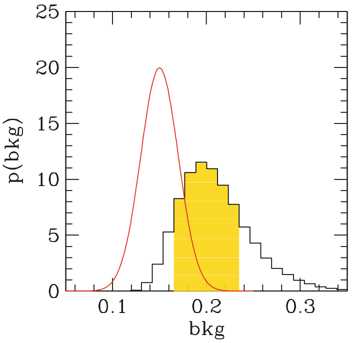

Fig. 6.28 Background rate probability distribution for the Borexino experiment accounting for all background terms. The histogram accounts for the non-negative nature of the rate, the curve indicates the Borexino approximation. The *shading* indicates the 68 % interval

Figure 6.27 shows the distribution of `bkg` (histogram) and the authors' approximation (red curve). The histogram is slightly asymmetric. The asymmetry originated from the fact that each individual background cannot be negative and some backgrounds have severely truncated distributions, for example `bkg.lihe`. For the latter source of background events, the unphysical (negative) background rate is just 1.5 σ away from the most likely background rate. Our computation does account for this, the authors' analysis does not, and this explains the small difference between our (histogram) and their (solid curve) results visible in Fig. 6.27.

Let us now do a step further and account for the five sources with upper limits neglected by BOREXINO Collaboration et al. (2010). We adopt an exponential distribution for the five sources such that the 90 % upper limit corresponds to the value quoted in the paper. The choice of an exponential is motivated by the lack of any information in their paper about their probability distribution and because this is a natural choice (what is expected to be obtained for an observation with 0 observed background events). We are not claiming that an exponential is the correct distribution to be taken, only that it is a plausible choice in the absence of any better information. With these additional 5 sources of background, the background model reads:

```
model{
bkg.lihe ~ dnorm(0.03,pow(0.02,-2)) T(0,)
bkg.fastn ~ dexp(200)
bkg.fastn2 ~ dexp(58)
bkg.untagmu ~ dnorm(0.011,pow(0.001,-2)) T(0,)
bkg.acccoin ~ dnorm(0.080,pow(0.001,-2)) T(0,)
bkg.timcor ~ dexp(88)
bkg.fot ~ dexp(765)
bkg.spontfi ~ dnorm(0.003,pow(0.0003,-2)) T(0,)
bkg.scin    ~ dnorm(0.021,pow(0.002,2)) T(0,)
bkg.buff    ~ dexp(37)
bkg <- bkg.lihe+ bkg.fastn + bkg.fastn2 + bkg.untagmu
       + bkg.acccoin + bkg.timcor + bkg.fot
       +bkg.spontfi + bkg.scin + bkg.buff
}.
```

Figure 6.28 shows the distribution of bkg (histogram) and the authors approximation (red curve). The two distributions are centered at fairly different values, our mean background value, 0.21, is 3 of the author's sigma away from their derived value of 0.15. The reason for this difference comes from the authors neglecting of upper limits in their analysis, even when the neglected background terms are larger than terms considered. For example, the (α,n) reaction in the buffer has an upper limit of 0.061 (and was neglected), but all but one of the considered background terms have smaller rates (and were considered).

In conclusion, we expect a background rate of 0.21 ± 0.04 in the Borexino search of anti-neutrinos. To be precise this is the estimate for an experiment ran as long as the background, 100 ton yr. The Borexino search has an exposure of 252.6 ton yr, so the background will be 2.526 larger. Borexino observed 21 events (obstot= 21). The signal to background ratio, computed as the authors do but with our estimate of the background, is then 39 (=(21−0.21*2.526)/(0.21*2.526)). The authors quote a larger value, 50, because of the underestimated background. As will be clear in a moment, what we refer to as signal in this section will partially be background in the next.

6.3.4.2 Discriminating Natural from Human-Induced Neutrinos

The Borexino experiment does not simply detect antineutrinos, but makes a step further, it attempts to discriminate between antineutrinos coming from natural reactions and those coming from human-induced reactions (i.e., in nuclear reactors). These authors estimate that neutrinos from reactors account for 9.4 ± 0.6 plus 5.0 ± 0.3 events in the Borexino experiment (and in presence of neutrino oscillations). The new background model reads:

```
model{
bkg.lihe ~ dnorm(0.03,pow(0.02,-2)) T(0,)
bkg.fastn ~ dexp(200)
bkg.fastn2 ~ dexp(58)
bkg.untagmu ~ dnorm(0.011,pow(0.001,-2)) T(0,)
bkg.acccoin ~ dnorm(0.080,pow(0.001,-2)) T(0,)
bkg.timcor ~ dexp(88)
bkg.fot ~ dexp(765)
bkg.spontfi ~ dnorm(0.003,pow(0.0003,-2)) T(0,)
bkg.scin    ~ dnorm(0.021,pow(0.002,2)) T(0,)
bkg.buff    ~ dexp(37)
bkg <- bkg.lihe+ bkg.fastn + bkg.fastn2 + bkg.untagmu
       + bkg.acccoin + bkg.timcor + bkg.fot
       +bkg.spontfi + bkg.scin + bkg.buff
bkg.react1 ~ dnorm(5.0, pow(0.3, -2))
bkg.react2 ~ dnorm(9.4, pow(0.6, -2))
bkg.tot <- 2.526 * bkg + bkg.react1 + bkg.react2
}.
```

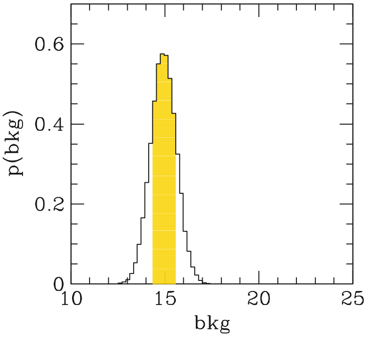

Fig. 6.29 Background rate probability distribution for the Borexino experiment, also included in the background are anti-neutrinos from nuclear reactors. The *shading* indicates the 68 % interval

Figure 6.29 shows the updated distribution of bkg: it is, unsurprisingly, (almost) a Gaussian centered on 14.9 and with $\sigma = 0.68$, because the last two background terms, assumed to be Gaussian, as the authors assumed, are one order of magnitude larger than the sum of all the other background terms and thus dominate the background.

We emphasize that the distribution above concerns bkg, i.e., the true background rate; it is a measure of how well we known the true background, not a statement on the distribution of observed values of the background. The latter will have the usual Poisson fluctuations, and thus their probability distribution will have at least a width of 4 counts (vs $\sigma = 0.68$ for the uncertainty of the true rate).

The Borexino experiment detected 21 events. We can finally compute the posterior probability of the amplitude of the geo-antineutrinos signal. For this aim, we just need to add to the model above the observation obstot= 21, the signal likelihood, taken to be Poisson, and the prior, taken to be uniform, i.e.,

```
obsbkg ~ dpois(bkg.tot)
obstot ~ dpois(s + bkg.tot)
s ~ dunif(0,100).
```

To sum up, the whole model reads:

```
model{
obsbkg ~ dpois(bkg.tot)
obstot ~ dpois(s + bkg.tot)
s ~ dunif(0,100)
bkg.lihe ~ dnorm(0.03,pow(0.02,-2)) T(0,)
bkg.fastn ~ dexp(200)
bkg.fastn2 ~ dexp(58)
bkg.untagmu ~ dnorm(0.011,pow(0.001,-2)) T(0,)
bkg.acccoin ~ dnorm(0.080,pow(0.001,-2)) T(0,)
bkg.timcor ~ dexp(88)
bkg.fot ~ dexp(765)
bkg.spontfi ~ dnorm(0.003,pow(0.0003,-2)) T(0,)
bkg.scin    ~ dnorm(0.021,pow(0.002,2)) T(0,)
bkg.buff    ~ dexp(37)
bkg <- bkg.lihe+ bkg.fastn + bkg.fastn2 + bkg.untagmu
```

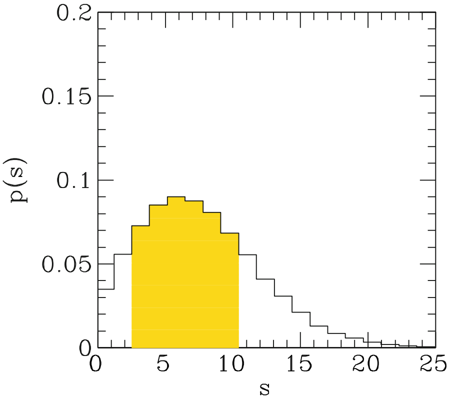

Fig. 6.30 Probability distribution of the number of geo-neutrinos of the Borexino experiment. Events of all light yield are considered. The *shading* indicates the 68 % interval

```
        + bkg.acccoin + bkg.timcor + bkg.fot +bkg.spontfi
        + bkg.scin + bkg.buff
bkg.react1 ~ dnorm(5.0, pow(0.3, -2))
bkg.react2 ~ dnorm(9.4, pow(0.6, -2))
bkg.tot <- 2.526 * bkg + bkg.react1 + bkg.react2
}.
```

Figure 6.30 shows the posterior distribution of the intensity of geo-antineutrinos: it has a mean of 7.5 but a large dispersion, 4.3, or quoting 68 % (highest posterior) intervals: [3.1, 10.7] (often quoted in articles as $7.6^{+3.1}_{-4.5}$). Figure 6.30 shows that a zero value for the signal is only half as likely as the most likely values, i.e., that it is not unlikely given the available data, confirming that geo-neutrinos are not detected with certainty using these data.

6.3.4.3 Improving Detection of Geo-Neutrinos

To improve the (so far unsuccessful) detection of geo-neutrinos, the authors applied stronger data cuts. Most of the neutrinos coming from reactors produce a light yield[3] larger than 1300 p.e. Therefore, the authors consider only the events (data) with $Q_{prompt} < 1300$ p.e. Under these new conditions, they observed 15 events when 5.0 ± 0.3 are expected from reactors and 0.31 ± 0.05 from background events.

The modeling of the (new) background is

```
bkg.react ~ dnorm(5.0, pow(0.3,-2))
bkg ~ dnorm(0.31, pow(0.05,-2))
bkg.tot <- bkg + bkg.react
obsbkg ~ dpois(bkg.tot).
```

[3] This is one of the many data features useful for discriminating interesting events from background events.

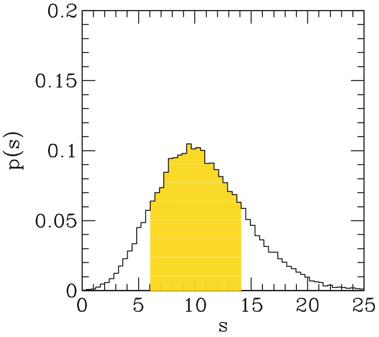

Fig. 6.31 Probability distribution of the number of geo-neutrinos of the Borexino experiment. Events with light yield < 1,300 p.e. are considered. The *shading* indicates the 68 % probability interval

As for the observation of 21 neutrinos, we can compute the posterior probability of the amplitude of the geo-antineutrinos signal by just adding, to the model above, the observation obstot= 15, with the likelihood, taken to be Poisson, and the prior taken to be uniform, i.e., adding

```
obstot ~ dpois(s + bkg.tot)
s ~ dunif(0,100).
```

To sum up the model reads:

```
model{
obstot ~ dpois(s + bkg.tot)
s ~ dunif(0,100)
bkg.react ~ dnorm(5.0, pow(0.3,-2))
bkg ~ dnorm(0.31, pow(0.05,-2))
bkg.tot <- bkg + bkg.react
obsbkg ~ dpois(bkg.tot)
}.
```

Figure 6.31 shows the posterior distribution of the incoming intensity of geo-antineutrinos: we found $10.6^{+3.3}_{-4.6}$ (mean and 68 % highest posterior intervals). The authors find a similar result, $9.9^{+4.1}_{-3.4}$, using a different analysis that makes use of data not tabulated in their article, and thus not usable by us. As it is fairly obvious in Fig. 6.31, the probability that the incoming flux of geo-neutrinos is zero is very small.

6.3.4.4 Concluding Remarks

This example shows the simplicity of accounting for a complex statistical model, i.e., how one may reach the scientific goal (determine the background rate, the probability of having detected geo-neutrino, etc.) without needing to use approximations, such as Gaussian distributions for positively defined quantities and neglecting upper limits. Some of these approximations have an important impact on some of the measured quantities.

6.3.5 Upper Limits from Counting Experiments

In this section, we will illustrate how to compute upper limits by re-analyzing some WIMP (Weakly Interacting Massive Particles) experiments. In such re-analysis, we put ourself in the authors' place at the time they published their work, making the same assumptions as they do (even if later papers revisited the original analysis).

At the time of this writing, we know that about 85 % of the Universe's matter is dark. We know what it is not (it is not-baryonic, non-luminous, and non-relativistic), but we do not know what it is. Amongst the many proposed Dark Matter candidates, Weakly Interacting Massive Particles (WIMP) occupy a special place, as they arise naturally from well-motivated extensions of the standard model of particle physics. Experimental searches (e.g., Ahmed et al. 2009) have not neatly detected WIMPs, but only reported upper limits on how much it interacts with ordinary matter (technically speaking, most experiments measured the WIMP-nucleon elastic-scattering spin-independent cross-section). In practice, WIMP experiments consist of applying cuts that select events compatible with WIMP and exclude as much as possible events compatible with interactions with known particles.

As it is often the case, there are other sources that interact with ordinary matter, i.e., a background, that we need to account for. Experiments have been built in a way that the background is usually expected to produce a few events at most, but the background rate is poorly known, basically because one cannot run a measurement of the background alone, i.e., with the WIMP signal switched off.

6.3.5.1 Zero Observed Events

Let us start by considering experiments that reported zero observed events (`obstot`$=0$). Because information concerning the background rate is, apart from the large emphasis on its importance, often coarsely reported because the difficulty of such a measurement, we now proceed to understand the importance of a precise characterization of the background rate when `obstot`$=0$. Very little, indeed: when `obstot`$=0$ the 90 % upper limit to the true number of WIMP events is 2.3 independent of the true background value, at least for background values between 0 and 10 (these values amply cover the plausible number of background events in real experiments). This result is obtained with the following model, that is the one introduced in Sect. 6.1.2, and simplified for the current case:

```
model{
obstot ~ dpois(s+bkg)
s ~ dunif(0,10)
```

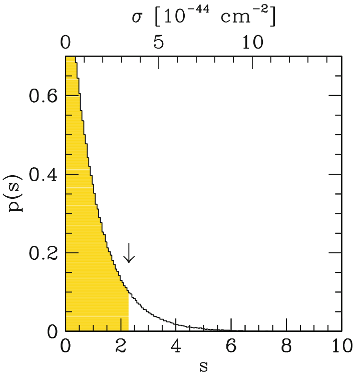

Fig. 6.32 Posterior probability of the cross-section for the XENON-100 experiment. The *arrow* indicates the 90 % upper limit

```
bkg <-0
# bkg <-10
}
```

and by comparing the 90 % upper limit on s derived for different bkg values. The problem is simple enough to have an analytic solution, which we leave for the reader as an exercise.

The independence of the 90 % upper limit on s with the true value of the background gives a (partial) justification of the procedure frequently used in the WIMP domain to ignore the information on the background value for deriving the upper limit. Of course, when the observed number of events is non-zero, the above justification does not hold.

Let us illustrate the case in detail for the XENON100 experiment (Aprile et al. 2010). They observed zero events in an experiment with a low background. Given the irrelevance of the background rate for deriving a 90 % upper limit for s in such conditions, we assume a uniform distribution for bkg,

```
bkg ~ dunif(0,10)
```

emphasizing that identical values would be obtained for another background prior. With this prior, our S+N model of Sect. 6.1.2 becomes:

```
model{
obstot ~ dpois(s+bkg)
s ~ dunif(0,10)
bkg ~ dunif(0,10)
}
```

and gives $p(\mathrm{s})$, the posterior distribution of the true number of WIMP events. This can be expressed in more familiar (to physics researchers) units by multiplying s by the count-to-physical conversion factor for a WIMP mass of 55 GeV, 1.48×10^{-44} cm^2. Figure 6.32 shows the posterior distribution for the WIMP cross section for a WIMP of 55 Gev. The posterior resembles an exponentially decreasing

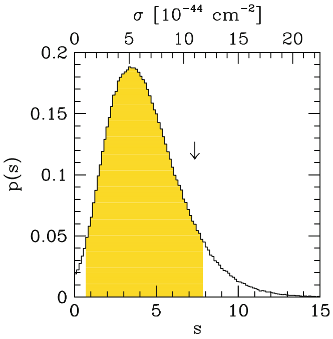

Fig. 6.33 Posterior probability of the cross-section for the EDELWEISS-II experiment. The *arrow* indicates the 90 % upper limit

function, implying that the cross-section is not large, but its value has not yet been identified, but only coarsely determined. The 90 % upper limit to the true value of number of WIMP events is 2.3, i.e., 3.4 10^{-44} cm^2, identical to the value derived by the authors (using a different procedure).

6.3.5.2 Non-zero Events

We now move to experiments having detected some events compatible with a WIMP. Let us now consider the preliminary results of EDELWEISS-II (Armengaud and et al. 2010). This experiment observed four events compatible with a WIMP, of which "1.6 events (90 % CL) of known origin [background, NDR] are expected for this WIMP search." This is the only information of statistical nature given for the background rate and, alone, it does not fully specify the prior on the background rate: there are an uncountable set of distributions for which the 90 % upper limit is 1.6.

We interpret the author's statement as the true value of background event is exponentially distributed with scale equal to 1.45 because this distribution (prior) has a 90 % upper limit equal to 1.6. The assumed exponential shape is the natural outcome of an experiment measuring the background having recorded no event. The background prior is therefore given by:

```
bkg ~ dexp(1.45).
```

Assuming a uniform prior for the signal s, the full model therefore reads:

```
model {
obstot ~ dpois(s+bkg)
bkg ~ dexp(1.45)
s ~ dunif(0,20)
}.
```

Fitting this model we obtain $p(\mathrm{s})$, the posterior distribution of the true number of WIMP events. It can be expressed in more familiar (to physicists) physical units by using the count-to-physical conversion factor of 1.5×10^{-44} cm^{-2} for a WIMP mass of 80 GeV. Figure 6.33 shows the posterior distribution for the WIMP (spin-independent) cross section in both raw units (lower abscissa) and in the standard physical units (upper abscissa). The 90 % upper limit to the true value of number of WIMP events is 7.4 in raw units, i.e., 11.1×10^{-44} cm^{-2}, to be compared to the more stringent, but preliminary, result, 5×10^{-44} cm^{-2} derived by the authors (they used a method that makes no use of the information on the background).

6.4 Exercises

Exercise 1

Checking numerically sampled posteriors. Kraft et al. (1991) analytically computed the posterior interval estimates for a source of intensity `s` in the presence of a background `bkg` (all expressed in counts), similarly to what we did in Sect. 6.1.2. In their paper, however, `bkg` is assumed to be perfectly known and they only consider the case $C = 1$. As in Sect. 6.1.2, $\texttt{obstot} \sim \texttt{dpois}(\texttt{s} + \texttt{bkg})$. Adapt the code in Sect. 6.1.2 to deal with this simplified case. Then, numerically compute some of the 90 % posterior intervals given in Table 1 of Kraft et al. (1991). Note that Kraft et al. (1991) assume an improper prior (i.e., a prior that is flat on the interval $(0, \infty)$) for `s` and JAGS does not currently have a method for implementing this but we can approximate this prior by using $\texttt{s} \sim \texttt{dunif}(0, \texttt{a})$ where `a` is large, say 1×10^7.

Exercise 2

Checking numerically sampled posteriors. Prosper (1985) analytically derived the posterior distribution for a source intensity `s` in the presence of an unknown background `bkg` (both expressed in counts) similarly to what we did in Sect. 6.1.2. In his paper, Prosper (1985) assumed, as we do, that the experiment measures `obsbkg` of $\texttt{bkg}/C$ where C is known. Prosper introduced the following prior,

$$p(\texttt{s}, \texttt{bkg}) = \frac{1}{(\texttt{s} * \texttt{bkg})^{\nu}},$$

which gives the analytical form of the posterior:

$$p(\texttt{s}|\texttt{obstot}, \texttt{obsbkg}) = \frac{\exp(-\texttt{s}) \sum_{j=0}^{\texttt{obstot}} \binom{\texttt{obstot}}{j} \texttt{s}^{\texttt{obstot}-j-\nu} \frac{\Gamma(\text{obsbkg}+j-\nu+1)}{(1+C)^{\text{obsbkg}+j-\nu+1}}}{\sum_{j=0}^{\texttt{obstot}} \binom{\texttt{obstot}}{j} \Gamma(\texttt{obstot} - j - \nu + 1) \frac{\Gamma(\text{obsbkg}+j-\nu+1)}{(1+C)^{\text{obsbkg}+j-\nu+1}}}.$$

Suppose we observe obstot $= 23$ and obsbkg $= 6$ and assume $C = 0.5$. For $\nu = 0$ (i.e., a dunif(0,a) for a large), manipulate the code in Sect. 6.1.2 to numerically sample from the posterior distribution and compare your results to the analytical posterior.

Exercise 3

Checking numerically sampled posteriors. D'Agostini (2004) addressed the determination of a fraction in the presence of a background, similarly to what we did in Sect. 6.1.3.[4] He derived the analytical expression of the posterior in the simpler case of perfectly knowing the mean number of background galaxies nbkg (with C$= 1$) and a perfectly known expected fraction of background blue galaxies fbluebkg. Adapt the code in Sect. 6.1.3 to deal with this case and plot the numerically computed posterior for obsntot$= 12$, obsbluetot$= 9$, and compare it with the analytically computed posteriors shown in his Fig. 6.10 for the listed nbkg (named λ_b there) and fbkg (named p_b there). Hint: you may find the following code useful:

```
model {
obsntot~dpois(nbkg+nclus)
obsbluetot~dbin(f,obsntot)
f <- (fbkg*nbkg+fclus*nclus)/(nbkg+nclus)
nclus~ dunif(1,1e+7)
fclus ~ dbeta(1,1)
}.
```

Exercise 4

Recycle. The model for computing the location and scale of a distribution illustrated in Sect. 6.1.1 can be revisited, with minor changes, to compute the luminosity function of lenticular galaxies in the Coma cluster of galaxies (as well as the parameters of everything with a Gaussian distribution) whose data[5] are in Michard and Andreon (2008). Assume 0.1 mag errors for all galaxies and also allow for uncertainty on the luminosity function normalization phistar. Hint: you may find the following code useful:

```
model {
for (i in 1:length(obsmag)){
mag[i]~dnorm(magcent,prec)
obsmag[i] ~ dnorm(mag[i],pow(0.1,-2)
```

[4] It would perhaps be more correct to say that the first author of this book adopted a Bayesian approach after Giulio D'Agostini wrote, on his request, this paper.

[5] The data can be found at http://www.brera.mi.astro.it/∼andreon/BayesianMethodsForThePhysicalSciences/comaS0FL.dat.R.

```
}
howmany~dpois(phistar)
phistar~dunif(0,100)
sigma ~ dunif(0,3)
magcent~dnorm(-19,1/4)
}
```

Exercise 5

Recycle. Modify the location and spread example (Sect. 6.1.1) to allow a Student-t likelihood with ten degrees of freedom to increase robustness against outlier measurements. Then change the Gaussian scatter into a Student-t likelihood with ten degrees of freedom to increase robustness against contamination by non-member galaxies.

Exercise 6

Suppose we observe $\texttt{obsm}= 2.06, 5.56, 7.93, 6.56, 2.05$ from a Gaussian distribution with mean $\texttt{m}$ and variance $\texttt{s}^2$. Assume for your priors $\texttt{m} \sim \texttt{dunif}(-10,10)$ and $\texttt{s} \sim \texttt{dunif}(0,50)$. Sample and summarize the posterior distribution for $\texttt{m}$, $\texttt{s}$ and $\texttt{m}/\texttt{s}$.

Exercise 7

Britt et al. (2002) provide density measurements on 27 different asteroids. The densities are (in g/cm^3) 2.12, 2.71, 3.44, 2.76, 2.72, 0.96, 2.00, 3.26, 2.35, 1.20, 1.62, 1.30, 1.96, 2.60, 4.60, 2.67, 4.40, 1.80, 4.90, 2.39, 1.62, 1.47, 0.89, 2.52, 1.21, 0.90, 0.80. Assume that this data comes from the same Weibull distribution, i.e., $\texttt{dens} \sim \texttt{dweib}(\texttt{a},\texttt{b})$. Assume the priors $\texttt{a} \sim \texttt{dunif}(0,50)$ and $\texttt{b} \sim \texttt{dunif}(0,50)$. Plot the prior and posterior for the 0.10 and the 0.90 quantiles of the Weibull distribution. To make things easier, the formula for the quantiles for the JAGS parameterization of the Weibull distribution is

$$d_p = \left[-\frac{1}{\texttt{b}}\log(1-p)\right]^{1/\texttt{a}},$$

where d_p is the quantile and $p \in (0,1)$.

Exercise 8

Fun with discrete probability distributions. Suppose we have observed the following data from a binomial distribution: 4, 6, 6, 7, 10, 8, 5, 7, 8, 9, 6, 7, 8, 8, and 6. Denote these as obsx. Then we are assuming obsx $\sim$ dbinom(n, f). Typically we know the value of n. But for this exercise we will instead estimate this quantity. Suppose a priori we know that n must be greater than 0 but less than or equal to 15 and take a discrete uniform distribution on the integers $1, 2, \ldots, 15$ (i.e., a prior assigning equal mass to each of the integer values just mentioned). Plot the prior and posterior distribution of f and n. Hint: You may find the following code useful:

```
model{
##Likelihood
for(i in 1:length(obsx)){
obsx[i]~dbin(f,n)
}
##Prior
f ~ dunif(0,1)
n ~ dcat(pi)
}
```

where pi is a vector giving the weights of the discrete uniform prior (pi=(1/15, 1/15,. . . ,1/15)).

Exercise 9

In a study to measure the speed of light in air, Michelson in 1882 made the following measurements: 850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650, 760, 810, 1000, 1000, 960, and 960. The given values are offset from +299000 km s^{-1}. Assume that speed values, denoted as speed, follow a gamma distribution, speed $\sim$ dgamma(a, b) and assume a $\sim$ dunif$(0, 1000)$ and b $\sim$ dunif$(0, 1000)$. The mode of the gamma distribution is calculated using

$$\text{mode} = \frac{\mathtt{a} - 1}{\mathtt{b}}.$$

Plot the prior and posterior distribution for the mode and provide the median and a 95 % probability interval for the posterior mode. Comment on the results, in particular with respect to the accepted value of 705(+299000).

Exercise 10

The following heat measurements (in Watts) were made on a Pu-238 fuel cell, `obsheat`= 54.581, 54.572, 54.503, 54.667, 54.697, 54.588, 54.574, and 54.466. Assume that these measurements were made close enough in time that we can ignore the effects of radioactive decay. These measurements are assumed to follow `obsheat` $\sim$ `dnorm(heat,prec)` where `prec` $=$ `pow(rho` $*$ `heat,` -2`)`, `heat` is the item's true heat at the time the measurements were made and `rho` is a relative standard deviation. Assume `heat` $\sim$ `dnorm`(54.59, 1000) and `rho` $\sim$ `dunif`(0, 1). Plot the prior and posterior distribution of `heat` and `rho`.

Exercise 11

Find a recent research paper that illustrates frequentist methods for the `S` $+$ `N` problem, preferably one that is relevant to your research so you can sculpt your own prior. Redo the analysis but instead use Bayesian methods. Compare the results from your Bayesian analysis to the frequentist approach. Is your choice of prior appropriate for the problem?

References

S. Andreon. The stellar mass fraction and baryon content of galaxy clusters and groups. *Monthly Notices of the Royal Astronomical Society*, 407:263–276, 2010a.

S. Andreon and A. Moretti. Do X-ray dark, or underluminous, galaxy clusters exist? *Astronomy & Astrophysics*, 536(A37), 2011.

S. Andreon, J. Willis, H. Quintana, I. Valtchanov, M. Pierre, and F. Pacaud. Galaxy evolution in clusters up to $z = 1.0$. *Monthly Notices of the Royal Astronomical Society*, 353:353–368, 2004.

S. Andreon, H. Quintana, M. Tajer, G. Galaz, and J. Surdej. The Butcher-Oemler effect at $z \sim 0.35$: a change in perspective. *Monthly Notices of the Royal Astronomical Society*, 365:915–928, 2006.

E. Aprile, K. Arisaka, F. Arneodo, A. Askin, L. Baudis, A. Behrens, and et al. First Dark Matter Results from the XENON100 Experiment. *Physical Review Letters*, 105(13):131302, 2010.

E. Armengaud and et al. Preliminary results of a WIMP search with EDELWEISS-II cryogenic detectors. *ArXiv e-prints*, page 1011.2319, 2010.

B. Binder, B. F. Williams, M. Eracleous, T. J. Gaetz, A. K. H. Kong, E. D. Skillman, and et al. The Chandra local volume survey: The X-ray Point Source Population of NGC 404. *The Astrophysical Journal*, 763:128, 2013.

BOREXINO Collaboration, G. Bellini, J. Benziger, S. Bonetti, M. B. Avanzini, B. Caccianiga, and et al. Observation of geo-neutrinos. *Physics Letters B*, 687: 299–304, 2010.

D. T. Britt, D. Yeomans, K. Housen, and G. Consolmagno. Asteroid density, porosity, and structure. In W. F. Bottke Jr., W. F. A. Cellino, P. Paolicchi, and R. P. Binzel, editors, *Asteroids III*, pages 485–500, 2002.

J. P. Brodie, C. Usher, C. Conroy, J. Strader, J. A. Arnold, D. A. Forbes, and et al. The SLUGGS Survey: NGC 3115, a critical test case for metallicity bimodality in globular cluster systems. *The Astrophysical Journal Letters*, 759:L33, 2012.

G. D'Agostini. Inferring the success parameter p of a binomial model from small samples affected by background. *ArXiv Physics e-prints*, 2004.

S. Ettori, A. Morandi, P. Tozzi, I. Balestra, S. Borgani, P. Rosati, and et al. The cluster gas mass fraction as a cosmological probe: a revised study. *Astronomy & Astrophysics*, 501:61–73, 2009.

R. P. Kraft, D. N. Burrows, and J. A. Nousek. Determination of confidence limits for experiments with low numbers of counts. *The Astrophysical Journal*, 374: 344–355, 1991.

R. Michard and S. Andreon. Morphology of galaxies in the Coma cluster region down to $M_B = -14.25$. I. A catalog of 473 members. *Astronomy & Astrophysics*, 490:923–928, 2008.

T. Park, V. L. Kashyap, A. Siemiginowska, D. A. van Dyk, A. Zezas, C. Heinke, and et al. Bayesian estimation of hardness ratios: modeling and computations. *The Astrophysical Journal*, 652:610–628, 2006.

M. Postman, M. Franx, N. J. G. Cross, B. Holden, H. C. Ford, G. D. Illingworth, and et al. The morphology-density relation in $z \sim 1$ clusters. *The Astrophysical Journal*, 623:721–741, 2005.

W. H. Press. Understanding data better with Bayesian and global statistical methods. In J. N. Bahcall and J. P. Ostriker, editors, *Unsolved Problems in Astrophysics*, pages 49–60, 1997.

H. B. Prosper. A Bayesian analysis of experiments with small numbers of events. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 241(1):236–240, 1985.

M. Sun, G. Voit, M. Donahue, C. Jones, W. Forman, A. Vikhlinin. Chandra Studies of the X-Ray Gas Properties of Galaxy Groups. The Astrophysical Journal, 693:1142 (2009)

Chapter 7
Non-random Data Collection

In an ideal world, the sample we are observing is randomly selected from the population we wish to study, and so any inference we draw about the sample applies to the population; in other words, we can ignore the way in which the data were collected. For example, if we are owners of a coffee shop and wish to know the opinion customers have about the new espresso, we may be tempted to randomly ask customers coming to our store and record their thoughts. By sampling easier-to-collect customers, those who are still customers after the introduction of the new espresso, we are introducing a bias against those who left our coffee shop because they do not appreciate the new coffee. Furthermore, if a one-out-of-many schema is used to select customers for their opinions, frequent customers will be over-represented in the sample, compared to the whole population of customers.

In astronomy and physics data collection is often not a random selection and we cannot ignore the sampling scheme (think of the coffee example). Rarely are the recorded events a random sample of the quantity (or population) being studied. Often, the recorded events are those that are easier to observe. In most cases, instruments record events only in a finite range of intensities (e.g., weak signals are missed) or energy. In some cases, instruments precisely record events in a given range and completely miss events outside that range. In other cases, boundaries are fuzzy and some of the events that are near the boundaries are lost. More generally, events are missed, with varying degrees, in the whole range where the instrument is sensitive, because no instrument is 100 % efficient. At the same time, some of the events outside, but near the boundaries, are recorded. In other terms, selection is probabilistic, not sharp. In some occasions, recorded events are only those that satisfy some conditions on the observed value. Other times, recorded events are those that satisfy a condition on the true value. In some cases, events are recorded over a wide range (e.g., energy, wavelength, etc), but the researcher is only interested on those in a narrower range and is not interested in modeling the data outside that range. He hence applies a selection that usually needs to be accounted for. Ignoring these selection effects usually leads to biased estimates.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_7

These effects have different names depending on the field: censoring, truncation (usually by truncation one means a harder-than-censoring effect: even the number of "hidden" data points is unknown), efficiency, contamination, sensitivity, or selection function.

In this chapter we want to address these different cases that are all related to a non-random data collection. We will start from the simplest non-random data collection: all events below a threshold are not observed and we will move toward a more soft (and realistic) selection: one in which the probability of inclusion in the studied sample does not jump from zero to one abruptly, as in the first example, but smoothly. We also consider the case of inferring the parameters that describe two common mixtures of distributions with non-random data collection in astronomy: colors and velocities. We defer discussion of dealing with non-random data collection in a regression application to Sect. 8.8.

Unless otherwise stated, in this chapter we assume that all information about lost events/data is forever lost, including information about how many events are missing. This is quite common in astronomy: we ignore how many events (asteroids, planets, galaxies, clusters, AGNs, etc.) there are below the detection threshold. Yet, we want to perform inferences in this (obliged) situation. Of course, knowing how many events are missed makes it possible to more precisely infer the target quantity.

Before proceeding, let us emphasize that in most practical situations the non-random data collection is similar to, but rather different from, the case assumed in the widespread tools used in survival and reliability analysis: the latter assumes sharp, deterministic, selection (e.g., all measurements larger or smaller than a value are lost) when instead, as mentioned, selection is usually stochastic. Furthermore, survival and reliability analyses typically know the number of missing records (for example, the researchers know how many patients are used in the clinical trial), while this is not known in astronomy and physics. Adopting these techniques would not be applicable. To further complicate the issue, what is called truncation and censoring presents differences even in languages as similar as JAGS and openBUGS.

7.1 The General Case

Suppose that in absence of data collection issues (e.g., selection effects) the likelihood is $p(x|\theta)$ (e.g., the common Gaussian).

Consider a sample affected by some non-random data collection (e.g., selection effect), mathematically characterized by the function S. S could be as simple as a step function (zero below a threshold, and one above). In a realistic example, S is usually a smooth function, for example an error function of some input (e.g., energy, flux, etc.), that we collectively call x here, i.e., S is $S(det|x)$. It describes the probability that the data is collected (recorded, in the sample, etc.) if the value is x.

In the presence of selection effects, the likelihood of observing a value x is given by the likelihood above, $p(x|\theta)$, times the data collection (selection) function $S(det|x)$:

$$p(x|det,\theta) = \frac{p(x|\theta)S(det|x)}{Z(\theta,det)}, \tag{7.1}$$

where

$$Z(\theta,det) = \int p(x'|\theta)S(det|x')dx'. \tag{7.2}$$

The denominator is just the integral over the range of x and is requested because probabilities by definition must integrate (sum) to one. Said simply, the new likelihood is the selection-free likelihood, modified by the selection function which is then normalized to one so that we have the probability that describes the problem we are modeling. Notice that this is a straightforward application of Bayes theorem where we replace $p(x)$ in Eq. (2.8) with $S(det|x)$.

Reformulated in a more rigorous way, suppose we introduce a new binary variable (i.e., a variable taking only two values 1 and 0) D_i for observation i where $D_i = 1$ if i is observed and zero otherwise and D_i depends only on the value of the underlying x_i. The likelihood of observing a value x_i^{obs}, denoting other parameters of the model collectively by θ, can then be expressed as

$$f(x_i^{obs}|D_i = 1, x_i, \theta) = \frac{f(D_i = 1|x_i)f(x_i^{obs}|x_i,\theta)}{\int f(D_i = 1|z)f(z|x_i,\theta)dz} \tag{7.3}$$

where the integral in the denominator is across the range of x_i^{obs}, $f(x_i^{obs}|x_i,\theta)$ is the usual likelihood when there are no selection effects and $f(D_i = 1|x_i)$ is the sample selection function. The integral in the denominator of Eq. (7.3) gives the probability of selection (inclusion in the sample) given the values x_i and θ.

At this point, a computational problem may arise. In certain situations, for example if the numerator in Eq. (7.1) (or Eq. 7.3) is among the available distributions in JAGS (e.g., a truncated Gaussian), its coding in JAGS is straightforward: we only need to write the JAGS name of that distribution and proceed as usual. In general, when the numerator of Eq. (7.1) is not among the available distributions in JAGS, we will need to carefully code the likelihood by writing its mathematical expression. In such a case, great care should be practiced when implementing the correct equation (i.e., Eq. 7.1) without making mistakes (some mistakes are known in both the physics and astronomical literature). We also need to code a way to sample from the likelihood, and finally, we need to be able to compute the integral Z (for example, using the Newton method). This causes a (potentially) clear and short model which can be implemented in a few lines of code in JAGS (e.g., 5 lines long say) to grow into a program requiring several tens of lines of code and causes the model to lose most of its clarity, because 90 % of the code deals with the details of the technical implementation, and not with the statistics. For clarity, in this chapter models are presented twice: once in the short and clear form (not always understood by JAGS, and in the latter case starting by `symbolic-model`) and once in the long version (always understood by JAGS, starting by `model` as usual). For easier reading,

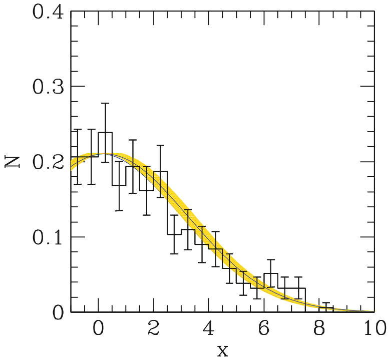

Fig. 7.1 Simulated data (represented by the *histogram*) from the truncated Gaussian distribution represented by the *curve*, with the fitted truncated Gaussian model (represented by the *shading about the curve*). The *shading* indicates the 68 % interval of the truncated Gaussian model

the longer code examples and the implementation details are postponed and illustrated at the end of the chapter. Although following this chapter may seem difficult (especially in later sections), we emphasize that not accounting for the selection effects in your model through the likelihood will lead to false conclusions and generalizations.

7.2 Sharp Selection on the Value

We begin with a simple case. Consider an instrument (or a survey) that misses all events for which some property of our population, denoted by x, is less than some threshold, x[i]< xtresh. It is also unknown how many events are missed by the instrument (or the survey).

Consider a Gaussian model truncated at the threshold xtresh$= -1$. Figure 7.1 shows simulated data consisting of roughly 300 values, x[i], represented by the histogram. Note that we want to fit the individual data values and not the histogram with arbitrary bins depicted in the figure. From the observed data, we want to recover the center of the (untruncated) Gaussian. In this very first example we assume that the x[i] are observed without error.

The truncated Gaussian is among the distributions programmed in JAGS and therefore xcent can be estimated by the simple model:

```
model {
 for (i in 1:length(x)) {
  x[i]~dnorm(xcent,pow(sigma.intr,-2)) T(-1,)
 }
 sigma.intr <- 3
 xcent ~ dunif(-10,10)
},
```

where `T(-1,)` is the JAGS function for implementing a sharp threshold (note the `T` for truncation). Conceptually, non-random data selection is not much different than other models we have so far presented. As always, one just needs to specify a prior and a likelihood and to let the computer compute.

Returning to Fig. 7.1, the recovered truncated Gaussian is displayed by the shading. We found `xcent`$= 0.23 \pm 0.24$, in good agreement with the value used to generate the data (`xcent`$= 0$). Alternatively, we may infer the value of `sigma.intr`, with `xcent` known. This may be achieved by replacing

```
sigma.intr <- 3
xcent ~ dunif(-10,10)
```

with

```
sigma.intr ~ dunif(0,10)
xcent <- 0 .
```

We found `sigma.intr`$= 3.15 \pm 0.14$, in agreement with the input value used to generate the data (3).

We suggest caution when attempting to infer about more than one parameter from truncated data such as those depicted in Fig. 7.1 because very often there are combinations of parameter values that fit the model just as good as other combinations. For our data set and for similar ones (e.g., heavily truncated ones), data can almost equally well be fitted by the blue curve, or a model with a lower `xcent` and an (appropriately) larger `sigma.intr`. Or said more intuitively, one cannot state the width of the distribution by not knowing where its center is.

The simulated data in Fig. 7.1 is generated by:

```
x ~ dnorm(0,pow(3,-2)),
```

where we discard values such that `x[i]`$<$ `xtresh`$= -1$.

7.3 Sharp Selection on the Value, Mixture of Gaussians: Measuring the Gravitational Redshift

In general relativity, the light emitted from a galaxy located well inside the potential well of a galaxy cluster is expected to be redshifted because photons must lose energy (change their frequency) in order to escape the potential well. The amount of energy lost is proportional to the depth of the potential well at the location where the photon is emitted. For a typical cluster, this effect, named gravitational redshift, is of the order of 10 km/s. The random motion of galaxies in clusters also induces a change of frequency, named the Doppler shift, which is on average two orders of magnitude larger than the gravitational redshift. However, the latter gives rise to a symmetric broadening of the observed velocity distribution (because it can be positive or negative), whereas the former shifts its centroid (because it only decreases frequencies), allowing us to disentangle the two effects and to test general relativity on a spatial scale never probed before (and not accessible elsewhere). In short,

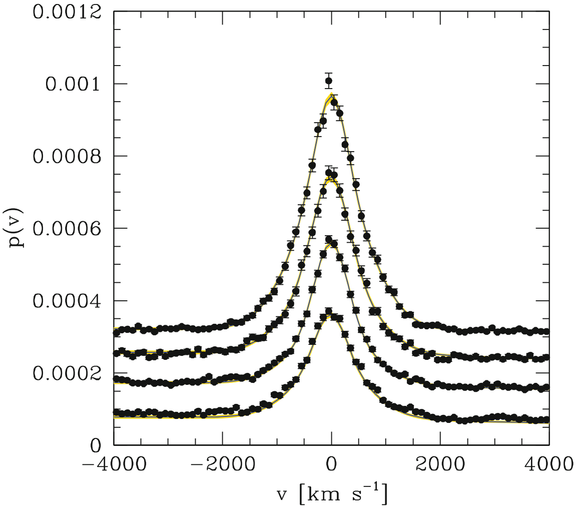

Fig. 7.2 Gravitational redshift fit. The *solid line* marks the mean fitted velocity distribution, whereas the *shaded region* (barely visible) indicates the 68 % interval. Points are observed values, with error bars set to $\sqrt{n}$. For display purposes only, the distributions for bins of cluster centric distance are vertically offset

the barycenter of the velocity distribution of a cluster of galaxies' center should appear more red-shifted (i.e., should display a larger velocity) compared to the outer regions.

In order to measure the shift, Wojtak et al. (2011) used relative velocities, v, of about 120,000 galaxies lying within ±4,000 km/s and within 6 Mpc from the barycenter of a (composite) cluster. Our forthcoming analyses incorporate the data from Wojtak et al. (2011) (which were kindly provided by the first author[1])

The sample consists of both cluster and background galaxies, the latter being galaxies satisfying the spatial and velocity constraints mentioned above but that are not gravitationally bound to the cluster. As remarked by the authors, close inspection of velocity histograms (Fig. 7.2) shows that the distribution of the background is almost uniform, with a slight excess of galaxies with negative velocities (because of observational effects). For computational simplicity, we model this background component using a Gaussian with a large sigma (larger than 5,000 km/s), with center, `lambda[1]` set at −4,000 km/s and truncated outside the [−4,000,4,000] km/s interval. A Gaussian with such a large sigma is indistinguishable from a linearly decreasing function in the considered range. We take a uniform prior for the background precision (one over sigma square), between $1/50{,}000^2$ (larger values ends in indistinguishable functions) and $1/5{,}000^2$ (because we want a sigma larger than 5,000 km/s). This reads:

[1] Data are available at http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/Wojtak.dat.R.

```
lambda[1] <- -4000
tau[1] ~ dunif(pow(50000,-2),pow(5000,-2)) .
```

As noted by the authors, close inspection of the velocity histogram in the range where it is dominated by cluster galaxies (say [−2000,2000] km/s) shows that its shape is close to a Gaussian, just a bit more peaked. Following Wojtak et al. (2011), we model the velocity histogram with a mixture (sum) of two Gaussians having the same centers, `lambda[2]=lambda[3]` and different sigmas. We take a uniform distribution for `lambda[2]` in a range large enough to certainly include the true distribution's center, [−500,500] km/s:

```
lambda[2] ~ dunif(-500,500)
lambda[3] <- lambda[2] .
```

For the sigmas of these two Gaussians (modeling the galaxy cluster distribution), we take uniform priors on their squared reciprocal (for computational simplicity):

```
tau[2] ~ dunif(pow(2000,-2),pow(100,-2))
tau[3] ~ dunif(tau[2],pow(100,-2))
```

with the component with index 3 taken to have the smallest sigma. Of course, these distributions are also truncated outside the [−4000,4000] km/s range. In summary, the model consists of a mixture of three Gaussians truncated outside the [−4000,4000] km/s interval:

```
model {
for (i in 1:length(v)){
v[i] ~ dnorm(lambda[I[i]],tau[I[i]]) T(-4000,4000)
I[i] ~ dcat(p[])
}
```

where `I[i]` is the (usual for mixtures) indicator variable that tells which component galaxy *i* belongs to.

Finally, the prior for the proportions (relative abundances) of the three components is

```
alpha[1] <-1
alpha[2] <-1
alpha[3] <-1
p[1:3] ~ ddirch(alpha[1:3])
```

to have `I[i]` uniformly distributed and $p[1]+p[2]+p[3]=1$.

To summarize, the model reads:

```
model {
for (i in 1:length(v)){
v[i] ~ dnorm(lambda[I[i]],tau[I[i]]) T(-4000,4000)
I[i] ~ dcat(p[])
}
lambda[1] <- -4000
tau[1] ~ dunif(pow(50000,-2),pow(5000,-2))
lambda[2] ~ dunif(-500,500)
lambda[3] <- lambda[2]
```

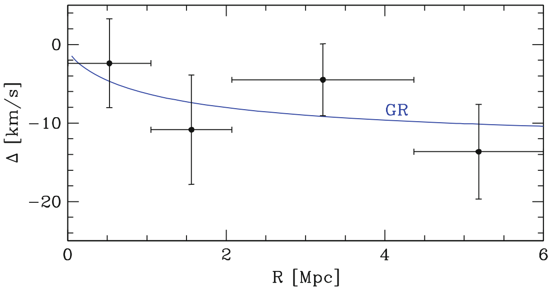

Fig. 7.3 Observed gravitational redshift (points with errors) as a function of the cluster's centric distance. The expectation from General Relativity is shown as *solid line*

```
tau[2] ~ dunif(pow(2000,-2),pow(100,-2))
tau[3] ~ dunif(tau[2],pow(100,-2))
alpha[1] <-1
alpha[2] <-1
alpha[3] <-1
p[1:3] ~ ddirch(alpha[1:3])
} .
```

We fit the four datasets, one per radial range considered by the authors: $0 < R < 1.1$ Mpc, $1.1 < R < 2.1$ Mpc, $2.1 < R < 4.4$ Mpc and $4.4 < R < 6.6$ Mpc.

Figure 7.2 shows the velocity distribution of the 120000 considered galaxies and the fitted models. The distributions are vertically shifted for display purpose.

Figure 7.3 shows the velocity shifts `lambda[2]` as a function of the distance from the cluster's center. Gravitational redshift manifests itself as a blue-shift (negative velocity) of galaxies with increasing cluster centric distances, indicating that the deepest part of the cluster (the center) are redshifted (have a larger mean velocity) than the outer regions. Again, this corresponds to what is expected from the gravitational-redshift result from general relativity (indicated by the curve).

Figure 7.4 shows the posterior distribution of three of the model parameters for the fit of the galaxies in the outermost radial range.

7.4 Sharp Selection on the True Value

In this section we add two grains of reality, measurements now have errors and these errors have different sizes for different points.

We again want to determine the center, `xcent`, of a Gaussian distribution with width `sigma.intr` after having observed 100 values `obsx[i]`, with Gaussian heteroscedastic errors `err.x[i]`. These 100 values are selected by `x[i]` > `xthresh`= 3. For this example, we simulated the data from a Gaussian with parameters `xcent`= 0 and `sigma.intr`= 3. Note that with this distribution and our threshold, our data is heavily truncated, missing about 80 % of the data (because the data are kept in the sample only if they are above about 1 `sigma.intr`).

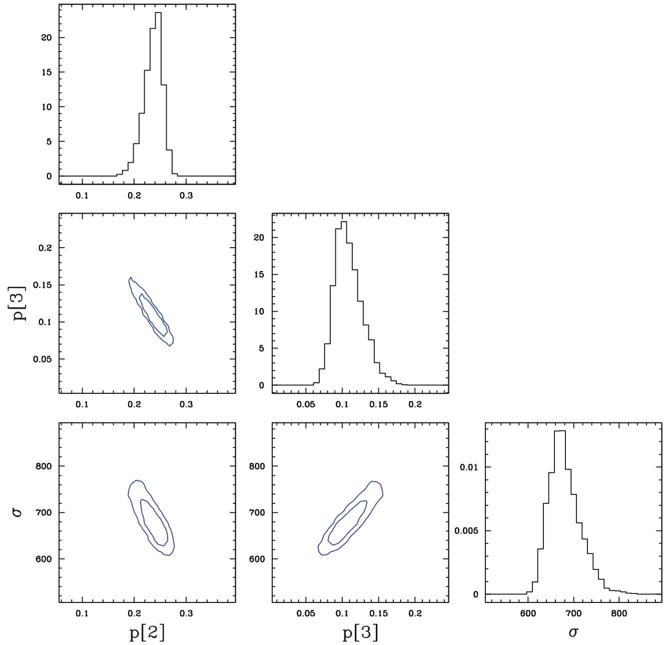

Fig. 7.4 Posterior probability distribution for the proportion of the first and second population (`p[1]` and `p[2]`) and for the dispersion of the second one ($\sigma[2]$) of the gravitational redshift fit in the outermost radial range. The *histograms* show the posteriors, marginalized over the other parameters. The contour plots on the off-diagonal panels refer to the 68 % and 95 % probability levels

From the 100 `obsx[i]` values in the sample (shaded histogram in the top-right panel of Fig. 7.5), we want to infer the value of `xcent` (far away from the sampled range!). The top-left panel shows what is not observed: the distribution (the curve) from which the `x[i]` values are drawn, the drawn `x[i]` values (histogram) and the selected `x[i]`'s (shaded histogram). The top-right panel shows all `obsx[i]` values (histogram) and those that pass the selection (shaded histogram). Only the latter are available, and from these we want to reconstruct the blue curve in the top-left panel. As in previous sections, we know nothing about how many observations are missed. The main complication added to this example is that `x[i]` values are never observed and we have instead observed `obsx[i]` with error (`err[i]`).

We only need to make a small modification to the model in Sect. 7.2, i.e., to model the Gaussian measurement errors (as is often done in this book). The updated model is:

```
model {
for (i in 1:length(obsx)) {
obsx[i] ~ dnorm(x[i], pow(err[i],-2))
x[i] ~ dnorm(xcent, pow(sigma.intr,-2)) T(3,)
}
```

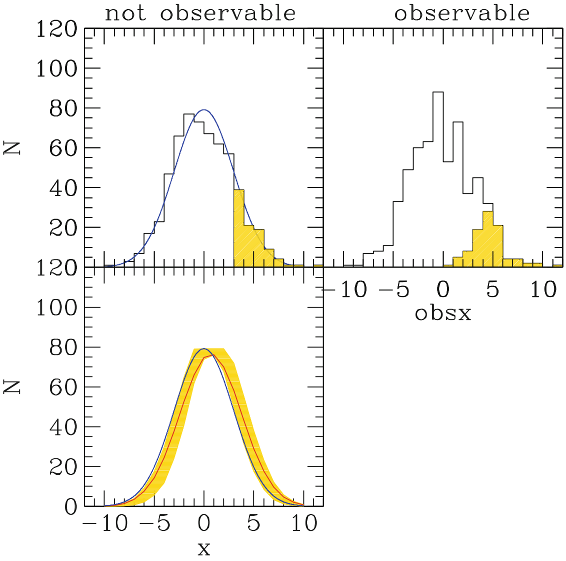

Fig. 7.5 Sharp truncation on the true value. *Top-left panel:* histogram of unobservable x and the distribution from which they are drawn (*blue curve*). The shading indicates events that pass the threshold x> 3. *Top-right panel:* histogram of observable obsx. The *shading* indicates events that pass the threshold x> 3. Only these are available for the analysis. *Bottom-left panel:* input distribution (*blue curve*), fitted model (*blue curve*), and 68 % interval of the model (*yellow shading*)

```
C <- 10
sigma.intr <- 3
xcent ~ dunif(-10,10)
}.
```

Again, to handle this non-random data selection, we just need to specify, as always, the prior (a uniform) and the likelihood.

By fitting the model, we found: xcent$= 0.38 \pm 0.79$, in good agreement with the value used to generate the simulated data (0). The error for xcent is large because we are attempting to determine the location of the peak of the distribution after only observing a tail of the distribution (the region around the peak is not observable). The bottom-left panel of Fig. 7.5 shows the input distribution (blue curve) and the inferred one (red curve) and its 68 % errors (shading).

The model used to generate the data reads:

```
model {
obsx ~ dnorm(x,prec.err)
x ~ dnorm(0, pow(3,-2))
prec.err ~ dunif(0.8, 1.2)
}
```

and we discarded obsx[i] values for which x[i]< 3. As indicated, we allowed the data obsx[i] to have different errors, which are drawn by selecting precisions from a uniform distribution bounded between 0.8 and 1.2.

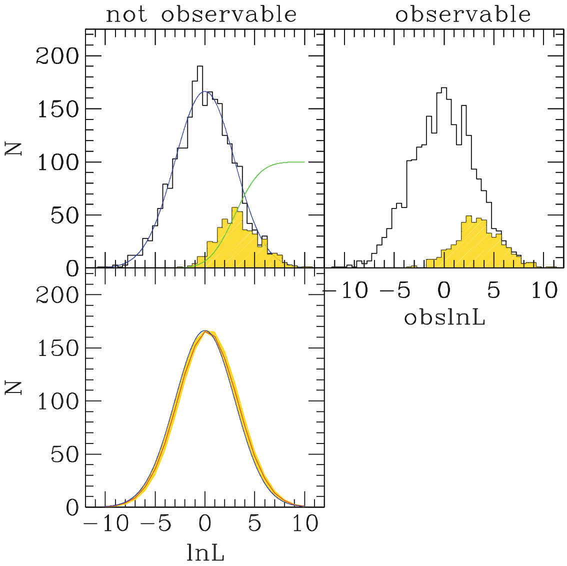

Fig. 7.6 Soft selection on the true value. *Top-left panel:* histogram of unobservable `lnL` and the distribution from which they are drawn (*blue curve*). The shading indicates events included in the sample. The probability, multiplied by 100 for readability, of being included in the sample is plotted in *green*. *Top-right panel:* histogram of observable `obslnL`. The *shading* indicates events included in the sample. Only these are available for the analysis. *Bottom-left panel:* input distribution (*blue curve*), fitted model (*blue curve*) and 68 % interval of the model (*yellow shading*)

7.5 Probabilistic Selection on the True Value

We are now ready to address a more complex case, we no longer have a sharp threshold, but instead the probability of detection goes smoothly from zero to one. In other words, some events are lost at values higher than the “threshold,” and events are included at values lower than the “threshold.” We are addressing a soft selection which is implemented using a selection function.

For example, suppose we want to estimate the mean (log of) X-ray luminosity of galaxy clusters, `lnLhat` (of clusters of a given mass or richness, but this does not matter here), disposing of a heavily censored sample of measurements, `obslnL`, each with its own error, `errlnL`, and knowing (e.g., from Monte Carlo simulations) that a cluster enters in the sample with probability given by an error function. This same function with a different parametrization is called ϕ in the statistics domain. For our example, we take an error function equal to:

$$p(\texttt{lnL}) = \phi\left(\frac{\texttt{lnL} - 3}{2}\right) \tag{7.4}$$

depicted as the green line in the top-left panel of Fig. 7.6.

Figure 7.6 illustrates the situation: the top-left panel shows the distribution of (log of) X-ray luminosities (blue continuous line), from which the unobservable (true values of) X-ray luminosities (histogram) are drawn. The probability (multiplied by 100 for readability) that an object (cluster) is included in the sample is indicated by a green line. About 480 clusters turn out to be included in the sample, and their histogram is shaded. Note how much of the population (`lnL[i]` values) is missed by the selection function, i.e., how one collects the brightest objects only. As mentioned, we only observe noisy versions of the `lnL[i]` included in the sample, `obslnL[i]`, shaded in the top-right panel, with errors `err.lnL[i]`. From these, we want to infer the location `lnLhat` of the Gaussian from which the `lnL[i]` data are drawn.

The (symbolic) fitting model is almost identical to the previous one, except for the fact that we have now replaced the sharp selection with the soft one, named here `phi((lnL-3)/2)` and changed the name to variables to reflect their current meaning. The model reads:

```
symbolic-model {
for (i in 1:length(obslnL)) {
 obslnL[i] ~ dnorm(lnL[i],pow(err.lnL[i],-2))
 lnL[i] ~ dnorm(lnLhat,pow(sigma.intr,-2)) phi((lnL-3)/2)
}
sigma.intr <- 3
lnLhat ~ dunif(-10,10)
} ,
```

where we adopted a uniform prior for `lnLhat`. As in the previous sections, to deal with a non-random data selection we just need to specify the prior and the likelihood. However, unlike the previous sections, this model cannot be easily fitted using JAGS, and its numerical implementation, described in Sect. 7.7, requires some efforts.

We found: `xcent`$= 0.26 \pm 0.22$, in good agreement with the value used to generate the simulated data (0). The bottom-left panel of Fig. 7.5 shows the input distribution (blue curve) and the inferred one (red curve) and its 68 % interval (shading).

The data are generated using:

```
model {
obsLnL ~ dnorm(lnL,prec.err )
lnL ~ dnorm(lnLhat, prec.intr)
prec.intr <- 1/9.
prec.err ~ dunif(0.8, 1.2)
lnLhat <- 0
}
```

and we discarded `obsLnL[i]` values according to Eq. (7.4). As indicated in the model, we allowed the data `obsx[i]` to have different errors by drawing precisions from a uniform distribution bounded between 0.8 and 1.2.

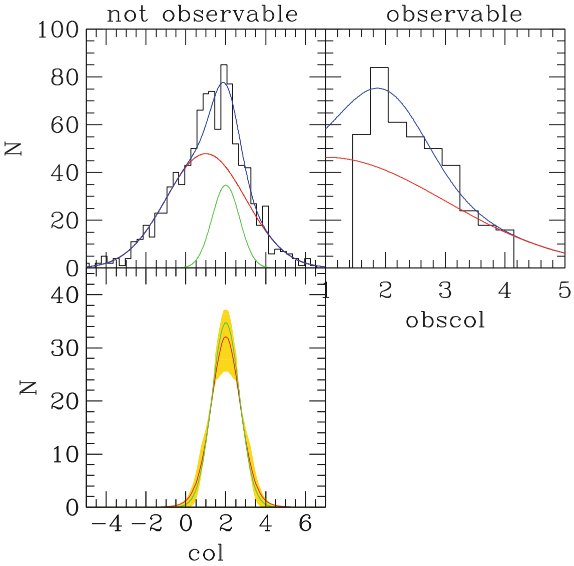

Fig. 7.7 Sharp selection on the observed value for mixtures. *Top-left panel:* histogram of unobservable `col` and the distributions from which they are drawn (*red and green curves*). The *blue curve* gives the sum of the two distributions. *Top-right panel:* histogram of observable `obscol`. The curves of the previous plot are reported. *Bottom-left panel:* input distribution (*green curve*), fitted model (*red curve*) and 68 % interval (*yellow shading*) for the distribution of the population of interest

7.6 Sharp Selection on the Observed Value, Mixture of Gaussians

As mentioned in the introduction of this chapter, sometimes we are not interested in modeling the distribution from which the data are drawn outside a given range (because these data are corrupt, or not available, too hard to model, not interesting, etc). Data outside the range are therefore simply ignored.

The top panel of Fig. 7.7 exemplifies the situation, for a (fictitious) quantity named `col[i]` (for color). Unobservable (true) data are drawn from a mixture (sum) of two Gaussians, centered on `lambda[1]` $= 2$ and `lambda[2]` $= 1$ and with σ's equal to $1/\sqrt{2.1}$ and 2. In the first author's mind, the second population is a nuisance (background) population, which contaminates the interesting population (the Gaussian with the larger `lambda` and a narrower distribution). As mentioned, values of `obscol[i]` outside the range $[1.5, 4]$ are ignored (or lost, corrupted, or whatever) and we only deal with values inside the range. Our simulated data consist of 240 values of `obscol[i]`, of which about 80 % comes from the background (but we do not have this information when fitting the model). To inject a grain of reality, `obscol[i]` have Gaussian errors `err[i]` which vary from point to point. Basically, we only have events shown in the top-right panel of Fig. 7.7 and from those observations we want to infer, with high precision, the parameters from the

Gaussian distribution with `lambda[1]`$=2$ and $\sigma = 1/\sqrt{2.1}$ (of course these are unknown).

In a mixture of two Gaussians there are five parameters that one may potentially infer: the centers and spreads of the two Gaussians and the proportion of events that belong to these distributions. One may infer all five of these parameters if the data allows it, or perhaps the two parameters related to the distribution of interest (e.g., the interesting population). In our case, we assume we know all but one of the parameters, either from previous data or from theoretical calculations, where the unknown parameter is the width of the Gaussian from the signal's distribution.

To infer the width of the Gaussian's signal, we just need to adopt previous models to the new setting. The symbolic model (not to be run in JAGS[2]) reads:

```
symbolic-model {
for (i in 1:length(obscol)) {
obscol[i] ~ dnorm(col[i], prec[i]) T(1.5,4)
col[i] ~ dnorm(lambda[type[i]], prec.intr[type[i]], p)
type[i]~dcat(p[])
}
# Prior
lambda[1] <- 2
lambda[2] <- 1
p[1] <- 0.2
p[2] <- 1 - p[1]
prec.intr[1] ~ dunif(0.1,10)
prec.intr[2] <- 0.25
} .
```

The fitted model adopts a uniform prior for the precision of the population of interest. The signal for our Gaussian is estimated to be 0.62 ± 0.27, in agreement with the input value, 0.69 $(= 1/\sqrt{2.1})$.

To generate data distributed as a sum of Gaussians, we use the JAGS distribution `dnormmix`. The model used to generate the simulated data is:

```
model{
obscol ~ dnorm(col, prec)
col ~ dnormmix(lambda, prec.intr, p)
lambda[1] <- 2
lambda[2] <- 1
p[1] <- 0.2
p[2] <- 1 - p[1]
prec.intr[1] <- 2.1
prec.intr[2] <- 0.25
err ~ dunif(0.05,0.15)
prec <- pow(err,-2)
},
```

where we discard all data outside the range $1.5 <$`obscol`< 4.

[2] A special thanks goes to Martin Plummer, the JAGS author, who clarifies that the coding below corresponds to a likelihood different from the one we are interested in. Our likelihood is given in Sect. 7.7.4.

7.7 Numerical Implementation of the Models

As mentioned in Sect. 7.1, the JAGS coding of the models may be quite involved when the data collection is non-random and the numerator of Eq. (7.1) is not among the distributions available in JAGS. In such a case, we need to code the likelihood by writing its mathematical expression, we need to code a way to sample it, and finally we need to compute by ourselves the integral Z (for example, using the Newton method). In this section we go through some of the models of the chapter, starting with how to implement the non-random data collection, first in cases addressed by the JAGS `T` function, and, in later subsections, in cases not easily dealt by JAGS.

7.7.1 Sharp Selection on the Value

The JAGS model adopted in Sect. 7.2 reads:

```
model {
for (i in 1:length(obsx)) {
x[i]~dnorm(xcent,pow(sigma.intr,-2)T(-1,)
sigma.intr <- 3
xcent ~ dunif(-10,10)
} .
```

For now we will ignore this elegant solution to prepare ourselves for more complex situations not addressed by JAGS.

In this case, Eqs. (7.1) and (7.2) require us to write the expression of a Gaussian, and also require that the integral above `xtresh`$= -1$ is equal to 1. To sample from a generic distribution (the Gaussian in our case) we use the zero trick (Sect. 3.2), and to compute the integrals we write our own subroutine in JAGS (deliberately ignoring that JAGS has a function returning this number). We implement the integration in the simplest way, the histogram (Newton) method: we measure the height of the function to be integrated on a grid `grid.x[k]`. Then we sum the products of these heights and the grid step, `step.grid.x`. The model, in JAGS, reads:

```
data  {
 # grid for integral evaluation
 for (k in 1:400){
  # start from -1, because integration start here
  grid.x[k] <- (k-21.)/20.
 }
 step.grid.x <-grid.x[2]-grid.x[1]
 # dummy variable for zero-trick
 # to sample from a distribution not available in JAGS
 for (i in 1:length(obsx)) {
  dummy[i] <-0
 }
}
```

```
model {
  # compute the integral at denominator
  for (k in 1:length(grid.x)) {
  normaliz[k] <- sqrt(prec/6.28)*
                exp(-prec*pow(xcent-grid.x[k],2)/2.)
  }
 tot.norm <- sum(normaliz[])*step.grid.x

 for (i in 1:length(obsx)) {
  # sampling from an unavailable distribution
  dummy[i] ~ dpois(loglike[i])
  # computing the sampled distribution
  numerator[i] <- sqrt(prec/6.28)*
                 exp(-prec*pow(xcent-obsx[i],2)/2.)
  loglike[i] <- -log(numerator[i]/tot.norm)+C
 }
C <- 100.
sigma.intr <- 3
prec <- pow(sigma.intr,-2)
xcent ~ dunif(-10,10)
} .
```

The first for loop prepares the grid grid.x[k] needed for the integral evaluation, the second loop prepares the dummy values for the zero trick. Then, we evaluate the integral of the Gaussian over $[-1,\infty)$.

The model, about four lines long in the elegant version, has grown by a factor of 8, losing much of its clarity. Yet, it is important to be able to apply such methods when an elegant solution is not available.

7.7.2 Sharp Selection on the True Value

The model described in Sect. 7.4 reads:

```
model {
for (i in 1:length(obsx)) {
 obsx[i] ~ dnorm(x[i], prec.err[i])
 x[i] ~ dnorm(xcent, pow(sigma.intr,-2)) T(3,)
}
C <- 10.
sigma.intr <- 3
xcent ~ dunif(-10,10)
} .
```

Again, we will ignore this elegant solution, and solve this problem using the brute force approach. We now implement the sharp selection with the JAGS function step(x-3), which takes zero value below 3 and 1 above. Equation 7.1 is coded in JAGS by computing separately the three terms of which it is composed, correcc (the function $S(det|x)$), tot.norm (the integral Z), and x[i] $\sim$ dnorm(xcent, pow(sigma.intr,-2)) (the likelihood in absence of selection effects).

```
data
{
 # grid for integral evaluation
 for (k in 1:400){
  grid.x[k] <- (k-200.)/20.
  correcc[k] <- step(grid.x[k]-3)
 }
 step.grid.x <-grid.x[2]-grid.x[1]
 # dummy variable for zero-trick
 # to sample from a distribution not available in JAGS
 for (i in 1:length(obsx)) {
  dummy[i] <-0
 }
}
model {
  # compute the integral at denominator
  for (k in 1:length(grid.x)) {
  normaliz[k] <- sqrt(prec.intr/6.28)*exp(-prec.intr*
                  pow(xcent-grid.x[k],2)/2.)*correcc[k]
  }
 tot.norm <- sum(normaliz[])*step.grid.x
 for (i in 1:length(obsx)) {
 obsx[i] ~ dnorm(x[i], prec.err[i])
 # the following would suffice, if no selection
 x[i] ~ dnorm(xcent, prec.intr)
 # but selection is there. Adding the additional likelihood term
 # sampling from an unavailable distribution
 numerator[i] <- step(x[i]-3)+1e-14
 loglike[i] <- -log(numerator[i]/tot.norm)+C
 # computing the sampled distribution
 dummy[i] ~ dpois(loglike[i])
 }
   C <- 10.
   sigma.intr <- 3
   prec.intr <- pow(sigma.intr,-2)
   xcent ~ dunif(-10,10)
}.
```

The first `for` loop prepares the grid `grid.x[k]` needed for the integral evaluation, and computes the data selection function. The second loop prepares the dummy values for the zero trick (as in the previous section). Then, we evaluate the integral of the likelihood times the data selection function over the whole real axis (approximated as $[-10,10]$). The likelihood times data selection function is computed in JAGS implementing the former in the symbolic way (`x[i]` $\sim$ `dnorm(xcent, prec.intr)`) and the second one using the zero trick. Since the logarithm of zero is undefined, we inelegantly add a very small number to zero. We challenge the skilled reader to find the elegant coding that does not require such a trick.

As in the previous section, a few lines of code has become so complex that it requires one line of comment (starting with an `#`) for every line of code.

7.7.3 Probabilistic Selection on the True Value

In this section, we would like to implement working code for the model illustrated in Sect. 7.5:

```
symbolic-model {
for (i in 1:length(obslnL)) {
 obslnL[i] ~ dnorm(lnL[i],pow(err.lnL[i],-2)
 lnL[i] ~ dnorm(lnLhat,pow(sigma.intr,-2)) phi((x-3)/2)
}
sigma.intr <- 3
lnLhat ~ dunif(-10,10)
}.
```

The product of a normal and a ϕ function is not among the JAGS distributions and it is not allowed in JAGS to write a line with a tilde and a product of a distribution and a function. The brute force approach is therefore compelling.

The current model is almost identical to the previous one, except for the fact that we replaced the previous sharp selection with the soft one. The latter replacement enters in three places: when we compute the soft selection $p(lnL)$ (the `correct[k]` line), in the integral computation, and in the numerator of the posterior. For our comfort, we also changed the name of the variables. The model reads[3]:

```
data
{
# grid for integral evaluation
for (k in 1:400){
 grid.x[k] <- (k-200.)/20.
 correcc[k] <- phi((grid.x[k]-3)/2)
}
step.grid.x <-grid.x[2]-grid.x[1]
# dummy variable for zero-trick
# to sample from a distribution not available in JAGS
for (i in 1:length(obslnL)) {
 dummy[i] <-0
}
}
model {
# compute the integral at denominator
for (k in 1:length(grid.x)) {
normaliz[k] <- sqrt(prec.intr/6.28)*exp(-prec.intr*
                 pow(lnLhat-grid.x[k],2 )/2.)*correcc[k]
}
tot.norm <- sum(normaliz[])*step.grid.x
for (i in 1:length(obslnL)) {
 obslnL[i] ~ dnorm(lnL[i],prec.lnL[i])
 # the following would suffice, if no selection
 lnL[i] ~ dnorm(lnLhat,prec.intr)
 # but selection is there. Adding the additional likelihood term
 numerator[i] <- phi((lnL[i]-3)/2)
```

[3] A special thanks goes to Merrilee Hurn, who helped us to write this model.

```
 # sampling from an unavailable distribution
 loglike[i] <- -log(numerator[i]/tot.norm)+C
 dummy[i] ~ dpois(loglike[i])
}
C <- 10.
sigma.intr <- 3
prec.intr <- pow(sigma.intr,-2)
lnLhat ~ dunif(-10,10)
}.
```

7.7.4 Sharp Selection on the Observed Value, Mixture of Gaussians

The symbolic model (not to be run with JAGS) we now want to implement is:

```
symbolic-model {
for (i in 1:length(obscol)) {
obscol[i] ~ dnorm(col[i], prec[i]) T(1.5,4)
col[i] ~ dnorm(lambda[type[i]], prec.intr[type[i]], p)
type[i]~dcat(p[])
}
# Prior
lambda[1] <- 2
lambda[2] <- 1
p[1] <- 0.2
p[2] <- 1 - p[1]
prec.intr[1] ~ dunif(0.1,10)
prec.intr[2] <- 0.25
}.
```

To infer the width of the Gaussian signal, we just need to adopt previous models to the new setting: here we have a sum of Gaussians instead of just one single Gaussian, and therefore both the `numerator` and the normalization should be evaluated for the mixture of Gaussians. To evaluate the integral inside the range we just need to limit the grid on which the integral has to be evaluated, $[1.5,4]$. The model reads:

```
data {
# grid for integral evaluation, from 1.5 to 4
 for (k in 1:125){
  grid.x[k] <- (k+74.5)/50.
 }
 step.grid.x <-grid.x[2]-grid.x[1]
 for (i in 1:length(obscol)) {
   ### fake data for zeros trick
   zero[i] <- 0
 }
}
```

```
model {
for (i in 1:length(obscol)) {
 ## Observations
 l1[i] <- p[1]*sqrt(prec.tot[i,1]/6.28)*exp(-prec.tot[i,1]*
          pow(obscol[i]-lambda[1],2)/2.)
 l2[i] <- p[2]*sqrt(prec.tot[i,2]/6.28)*exp(-prec.tot[i,2]*
          pow(obscol[i]-lambda[2],2)/2.)
 ## integration
  for (k in 1:length(grid.x)) {
  n1[i,k] <- sqrt(prec.tot[i,1]/6.28)*exp(-prec.tot[i,1]*
             pow(lambda[1]-grid.x[k],2)/2.)
  n2[i,k] <- sqrt(prec.tot[i,2]/6.28)*exp(-prec.tot[i,2]*
             pow(lambda[2]-grid.x[k],2)/2.)
  }
 tot.norm[i] <- sum(p[1]*n1[i,]+p[2]*n2[i,])*step.grid.x

   for (j in 1:2) {
   sigma[i,j] <- sqrt(1/prec.intr[j] + 1/prec[i])
   prec.tot[i,j] <- pow(sigma[i,j],-2)
   }
 zero[i] ~ dpois(loglike[i])
 loglike[i] <- -log((l1[i]+l2[i])/tot.norm[i]) + C
}
C <- 10 ## Offset to ensure positivity of loglike[i]
## Prior
lambda[1] <- 2
lambda[2] <- 1
p[1] <- 0.2
p[2] <- 1 - p[1]
prec.intr[1] ~ dunif(0.1,10)
prec.intr[2] <- 0.25
}.
```

Note that the integration should be performed per each datum at each iteration of the chain, resulting in a slow code. The model can be written in a more compact and efficient form (e.g., Gaussian integrals are known by JAGS), however, when written in the current form the model is ready for implementing a soft (probabilistic) data selection.

7.8 Final Remarks

From a conceptual point of view, accounting for a non-random data collection is simple: one needs, as always, to write the prior and the likelihood of the model being fitted. However, the coding in JAGS of models accounting for selection effects may be somewhat involved when the likelihood is not among the distributions offered by JAGS. Therefore, we strongly invite readers to be extremely cautious and attentive to numerical-related issues. A non-exhaustive list of points to pay attention includes the integration range (large enough?), the step (small enough?), the constant *C* (large enough to have a positive rate, but not too large to exceed the computer's numerical precision?), and double checking that the written likelihood is the likelihood of the model being fitted (i.e., that no mistakes have been done in writing it).

A number of deterministic or probabilistic cases have been presented, both in the case of single distributions, or mixtures, and in presence or absence of errors. They are quite general and easy to adapt to numerous applications, for example the model of Sect. 7.6 is relevant to many applications in astronomy, such as in computing the width of a contaminated sample whose data have been truncated (e.g., velocity dispersion of a cluster of galaxies or the luminosity function of stars in globular clusters). One more model, dealing with a regression problem affected by non-random data collection, is described in Sect. 8.8.

Reference

R. Wojtak, S. H. Hansen, and J. Hjorth. Gravitational redshift of galaxies in clusters as predicted by general relativity. *Nature*, 477:567–569, 2011.

Chapter 8
Fitting Regression Models

In this chapter we introduce regression models, i.e., how to fit (regress) one, or more quantities, against each other through some sort of functional relationship. For example, fitting the amount of a radioactive isotope remaining in an item with time or the relation between interesting object features, such as velocity dispersion, luminosity, or temperature. Before we address this topic, however, we need to clear up some common misconceptions in astronomy and physics. Then, we show some examples of regression fitting, starting with examples where data collection has no influence on the conclusions. Next, we provide some rather common examples for which one needs to account for the data-collection scheme, e.g., the data structure introduces a bias in the fitted parameters (recall Chap. 7). We continue this chapter by illustrating how to deal with upper and lower limits in regression problems and how to predict the value of an unavailable quantity by exploiting the existence of a trend with another available quantity. We conclude this chapter by presenting some complex regressions problems.

8.1 Clearing Up Some Misconceptions[1]

8.1.1 Pay Attention to Selection Effects

Very often, easier-to-acquire data/objects are over-represented in samples (and difficult-to-acquire ones are under-represented). We already devoted a whole chapter to the non-random data collection (see Chap. 7). Unfortunately for the researcher, this non-random data collection often has an important effect on the determination of the parameters describing the trend between the quantities of interest.

Consider the data in Fig. 8.1, a case pointed out by Sandage (1972). There appears to be a trend, where the ordinate y (radio power, $logL_R$) increases with increas-

[1] Part of the material of this section has been drawn from Andreon and Hurn (2013).

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_8

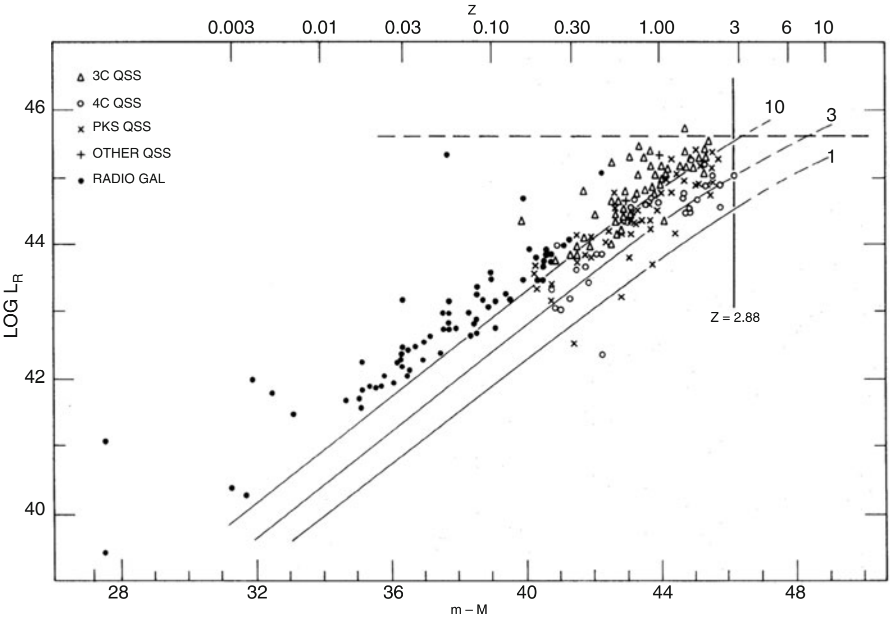

Fig. 8.1 An example of a non-random sample (from Sandage 1972). Radio power (basically, intrinsic luminosity) of radio sources as a function of their distance (distance modulus on the lower axis, redshift on the top axis) for various astronomical radio sources coming from various surveys. The tight relation is largely due to the non-random selection of the sample: points in the bottom-right corner are too faint to be included in the sample. *Astronomical interest:* This plot, named the Hubble diagram, allows one to measure the evolution of the studied population (radio sources in this plot), and if the latter is known, a cosmological parameter, see also Fig. 8.37. Reproduced from Sandage (1972) with permission

ing abscissa z (distance). The trend is not real, in the sense that this is a selection effect: objects with low y (radio power) and large z (distance) do not enter in the sample because a too smal volume has been explored. At low z (distance), the (Universe) volume is small (it goes with z^3 at the first order at low z), and rare objects (those with high y) are missing just because a too small volume has been explored. Therefore, a correlation between quantities is there, but this is not a measure of the properties of the objects under study, but of the way the sample has been collected. The latter has to do (also) with the observer, it is not a property of the objects under study. We emphasize that the problem of the selection cannot be easily fixed because we usually know nothing about the missed objects (because they are not observed), their abscissa, ordinate, and also how many are missed.

This is the standard case of X-ray astronomical observations which, because of their low depth, have faint objects under-represented. An example is illustrated in Fig. 8.2. At all temperatures T, clusters brighter-than-average are easier to detect (simply because they are brighter) and therefore they are over-represented in most samples, while those fainter-than-average are under represented. Therefore, at a given T the mean X-ray luminosity L_X of clusters in the sample is higher than the

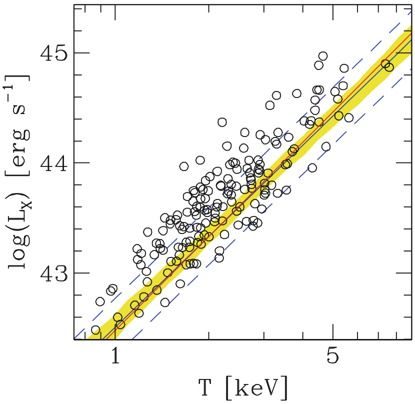

Fig. 8.2 A simulation (from Andreon and Hurn 2013) built around a true $L_X - T$ data set and selection function (from Pacaud et al. 2007), illustrating how the observed data may be unrepresentative compared to the underlying scaling relation due to selection effects. The *solid blue line* shows the true relationship but, due to the selection effects, most points lie above the line. Reproduced from Andreon and Hurn (2013) with permission

average L_X of the population at the same T. The selection is stochastic: it is possible to miss an object above the threshold and yet still record some objects below the threshold. The selection is profound, we ignore what we miss (both abscissa and ordinate) and how many objects are missed. While this effect is self-evident, its consequences are not always fully appreciated (Andreon and Moretti 2011).

Therefore, as soon as you suspect the existence of a trend in your data, please check if the possible trend may be due to some selection effects. Ask yourself *why* a datum is in your figure and whether it would be in the figure also if it were in a completely different part of the plot. If the probability that an object is in the sample depends on where it falls in the figure, you probably cannot ignore the effects of the data collection. If you are using an uncontrolled data set ("what is available"), remember that:

- difficult-to-acquire measurements are under-represented in plots (generally speaking).
- enlarging the sample size does not remove (usually) the selection effect when the newly introduced data shares the same selection bias with the available data.

8.1.2 Avoid Fishing Expeditions

Nowadays, there are extensive catalogs with many entries and many features per entry, for many categories of objects. It is hard to resist the temptation of looking for the existence of trends among object features. However, such a fishing expedition is a very risky one.

Let us suppose that each object has 10 features. There are 45 different pairs of variables to test for trends. Again suppose, for a moment (and nowhere else in this book), that we are using some non-Bayesian measurements of significance about the existence of a trend. If the catalog were filled with random numbers, then, on average, we expect to find two trends significant at 95 %, because "95 % significant" means that one in 20 is found by chance if no trend is there. Because we tested 45 pairs of variables, we should almost always find (at least) one pair of variables with a trend of 95 % significance. This unwanted feature occurs even if the sample is formed by a large sample, say, three billion entries (objects). A large sample does not guarantee that a trend is true. This unwanted feature of fishing expeditions occurs even if you, having found a trend at the first pair of object features, stop and do not consider any other of the 44 pairs of features.

Therefore, avoid fishing expeditions and do not hope that a large sample size will protect you from finding a trend that does not exist in reality.

8.1.3 Do Not Confuse Prediction with Parameter Estimation

There are three possible reasons why two quantities are regressed on each other:

- One of the two variables, y, is difficult/costly/impossible to measure. The other variable, x, is cheaper to measure and informative about the quantity that is difficult to measure, because there is a trend between x and y with small scatter. Therefore, one may measure x to estimate y. This task is called prediction (and sometimes calibration).
- We measured both x and y. The latter is a function of the former, and we want to measure the parameters describing the relation between the two variables. This task is named parameter estimation.
- Finally, by the available x and y we may want to infer if the relation between them is given by one mathematical form, or another: e.g., if $y \propto x^2$ or $y \propto sin(x)$. This task is named model selection.

These are three different tasks that should not be confused. In general, three different procedures have to be followed to reach the target purpose, one single procedure is not enough. One should not use the fit performed for parameter estimation to predict values (and vice versa), because these differ, as we now show.

8.1.3.1 Prediction and Parameter Estimation Differ!

Let us consider the case where the prediction of a variable y is linearly related to a predictor variable x which is itself a random variable. This is the situation in which researchers are often confronted when we want to predict a quantity as a function of another and for both quantities we must collect observational data.

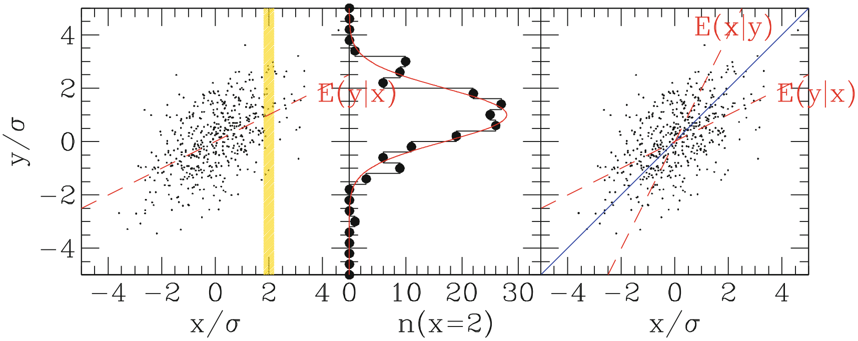

Fig. 8.3 *Left panel*: 500 points drawn from a bivariate Gaussian, overlaid by the line showing the expected value of y given x. The *yellow vertical stripe* captures those y for which x is close to 2. *Central panel*: Distribution of the y values for x values in a narrow band of x centered on 2, as shaded in the left panel. *Right panel*: as the left panel, but we also add the lines joining the expected x values at a given y, and the $x = y$ line. Reproduced from Andreon and Hurn (2013) with permission

For definitiveness, we will consider galaxy cluster mass (the holy grail of astronomy) as a function of the number of its galaxies (richness). Of course, the reader may choose their own favorite relation.

Figure 8.3 shows a set of 500 points drawn from a bivariate Gaussian where marginally both x and y are standard Gaussians with mean 0 and variance 1 and x and y have correlation $1/2$. The blue solid line in the right panel has slope 1 and intercept zero. This line seems, by eye, to be an acceptable description of the plotted data: it does not look odd for any reason, it captures the trend pointed out by the data.

Superimposed on the left-hand panel of Fig. 8.3 is the line giving the theoretical conditional expectation of y given x (this is known theoretically for this bivariate Gaussian to be $y = 0.5x$). We pretend to ignore the above result and re-derive it ourselves. Consider a small range in x close to 2, shaded in the figure. It is clear from the histogram of the points falling in the shaded area that their average is closer to the value predicted by the red line (1 in this case) than the value predicted by the blue, $y = x$, line (2 in this case). By repeating the calculation for other small ranges of x values, and connecting the found $(x, E(y))$ pairs, we obtain the red dashed line. This is the line that allows one to best predict y given x. To further illustrate the point, Fig. 8.4 shows the residuals from this line (with slope $1/2$ (i.e., $y - 1/2x$) and from the $y = x$ line (i.e., $y - x$). There is no doubt that the former is best for predicting values, while the latter is best to describe the trend between x and y. Therefore, different lines should be used for different purposes (prediction vs. parameter estimation) and one should therefore refrain from predicting y using the blue line.

Notice that the red line appears too shallow with respect to the trend identified by the points, which perhaps might be captured by the $x = y$ line shown in blue in the right-hand panel. However, if what we want to do is to predict a y given an x value, this "too shallow" line is more appropriate.

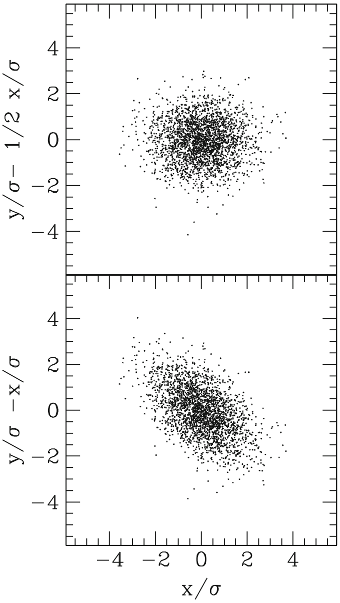

Fig. 8.4 Residuals from the $y = 1/2x$ (*top*) or $y = x$ relation

The data shown in Fig. 8.3 are generated by the model:

```
model {
val[1:2] ~ dmnorm(mu[],prec[,])
} ,
```

where the zero mean mu and the inverse covariance matrix prec (with variance 1 and correlation 0.5) are given by

```
prec <-structure(c(1.33333,-0.666667,-0.666667,1.33333),
                 .Dim = c (2,2))
mu <- c(0,0)
```

Note that JAGS parameterizes the multivariate normal in terms of the precision matrix. The precision matrix is defined as the inverse of the covariance matrix. The covariance matrix for the above example is

$$\texttt{covar} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

Then using common linear algebra methods, finding the inverse of this matrix yields the precision matrix, which is given in the JAGS code above,

$$\texttt{prec} = \texttt{covar}^{-1}.$$

Table 8.1 Data for efficiency determination

E[i]	nrec[i]	ninj[i]	E[i]	nrec[i]	ninj[i]	E[i]	nrec[i]	ninj[i]
2	23	96	38	221	235	74	41	41
6	71	239	42	210	217	78	32	32
10	115	295	46	182	185	82	32	32
14	159	327	50	136	140	86	31	31
18	200	345	54	119	121	90	22	22
22	221	316	58	79	79	94	18	18
26	291	349	62	81	81	98	11	11
30	244	281	66	61	61			
34	277	298	70	44	45			

8.1.3.2 Direct and Inverse Relations also Differ

Now we want to show that the line allowing one to predict y from x differs from the line giving predictions of x from y. Similarly to the previous section, we select a small range in y, take the mean of the corresponding x values, and connect the pairs $(y, E(x|y))$ (or one may consult a statistical book written by an analytically skilled author to find that this line has slope 2). By this procedure, we find the steep line shown in the right panel of the Fig. 8.3. This new line differs, $y = 2x$, from both the $E(y|x)$ line used for prediction earlier ($y = 1/2x$) and from the $y = x$ line! This new line is the most appropriate line for predicting x from its y value.

To summarize, the lines $E(x|y)$ and $E(y|x)$ differ, and these two differ from the $y = x$ line. If

$$y = ax + b, \tag{8.1}$$

then in general

$$x \neq y/a - b/a \tag{8.2}$$

because the slope of the inverse relation (the one allowing prediction of x given y) differs from the slope of the direct relation (the one allowing prediction of y given x).

Unfortunately, these differences are often overlooked. Therefore, gentle reader, before proceeding in your reading, we ask you to repeat our calculation, because it is really important not to overlook these differences. If you have some past experience in fitting data, you should now know one of the reasons why fitting procedures return different slopes for a single data set.

8.1.3.3 Summary

Do not proceed to fit a regression model before

1. checking for the presence of selection effects;
2. ignoring what you are looking for (i.e., do not embark on in fishing expeditions);
3. understanding what you want: y vs. x, x vs. y, predict y given x or predict x given y.

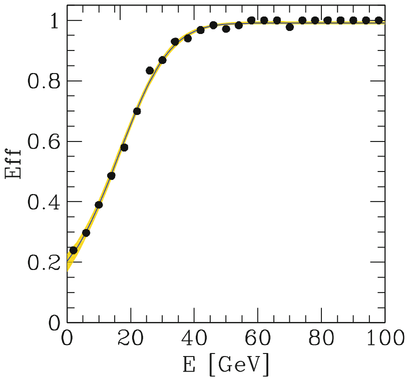

Fig. 8.5 Efficiency of the considered (CERN) experiment, as a function of energy. The *shading* indicates the 68 % probability interval

8.2 Non-linear Fit with No Error on Predictor and No Spread: Efficiency and Completeness[2]

Now that those common misconceptions are cleared up, we can now start fitting quantities against each other. Let us start from the simplest possible case: the predictor quantity is perfectly known, shows a tight relation with another quantity, and we want to learn about the parameters of the fit describing the trend between predictor and target. By saying that the relation is tight we mean that we have no reason to suspect that the data scatter around the fitted trend by more than their errors.

One of the standard works in astronomy and physics is estimating the efficiency of a selection or the completeness of a catalog. For example, some physics experiments are producing an extremely high event rate so that not all of the events can be recorded, which is quite typical of many of the experiments held at CERN. In such cases, events considered interesting through some quick, approximate, event reconstruction are only recorded. For example, the reconstructed event may have a missing energy of > 20 GeV, suggestive of a possible semi-leptonic W decay at LHC (a hot topic at the time of the writing of this book), and in such a case this (based on quick and dirty computations) missing energy triggers the recording of the event for a later (off-line) more detailed analysis. The trigger efficiency is useful for checking that the selection is correctly working (e.g., not discarding too many interesting events, and not including too many non-interesting events), to maximize the quality of the measurand quantity, and to infer the true rate of interesting events from the recorded ones. Similarly, in astronomy, star, galaxy or other populations are not 100 % selected (at all luminosities, masses, etc.), and the efficiency of the selection needs to be determined.

For definitiveness, we consider here a physics experiment, but our analysis can be applied to astronomical cases with no changes at all. For clarity, we suppose that

[2] This example draws material from (the end-course assessment relation of) Andrea Benaglia.

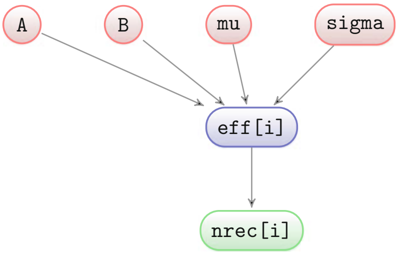

Fig. 8.6 Graph showing the stochastic relationship between quantities involved in fitting the efficiency of a CERN experiment

efficiency, `eff[i]`, depends on energy only, and we recorded, at the energy `E[i]`, `nrec[i]` events out of `ninj[i]` injected:

```
nrec[i] ~ dbin(eff[i],ninj[i]) .
```

The data are listed in Table 8.1. Inspection of the experimental data, shown in Fig. 8.5, suggests that the efficiency in our (not completely) hypothetical experiment of the semi-leptonic W decay at LHC can be well described by an error function (which is a `phi` function in the statistical literature and in JAGS, after a simple change of variable):

```
eff[i] <- A + (B-A)*phi((E[i]-mu)/sigma).
```

We adopt uniform priors for the parameters over a wide range certainly including the true value, but not so large to include unphysical values:

```
A~dunif(0,1)
B~dunif(0,1)
mu~dunif(0,100)
sigma~dunif(0,100).
```

The complete model, illustrated in Fig. 8.6, then reads:

```
model {
for (i in 1:length(nrec)) {
 nrec[i] ~ dbin(eff[i],ninj[i])
 nrec.rep[i] ~ dbin(eff[i],ninj[i])
 eff[i] <- A + (B-A)*phi((E[i]-mu)/sigma)
}
A~dunif(0,1)
B~dunif(0,1)
mu~dunif(0,100)
sigma~dunif(0,100)
}.
```

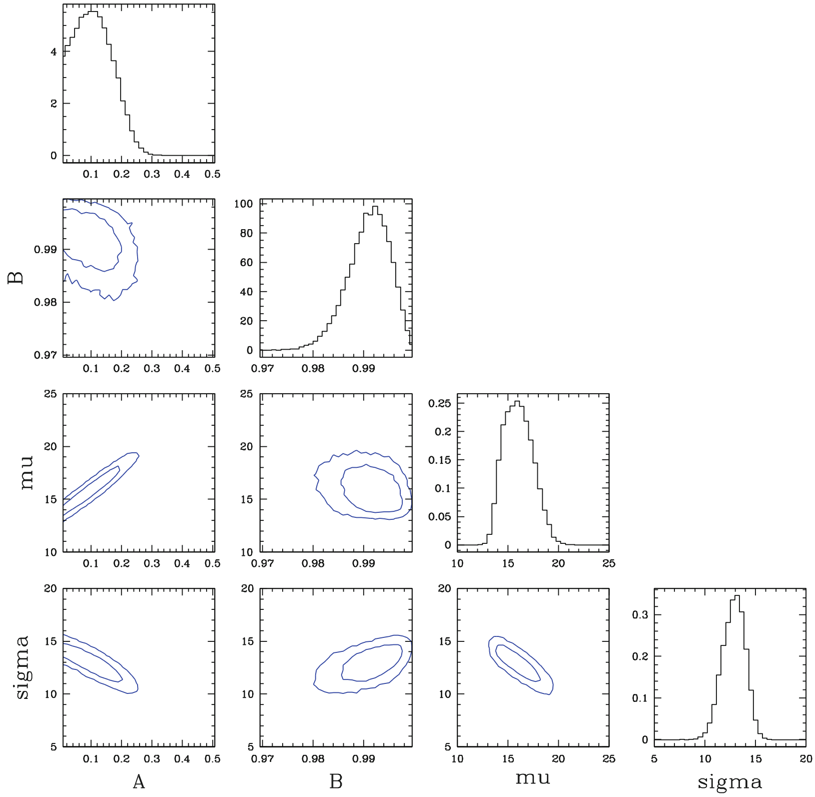

Fig. 8.7 Probability distributions of single parameters (histograms, top panel of each column) and of joint parameters (68 % and 95 % probability contours, remaining panels) for the efficiency estimation example. The jagged nature of some contours is due to numerical reasons (the finite length of the plotted chain)

Fitting this model on the data gives:

$$\texttt{eff}(E) = 0.11 \pm 0.06 + (0.88 \pm 0.06)\phi \left(\frac{\texttt{E} - 16.0 \pm 1.4}{12.8 \pm 1.1} \right) \tag{8.3}$$

which is plotted in Fig. 8.5. The fitted model smoothes irregularities due to (binomial) fluctuations and allows us to interpolate at energies between bin centers. The analysis does not require that we have data at all energies, we may well have some missing bins with no change at all to be applied to the code. Parameters show strong covariance, illustrated in Fig. 8.7: different parameter sets describe equally well the data.

The code above can be easily extended to the case of efficiency or incompleteness which depends on two quantities, as is usually the case of (astronomical) spectroscopic completeness: the success rate usually decreases for fainter galaxies, but also for red ones.

Astronomy is full of examples for which this fitting model, or a simple modification of it (changing the expression of `eff[i]`) can be applied, such as deriving the parameters describing the completeness of a catalog as a function of magnitude, color or redshift, or for deriving the trend with redshift (or any other precisely measured quantity) of the fraction of red galaxies, of early-type galaxies, of stars or AGN of a given type, etc.

8.3 Fit with Spread and No Errors on Predictor: Varying Physical Constants?

Very often when there is a trend between two quantities, but the relation is not tight, the data scatter around the trend by more than allowed by their error. The purpose of this section is to illustrate how to account for the stochasticity between predictor and target quantity in a fit.

An interesting question to address is whether constants in physics are really constants or are, instead, variable, in the sense that they may take a different value somewhere else in the Universe or at a different time. In order to address this possibility, Wendt and Molaro (2011) observed an object, QSO 0347-383, in the very distant universe (at redshift about 3) when the Universe was 11 Gyr younger than today (i.e., when it was 90 % younger). They measured the ratio between the wavelength of several spectral lines and compared them with the wavelength measured in the laboratory: $(\lambda_{obs} - \lambda_{lab})/\lambda_{lab} = \Delta\lambda/\lambda$. They exploit the fact that the difference between energy levels, and thus wavelength of the emitted photon, does depend on physical constants, such as the ratio of the proton-to-electron masses, μ, but with different coefficients of proportionality: some transitions have energy differences independent of proton-to-electron masses, some other transitions, however, instead do depend. This is so because the energy difference between two consecutive levels of the rotational spectrum of a diatomic molecule scales with the mass ratio μ, whereas the energy difference between two adjacent levels of the vibrational spectrum is proportional to $\mu^{1/2}$. If physical constants take values different in QSO 0347-383 than they have today here on Earth, then $\Delta\lambda/\lambda$ should be shifted in wavelength by an amount that is proportional to the sensitivity `k` of the transitions producing the wavelength lines. For a proton-to-electron mass ratio that always takes the same value, $\Delta\lambda/\lambda$ is independent of `k` and it is equal to $1+z$, by definition of z (redshift). In essence, measuring a variation of physical constants is a measurement of the slope of $\Delta\lambda/\lambda$ with `k`: if the slope is equal to zero, the proton-to-electron mass ratio has not changed between us and QSO 0347-383, whereas if `k`$\neq 0$ it has changed. Figure 8.8 displays recent data for QSO 0347-383 from Wendt and Molaro (2011): the abscissa reports `k`, the sensitivity of spectral features to the value of physical constants (larger `k` means increased sensitivity). The ordinate plots `y`$= 1000(\Delta\lambda/\lambda - z - 1)$, with $z = 3.0248$. These two coefficients have been introduced to make numbers easier to read: we first remove five identical digits present in all values, and we then multiply the resulting number by 1000, to remove unnecessary zeroes in front of the interesting part of the numbers.

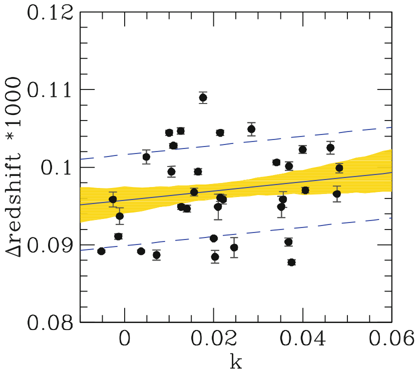

Fig. 8.8 Scaling between $\Delta\lambda/\lambda$ and k, observed data, the mean scaling (*solid line*) and its 68 % uncertainty (*shaded yellow region*) and the mean intrinsic scatter (*dashed lines*) around the mean relation. The distance between the data and the regression line is due in part to the observational error on the data and in part to the intrinsic scatter

As mentioned, we expect a linear relation between $\Delta\lambda/\lambda$ and k and therefore we adopt a linear regression model:

```
z[i] <- alpha+0.1+beta*(k[i]-0.03) ,
```

where we center quantities near their average (0.03 for the abscissa, 0.1 for the ordinate) to numerically help with the MCMC, in particular this reduces the covariance between parameter estimates and simplifies the interpretation of the found results.

Even a casual inspection of the figure shows that data scatter (vertically) by more than their errors. Because all measurements pertain to a single object, the observed spread is not due to a population effects, i.e., to intrinsic difference between objects, but to some sort of systematic error not yet identified. We therefore allow for an intrinsic scatter `intrscat` taken to be Gaussian.

```
y[i] ~ dnorm(z[i],pow(intrscat, -2)) .
```

Errors, `err.y[i]`, are also Gaussian:

```
obsy[i] ~ dnorm(y[i],pow(err.y[i], -2)).
```

We assume a uniform prior for the intrinsic scatter, for the intercept (formally, a Gaussian of large variance), and a uniform prior on the angle *b*, i.e., a Student-t distribution on the angular coefficient `beta`. The latter choice is motivated by the fact that the slope of the regression should not depend on the convention used (by humans) to measure angles (e.g., as in astronomy from *y* clockwise or as in mathematics, from *x* counterclockwise).

```
intrscat ~ dunif(0,3)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1) .
```

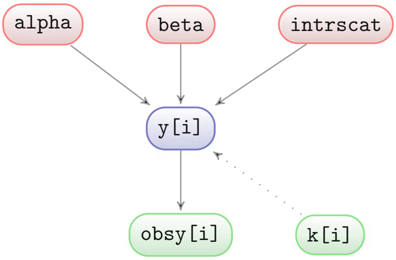

Fig. 8.9 Graph showing the stochastic relationship between quantities involved in estimating whether physical constants are indeed constant. *Solid arrows* indicate stochastic relations, *dotted arrows* represent deterministic relations

Table 8.2 Data for the possibly varying physical constant fit, from Wendt and Molaro (2011)

k[i]	obsy[i]	err.y[i]	k[i]	obsy[i]	err.y[i]	k[i]	obsy[i]	err.y[i]
0.0462	0.10250	0.0008233	0.0352	0.09489	0.001373	0.0165	0.09942	0.00037
0.0215	0.10440	0.0003033	0.0037	0.08917	0.00018	0.0156	0.09680	0.0005033
0.0210	0.09489	0.001663	0.0375	0.08774	0.0003267	0.0126	0.10470	0.00042
0.0482	0.09990	0.0006033	0.0369	0.10010	0.0005567	0.0105	0.09942	0.0007033
0.0477	0.09656	0.00102	0.0341	0.10060	0.0002833	0.0110	0.10280	0.0002367
0.0140	0.09465	0.0004267	0.0285	0.10490	0.00085	0.0100	0.10440	0.0003233
0.0127	0.09489	0.00033	−0.0052	0.08917	0.0001567	0.0072	0.08869	0.00066
0.0368	0.09036	0.0004967	0.0246	0.08965	0.001263	0.0049	0.10130	0.00092
0.0109	0.10280	0.0003233	0.0221	0.09584	0.00056	−0.0011	0.09370	0.00106
0.0406	0.09704	0.0003667	0.0203	0.08845	0.0008167	−0.0014	0.09108	0.0003333
0.0400	0.10230	0.0005133	0.0215	0.09608	0.0004633	−0.0026	0.09584	0.0009433
0.0356	0.09584	0.001007	0.0200	0.09084	0.0001133	0.0176	0.10900	0.0007267

The model, whose graph is shown in Fig. 8.9 reads:

```
model {
intrscat ~ dunif(0,3)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
for (i in 1:length(x)) {
 # modeling ordinate
 obsy[i] ~ dnorm(y[i],pow(err.y[i], -2))
 y[i] ~ dnorm(z[i],pow(intrscat, -2))
 # modeling ordinate vs x
 z[i] <- alpha+0.1+beta*(k[i]-0.03)
 }
}.
```

Figure 8.8 shows the scaling between y and k, the observed data, the mean scaling (solid line) and its 68 % uncertainty (shaded yellow region) and the mean intrinsic scatter (dashed lines) around the mean relation. The data are listed in Table 8.2.

Figure 8.10 shows the posterior distribution of key parameters: slope, intercept, and intrinsic scatter. The intrinsic scatter is large, comparable to the variation of the mean model between the minimal and maximal k. This large scatter makes the slope determination rather difficult, because the expected variation due to the k dependency is noised by the intrinsic scatter. In formulae (and after converting numbers back to the original scale):

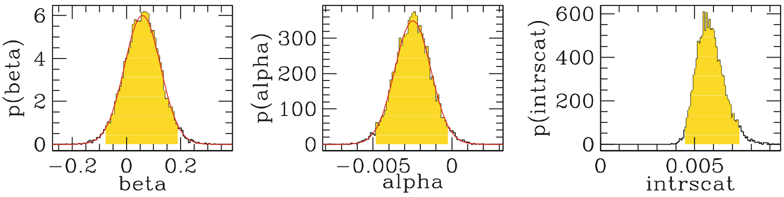

Fig. 8.10 Posterior distribution for the key parameters of the regression model used for testing whether physical constant are indeed constant

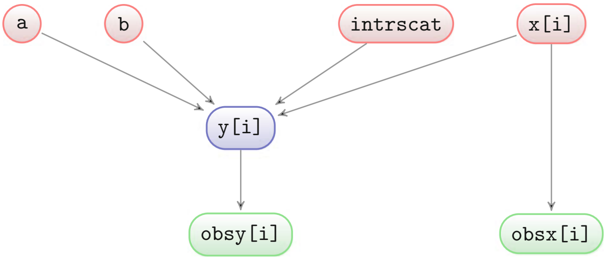

Fig. 8.11 Graph showing the stochastic relationship between quantities involved in estimating a relation between two quantities, each subject to errors, and in presence of an intrinsic scatter (e.g., the Magorrian relation)

$$\Delta\lambda/\lambda = 3.0248 + (0.98 \pm 0.01) \times 10^{-4} + (0.6 \pm 0.7) \times 10^{-4}(k - 0.03) \qquad (8.4)$$

with an intrinsic scatter of $5.8 \pm 0.7 \times 10^{-6}$. We quote for the regression coefficients posterior mean ± the standard deviation.

Back to the original question, whether there is any evidence for varying proton-to-electron mass (i.e., $k \neq 0$), we note that while the maximum a posteriori (and mean) slope is non-zero, the slope value of zero is almost as probable (30 % less only) than the most probable slope (left panel of Fig. 8.10). Furthermore, the model with the variable beta has one more parameter (yielding an extra degree of flexibility), that can be adjusted to increase the quality of the fit. Its better performance (larger value of the maximal posterior) is therefore partially due to the larger flexibility of the model.

8.4 Fit with Errors and Spread: The Magorrian Relation

Very often one wants to fit a trend between two quantities but *both* of these quantities are subject to errors, and the scatter around the trend is larger than errors allow, i.e., there is an intrinsic scatter. To illustrate this regression model, let us consider the problem (and data set) in Tremaine et al. (2002).

Tremaine et al. (2002) measured the trend between the log of the galaxy's velocity dispersion x and the log mass y of the black hole sitting in their center. Errors on these quantities are assumed to be Gaussian and are different from point to point, errx[i] and erry[i]:

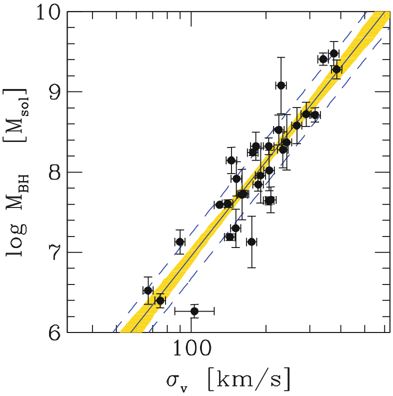

Fig. 8.12 Magorrian relation. The *solid line* marks the mean relation, its 68 % uncertainty is *shaded* (*in yellow*). The *dashed lines* show the mean relation plus or minus the intrinsic scatter `intrscat`. Error bars on the data points represent observed errors for both variables

```
obsx[i] ~ dnorm(x[i],pow(errx[i],-2))
obsy[i] ~ dnorm(y[i],pow(erry[i],-2)).
```

Figure 8.12 shows the data, which are listed in Table 8.3. Inspection of the plot shows that the (vertical) scatter is larger than the plotted errors, i.e., an intrinsic scatter is there. In this case the scatter may be a manifestation of an intrinsic variety of studied objects (i.e., a population effect). Therefore we need to adopt a regression model with a (an intrinsic) scatter, `intrscat`, taken to be Gaussian here for simplicity. Following the approach by the authors, the trend between black hole and galaxy velocity dispersion is taken to be linear on a log scale:

```
y[i] ~dnorm(b+a*(x[i]-2.3), prec.scat),
```

where we center the `x` quantity (log of velocity dispersion) near its average (200 km/s) because again, this is numerically convenient, reduces the covariance between estimated parameters, and simplifies the interpretation of the found results.

We assume uniform priors for the reciprocal of the square of the intrinsic scatter (formally, a gamma function with small parameters), for the intercept (formally, a Gaussian of large variance), and an uniform prior on the angle *b*, i.e., a Student-t distribution on the angular coefficient `beta`

```
intrscat <- 1/sqrt(prec.intrscat)
prec.intrscat ~ dgamma(1.0E-2,1.0E-2)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
}.
```

Finally, for `x` we adopt a uniform prior, between minus and plus infinity, approximated as:

```
x[i] ~ dunif(-1.0E+4,1.0E+4).
```

Table 8.3 Data for the Magorrian relation, from Tremaine et al. (2002)

x[i]	errx[i]	y[i]	erry[i]	x[i]	errx[i]	y[i]	erry[i]	x[i]	errx[i]	y[i]	erry[i]
2.01	0.08	6.27	0.08	2.16	0.02	8.15	0.16	2.57	0.02	9.48	0.15
1.88	0.02	6.40	0.09	2.31	0.02	8.02	0.25	2.21	0.02	7.73	0.04
2.20	0.02	7.72	0.31	2.16	0.02	7.19	0.04	2.18	0.02	7.92	0.21
2.32	0.02	7.65	0.16	2.26	0.02	8.32	0.17	2.59	0.02	9.28	0.12
2.31	0.02	7.64	0.05	2.11	0.02	7.59	0.01	2.25	0.02	8.24	0.04
2.18	0.02	7.30	0.24	2.50	0.02	8.71	0.09	1.95	0.02	7.13	0.15
2.24	0.02	7.13	0.32	2.38	0.02	8.37	0.34	2.37	0.02	8.28	0.22
2.15	0.02	7.61	0.05	2.35	0.02	8.53	0.19	2.46	0.02	8.72	0.15
2.36	0.02	9.08	0.35	2.27	0.02	7.85	0.08	2.42	0.02	8.58	0.22
2.31	0.02	8.32	0.11	2.28	0.02	7.96	0.34	1.83	0.02	6.53	0.17
2.53	0.02	9.41	0.08								

The model is similar to the one adopted for the possibly varying physical constant (compare the current graph, Fig. 8.11, with the graph of the previous model, Fig. 8.9): we only need to model here the error on the predictor (in the previous section the predictor was assumed to be perfectly known) and adopt a prior for the predictor.

In summary, our model reads:

```
model {
for (i in 1:length(obsx)) {
 x[i] ~ dunif(-1.0E+4,1.0E+4)
 obsx[i] ~ dnorm(x[i],pow(errx[i],-2))
 y[i] ~dnorm(b+a*(x[i]-2.3), prec.scat)
 obsy[i] ~ dnorm(y[i],pow(erry[i],-2))
}
prec.scat ~ dgamma(1.0E-2,1.0E-2)
intrscat <- 1/sqrt(prec.scat)
b ~ dnorm(0.0,1.0E-4)
a ~ dt(0,1,1)
}.
```

We found for the relation between the black hole mass M_{BH} ($\log_{10}$ scale, alias y) and galaxy velocity dispersion σ_v (also $\log_{10}$, which we called x):

$$\log M_{BH} = (3.9 \pm 0.3)\ (\log \sigma_v - 2.3) + 8.13 \pm 0.06, \qquad (8.5)$$

where, as usual, we quote for the regression coefficients the posterior mean ± one standard deviation.

Figure 8.12 shows the relation between the black hole mass M_{BH} and galaxy velocity dispersion, the observed data, the mean scaling (solid line) and its 68 % uncertainty (shaded yellow region), and the mean intrinsic scatter (dashed lines) around the mean relation. The ± 1 intrinsic scatter band is not expected to contain 68 % of the data points, because of the presence of measurement errors.

Figure 8.13 shows the posterior marginals for the model parameters: slope, intercept, and intrinsic scatter intrscat. The intrinsic scatter in black hole mass at a given galaxy velocity dispersion, intrscat=$\sigma_{lgM_{BH}|\sigma_v}$, is 0.28 ± 0.05 dex. This is the intrinsic scatter, i.e., the term left after accounting for measurement errors. It is clearly non-zero (see right panel of Fig. 8.13).

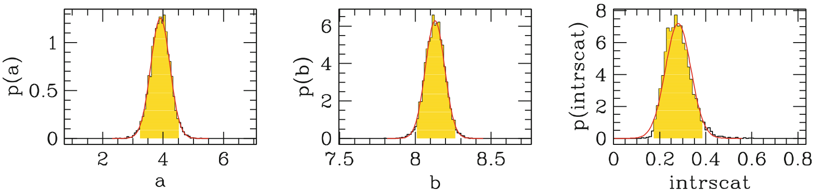

Fig. 8.13 Posterior probability distribution for the parameters of the Magorrian relation. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation to the posterior. The *shaded* (*yellow*) *range* shows the 95 % probability interval

8.5 Fit with More Than One Predictor and a Complex Link: Star Formation Quenching

Very often a quantity is affected not just by one variable, but by several. Therefore, a researcher may be interested in looking for trends not between pairs of quantities, but between n-tuples. We need to consider the joint effect of pairs (or triple or quadruple, etc.) of predictors on the target (interesting) quantity.

We illustrate this case using a classical galaxy evolution problem: the color of a galaxy (which measures the quenching rate) does depend on many factors: on (the log of) galaxy mass, `lgM`, on environment (the distance from the cluster center, usually expressed in units of a reference dimension, r_{200}) and on redshift, `z`, at the very least. Indeed, the galaxy color likely depends on a (an unknown) combination of these quantities, and, in order to disentangle one dependence from another, we need to fit a functional dependence on the three parameters simultaneously.

We use the data for 26 clusters of galaxies in Raichoor and Andreon (2012), with $0 < \mathtt{z} < 2.2$. The data consists of the total number of galaxies and the number of blue galaxies in the cluster's line of sight (`obsntot[i]` and `obsnbluetot[i]`, respectively), and in a reference line of sight (`obsnbkg[i]` and `obsnbluebkg[i]`, respectively). The latter has a `C[i]` times larger solid angle. Measurements are performed in three different cluster circular annuli, $r < r200/2$, $r200/2 < r < r200$ and $r200 < r < 2 * r200$, for galaxies in four different mass bins, `lgM`$= 11.47, 10.94, 10.54$, and $10.14 M_{\odot}$. The data table has 294 lines (each cluster has, on average, measurements at three or four different mass bins and at three different cluster-centric radii),[3] and its content is illustrated in Fig. 8.14.

The determination of the fraction of red galaxies in each cluster in the presence of a background has been already addressed in Sect. 6.1.3, and we summarize it here:

```
obsnbkg[i]~dpois(nbkg[i])
obsntot[i]~dpois(nbkg[i]/C[i]+nclus[i])
obsnbluebkg[i]~dbin(fbkg[i],obsnbkg[i])
```

[3] The table is given in electronic format at the link http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/Raichoor_Andreon12.dat.

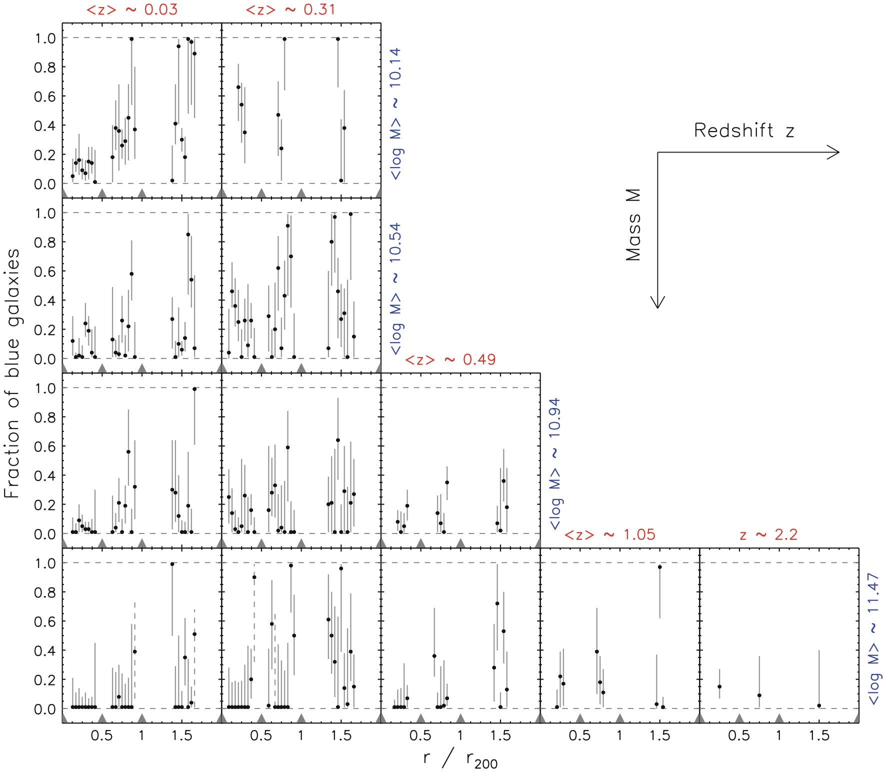

Fig. 8.14 Blue fraction for individual clusters as a function of cluster-centric distance (r/r200) for different bins of redshift (increasing rightward) and galaxy mass (increasing downward). Error bars represent the 68 % probability interval. Reproduced from Raichoor and Andreon (2012) with permission

```
obsnbluetot[i]~dbin(f[i],obsntot[i])
f[i] <- (fbkg[i]*nbkg[i]/C[i]+fclus[i]*nclus[i])/
        (nbkg[i]/C[i]+nclus[i]) .
```

The first two lines just state that counts fluctuate in a Poissonian way, the third and fourth lines state that the fractions have binomial errors, while the last line is just plain algebra.

We now need a model that allows us to describe how the fraction of red galaxies depends on redshift, mass, and environment. Following Raichoor and Andreon (2012) we “bend” the real axis on the [0,1] interval by using the `ilogit` function (which is the inverse of the logit function), which is defined by:

$$\texttt{ilogit}(y) = \frac{e^y}{1+e^y}. \tag{8.6}$$

For simplicity, we initially assume no cross-terms, i.e., no joint dependency on pairs of parameters among environment, mass and redshift (no terms proportional

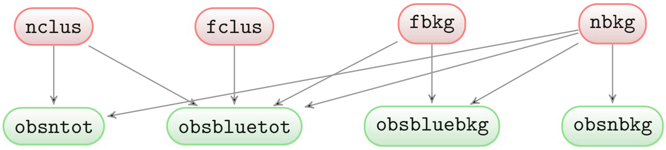

Fig. 8.15 Graph showing the stochastic relationship between quantities involved in estimating the parameters of the fit to the mass-environment-age star formation quenching. Deterministic links (to radius and redshift) are not drawn

to environment times mass, for example), and a linear relation between all terms, centering terms as usual at some value inside the studied range:

```
fclus[i] <- ilogit(lgfclus0+alpha*log(r200[i]/0.25)+
            beta*(lgM[i]-11)+gamma*(z[i]-0.3)) .
```

The coefficient `alpha` describes the amplitude of environmental dependency, the coefficient `beta` is the amplitude of the mass dependency, and `gamma` is the amplitude of the redshift dependency. To speed up computation and for simplicity in interpreting the found results, we center (i.e., subtract or normalize) quantities at a value where they are well measured (redshift: `z`$= 0.3$, mass: `lgM`$=10^{11}$ $M_{\odot}$, and cluster distance: $0.25r200$). The coefficient `lgfclus0` gives the blue fraction at these values.

This model does not fit the data (in particular the values in the top panels of Fig. 8.16 are not well fitted by this model), and the authors were forced to introduce a more complex model with a cross-term, making the rate of evolution dependent on mass (i.e., different for galaxies of different masses):

```
fclus[i] <- ilogit(lgfclus0+alpha*log(r200[i]/0.25)+
            beta*(lgM[i]-11)+gamma*(z[i]-0.3)+zeta*
            (lgM[i]-11)*(z[i]-0.3)) .
```

The parameter `zeta` measures how much faster/slower is the evolution of low mass galaxies compared to `lgM=11` galaxies.

Note that the increased complexity of the link function does not add "special" difficulties in the Bayesian approach: the latter is not restricted to fit linear relations between quantities, and indeed our first regression fitting, the determination of the efficiency (Sect. 8.2), used a non-linear link function.

We now set priors for parameters:

```
nbkg[i] ~ dunif(1,1e+7)
fbkg[i] ~ dbeta(1,1)
nclus[i]~ dunif(1,1e+7)
lgfclus0 ~dnorm(0,0.01)
alpha ~ dnorm(0,0.01)
beta ~ dnorm(0,0.01)
gamma ~ dnorm(0,0.01)
zeta ~ dnorm(0,0.01) .
```

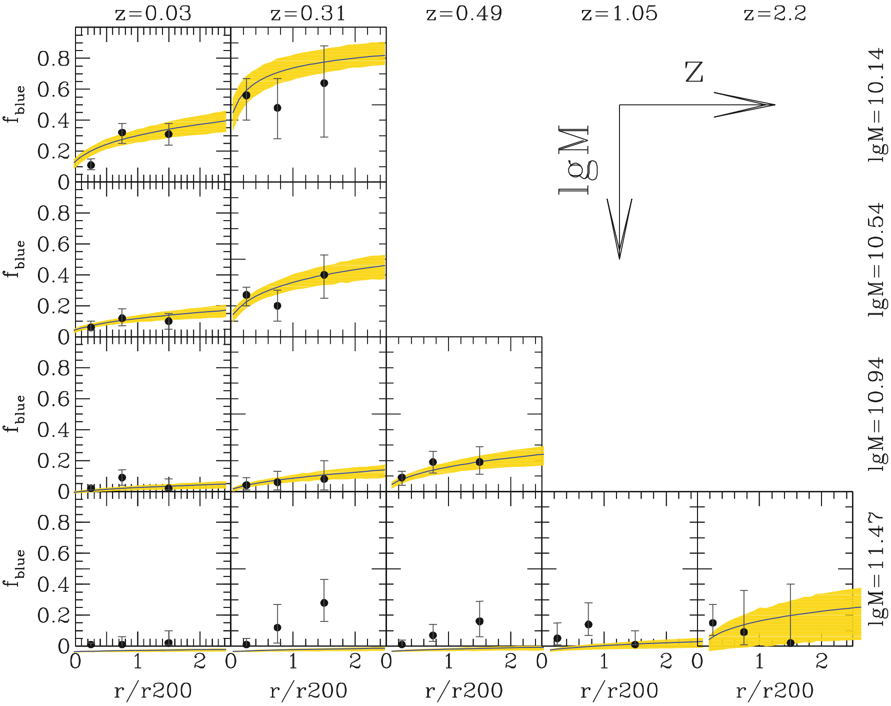

Fig. 8.16 Blue fraction for ensemble clusters as a function of cluster-centric distance (r/r200) for different bins of redshift (increasing rightward) and galaxy mass (increasing downward). Error bars represent the 68 % probability interval. The *solid line* and the *yellow shaded areas* represent the posterior mean and its 68 % probability interval

Therefore, the full model, illustrated in the Fig. 8.15, reads:

```
model {
for (i in 1:length(obsntot)){
obsnbkg[i]~dpois(nbkg[i])
obsnbluebkg[i]~dbin(fbkg[i],obsnbkg[i])
obsntot[i]~dpois(nbkg[i]/C[i]+nclus[i])
obsnbluetot[i]~dbin(f[i],obsntot[i])
f[i] <- (fbkg[i]*nbkg[i]/C[i]+fclus[i]*nclus[i])/
        (nbkg[i]/C[i]+nclus[i])
fclus[i] <- ilogit(lgfclus0+alpha*log(r200[i]/0.25)+
            beta*(lgM[i]-11)+gamma*(z[i]-0.3)+zeta*
            (lgM[i]-11)*(z[i]-0.3))
nbkg[i] ~ dunif(1,1e+7)
fbkg[i] ~ dbeta(1,1)
nclus[i]~ dunif(1,1e+7)
}
lgfclus0 ~dnorm(0,0.01)
alpha ~ dnorm(0,0.01)
beta ~ dnorm(0,0.01)
gamma ~ dnorm(0,0.01)
zeta ~ dnorm(0,0.01)
} .
```

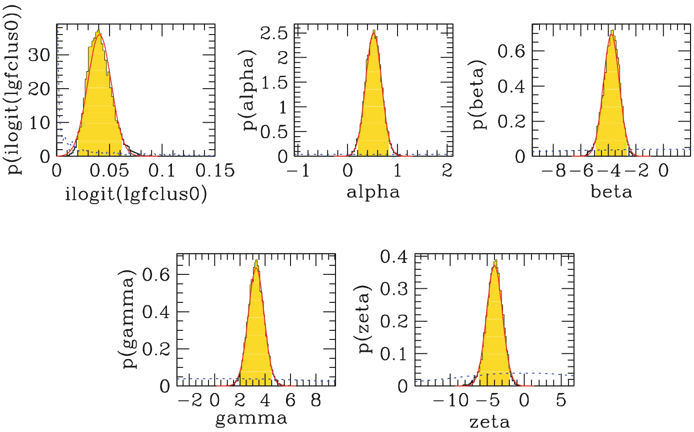

Fig. 8.17 Probability distribution for the parameters of the star formation quenching problem. The *black histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* shows a Gaussian approximation of the posterior. The *dashed curve* displays the adopted prior. The *shaded* (*yellow*) *range* shows the 95 % probability interval

Figure 8.16 shows the data, after co-adding data (to increase the signal to noise) concerning galaxies of similar mass, in similar environments and at similar redshifts, and the model fit, with 68 % errors. Posterior probability distributions of model parameters are shown in Fig. 8.17. The coefficient of the cross-term (`zeta`) avoids the zero value, i.e., this cross-term coefficient is needed (astronomers would say that mass quenching is dynamic).

Finally, Fig. 8.18 plots the data already shown in Fig. 8.16 for the inner radial bin in a more common form in astronomy, putting forward the just mentioned effect, galaxies of different masses (coded with different colors in the figure) quench (stop form stars, i.e., become red) at different redshifts.

8.6 Fit with Upper and Lower Limits: The Optical-to-X Flux Ratio

The presence of upper or lower limits does not change at all the fitting of a data set: a quantity does not have to be declared to be well measured (in some sense), or only have an upper limit (in some sense) to be used (although in general well-measured quantities are more informative than upper/lower limits). The analysis does not differ if the data includes upper limits, as opposed to some other approaches. Of course, the same is true for lower limits.

Let us consider the case of a linear fit between `magB`, the *B* band magnitude, and `lgfluxS`, the $\log_{10}$ of the flux in the [0.5, 2] keV band of a set of sources, that we know (or we suspect) to have a variety of intensity ratios. In the *B* and [0.5, 2] keV bands we have (see Sect. 6.1.2):

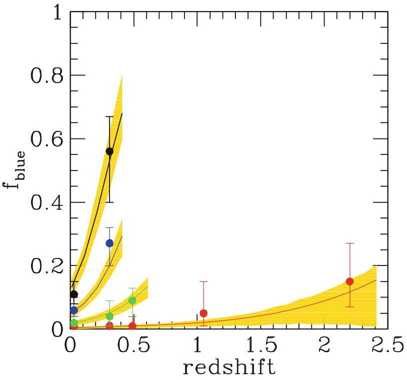

Fig. 8.18 Blue fraction for (coadded) clusters as a function of redshift in the inner cluster annulus for different galaxies' masses (*red, green, blue, and black points*), going from most to least massive. Error bars represent the 68 % probability interval. The *solid line, color-coded* as data, and the *yellow shaded areas* represent the posterior mean and its 68 % probability interval

```
obstotB[i] ~ dpois(sB[i]+bkgB[i]/CB[i])
obsbkgB[i] ~ dpois(bkgB[i])
obstotS[i] ~ dpois(sS[i]+bkgS[i]/CS[i])
obsbkgS[i] ~ dpois(bkgS[i]) ,
```

where `sB[i]` and `bkgB[i]` are the true counts of the source and background, and `CB[i]` is the ratio between the source and background solid angles in the *B* band (and similarly for the [0.5, 2] keV band). Observed values have a prefix `obs`. The analyzed sample, shown in Fig. 8.19, contains four different cases: upper limits in *B*, upper limits in [0.5, 2] keV, upper limits in both bands and no upper limits in any band.

Fluxes, magnitudes, and counts are related by (this is basic astronomy):

```
sB[i] <- pow(10,(zptB-magB[i])/2.5)
sS[i] <- pow(10,lgfluxS[i]-zptS[i]) ,
```

(the slightly different relations are due to slightly different habits in the *B* and X-ray domains), where the conversion from counts to flux in X-ray is object-dependent (i.e., has an `[i]`), as is usually the case in X-ray. To have a zero point differ from point-to-point in the *B* band too is easily implemented by adding a `[i]`.

As mentioned, we assume a linear relation with intrinsic scatter between `magB` and `lgfluxS`:

```
lgfluxS[i] ~ dnorm((magB[i]-22)*beta+alpha,pow(intrscat,-2)) .
```

We adopt a uniform prior for intrinsic scatter `intrscat`, and approximately uniform priors for `alpha` and `beta`,

```
intrscat ~ dunif(0,10)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1) ,
```

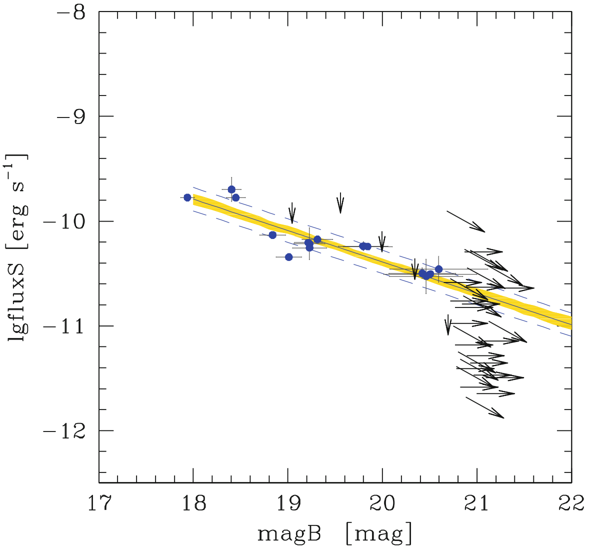

Fig. 8.19 Log of X-ray flux `lgfluxS` vs. `B` band magnitude `magB`. *Vertical, horizontal, and diagonal arrows* mark upper limits (see text for details). The *solid line* and the *yellow shaded areas* represent the posterior mean and its 68 % probability interval. The *dashed lines* delimit the mean model plus or minus the intrinsic scatter `intrscat`

and we assume uniform priors for the source and background intensities between zero and infinity, the latter taken here as 10^7 because we are dealing with fainter sources:

```
sB[i] ~ dunif(0,1e+7)
bkgB[i] ~ dunif(0,1e+7)
bkgS[i] ~ dunif(0,1e+7) ,
```

and a uniform prior over a wide range for the object's magnitude

```
magB[i] ~ dunif(18,25) .
```

Therefore, the fitted model reads:

```
model {
intrscat ~ dunif(0,10)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
zptB <-24
for (i in 1:length(obstotB)){
 obstotB[i] ~ dpois(sB[i]+bkgB[i]/CB[i])
 obsbkgB[i] ~ dpois(bkgB[i])
 obstotS[i] ~ dpois(sS[i]+bkgS[i]/CS[i])
 obsbkgS[i] ~ dpois(bkgS[i])
 magB[i] ~ dunif(18,25)
 sB[i] <- pow(10,(zptB-magB[i])/2.5)
 lgfluxS[i] ~ dnorm((magB[i]-22)*beta+alpha,pow(intrscat,-2))
 sS[i] <- pow(10,lgfluxS[i]-zptS[i])
 bkgB[i] ~ dunif(0,1e+7)
```

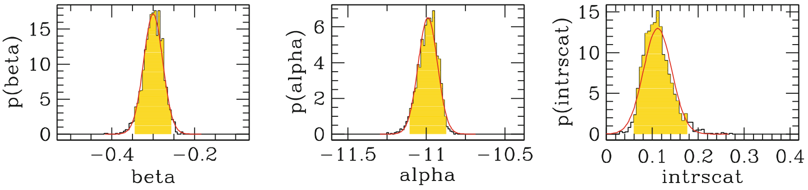

Fig. 8.20 Posterior distribution for the key parameters of the fit with upper limits

Table 8.4 Data of the first 10 entries of the upper limit example

obstotB[i]	obsbkgB[i]	obstotS[i]	obsbkgS[i]	zptS[i]
28	35	192	75	−13.72
37	30	277	74	−13.78
72	32	65	55	−11.44
46	48	813	70	−14.28
116	35	94	64	−11.69
47	40	75	64	−12.69
48	32	98	53	−11.95
84	38	549	65	−12.93
37	39	72	70	−11.82
51	40	56	75	−12.61

Note: The 10th object from the top, having obstotS[i] <obsbkgS[i] and the 1st, having obstotB[i] < obsbkgB[i] offer, obviously, only an upper limit to the source flux in the respective bands

```
 bkgS[i] ~ dunif(0,1e+7)
}
}.
```

The (fake) data (generated with the model listed at the end of this section) of 50 objects is shown in Fig. 8.19. When obstotB-obsbkg/CB$<$ $2\sqrt{\texttt{obstotB}+\texttt{obsbkgB}}$ or obstotS-obsbkgS/CS$<$ $2\sqrt{\texttt{obstotS}+\texttt{obsbkgS}}$ we plot an arrow, because the data only offer an upper limit of the object flux. Other criteria may be used to define the upper limit, but this would not change the analysis because such a definition is used only for drawing the plot; our fit does not use the arrows! As one may note, there are vertical arrows (X-ray upper limits), horizontal arrows (a lot, in the crowded region of the plot, these are *B* band upper limits), and diagonal arrows (upper limits in both bands). Indeed, 72 % of the flux measurements are upper limits.

We fit the data where the first 10 data points are listed in Table 8.4. The (posterior) probability distribution of the key parameters is shown in Fig. 8.20: we found a slope of -0.30 ± 0.02 (input was -0.3), an intercept of -10.99 ± 0.06 (input was -11.00), and an intrinsic scatter of 0.11 ± 0.03 (input was 0.10).

This example allows us to make a comment about prediction. As mentioned in Sect. 8.1.3, and throughout in Sect. 8.10, the existence of a tight relation between two quantities offers the possibility to predict one of them when the other is available. This is precisely the case of measurements with upper limits: objects with magB $\overset{<}{\approx}$ 21 mag are too faint to have a precisely determined *B* mag. However, owing to the tight relation with the X-ray flux, they may have a precisely inferred (predicted) *B* mag as good as intrscat/slope, $= 0.6$ mag in our fake example.

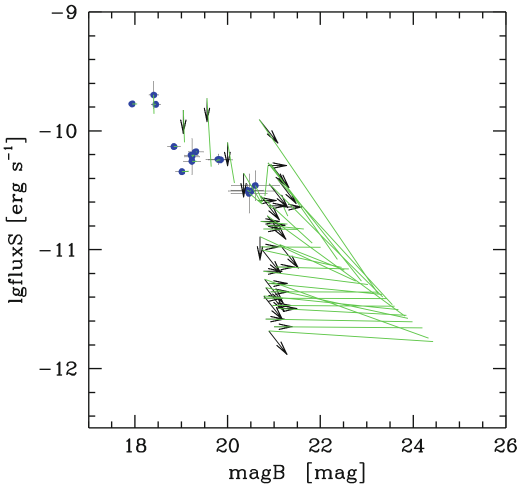

Fig. 8.21 Log of X-ray flux `lgfluxS` vs. *B* band magnitude `magB`. The *green lines* connect the observed values to the posterior mean (i.e., the predicted values). *Points and arrows* are described in Fig. 8.19

The same is true for predicting `lgfluxS` starting from `magB` values. Figure 8.21 visualizes how to change our knowledge from before the analysis (points and arrows) to after: the green lines connect the observed points to the posterior mean. To get the latter with JAGS, we only need, as for any other quantity, to save on disk the variable we are interested in, `magB` and `lgfluxS`. Clearly, the change will be mainly vertical when an upper limit of `lgfluxS` is available (vertical arrows), almost horizontal when an upper limit to `magB` is available (horizontal arrows), and oblique when we have upper limits on both quantities (oblique arrows).

To generate the sample we used the model:

```
model {
intrscat <- 0.1
alpha <- -11
beta <- -0.3
CB <-1
CS <-1
zptB <- 24
zptS ~ dnorm(-13,1)
 obstotB ~ dpois(sB+bkgB/CB)
 obsbkgB ~ dpois(bkgB)
 obstotS ~ dpois(sS+bkgS/CS)
 obsbkgS ~ dpois(bkgS)
 magB ~ dunif(18,25)
 sB <- pow(10,(zptB-magB)/2.5)
 lgfluxS ~ dnorm((magB-22)*beta+alpha,pow(intrscat,-2))
 sS <- pow(10,lgfluxS-zptS)
 bkgB ~ dunif(25,50)
 bkgS ~ dunif(50,70)
}.
```

We took `CB[i]` and `CS[i]` = 1 for all sources.

A fit between luminosities would proceed in an almost identical way, one only needs to change how to go from counts to luminosities (instead of fluxes).

8.7 Fit with An Important Data Structure: The Mass-Richness Scaling

In this section we illustrate the importance of not overlooking an important data structure. In particular, we focus on the danger of taking a uniform distribution for the quantity in the abscissa, implicitly assumed in many analyses published in astronomical or physics journals, when instead the quantity is not uniformly distributed. In particular, we focus on one real case of overlooking, noted in Andreon and Hurn (2010).

We want to infer the relation between mass and richness for a data set involving values of `obsn` as small as 3. We emphasized many times in this book that in nature there are many more small items (small richness `n` clusters) for each big item, and neglecting this fact may introduce a bias. This is currently the case. Accounting for the data structure, i.e., the important gradient in the abundance of clusters of a given `n`, is essential for determining the trend between richness and mass (i.e., the richness-mass calibration). To proceed as close as possible to the published work (whose reference is not given here for obvious reasons) we assume that only masses of stacks of clusters are available, i.e., we only have the average mass, `obslgMm`±`err`, of clusters of a given observed richness, because this is the case in question (and also what is often available in astronomy for poor clusters).

To have full control on what is going on, we generate fake data from a model with known values of the parameters (this is not possible with real data because in such cases the true values are unknown). We use input values for parameters adopting plausible values, as detailed below. Richness (`n`) are drawn from a Schechter (1976) function, with parameters $n^* = 10$ (1/scale, characteristic value) and $\alpha = -2$ (faint end slope), which are the values appropriate for a Jenkins et al. (2001) mass function and a linear mass-richness relation whose parameters are taken from Andreon and Hurn (2010). The large abundance gradient generated by the steep slope (-2) is the cause of the mentioned Malmquist-like bias: at a given `obsn`, the mean mass will be below the value one may naively expect, because most of these clusters will be, indeed, lower richness (mass) clusters scattered up by Poisson fluctuations.

We sample the Schechter (1976) function with the zero trick (see Sect. 3.2), because this distribution, with slope $\alpha = -2$, is not in the JAGS library. This obliges us to write the mathematical expression of the log of the likelihood:

```
data {
zeros<-0
C<-10
}
model {
n <- pow(10,lgn)
lgn ~ dunif(-1,3)
# -log likelihood
phi <- n/10^2-(-2+1-1)*lgn
zeros ~ dpois(phi+C)
} .
```

In this simulation, we neglect the background (to mimic the original analysis and to focus on the key points of this example, Sect. 8.10 does instead account for these) and thus

```
obsn ~dpois(n).
```

Individual `lgM` values scatter around a linear relation having slope `alpha` $= 1$ and intercept `beta`$=$ 14.4 with `sigmaintr`$=$ 0.20 (dex, Andreon and Hurn (2010)).

```
lgM ~ dnorm(alpha*(lgn-1.5)+beta,pow(sigmaintr,-2))
alpha <- 1
beta <- 14.4
sigmaintr <- 0.20.
```

As mentioned, the determination of the mass of these, low richness, individual clusters is outside our current reach and thus stacked analysis are often adopted. These stacked analysis compute the mean mass of clusters of a given `obsn`, and we do the same using the `lgM` values. Figure 8.23 shows the obtained richness-mass data points. In our simulation, we adopt simple mass errors, given by `err`$=$ $0.20/\sqrt{N}$, where N is the number of clusters in each `obsn` bin.

To sum up, the code used to generate the data reads:

```
data {
zeros<-0
C<-10
}
model {
# true model
lgM ~ dnorm(alpha*(lgn-1.5)+beta,pow(sigmaintr,-2))
obsn ~dpois(n)
n <- pow(10,lgn)
lgn ~ dunif(-1,3)
# -log likelihood
phi <- n/10^2-(-2+1-1)*lgn
tmp <- phi+C
zeros ~ dpois(phi+C)
alpha <- 1
beta <- 14.4
sigmaintr <- 0.2
}.
```

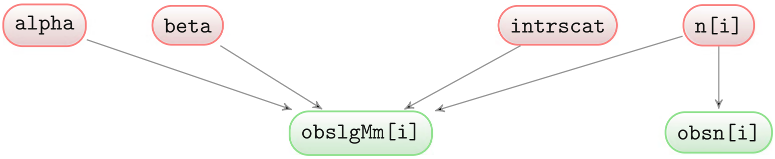

Fig. 8.22 Graph showing the stochastic relationship between quantities involved in estimating the parameters of the fit between richness and mass of stacks of clusters

The fitted model is very similar to the generating code and illustrated in Fig. 8.22. First, we need to model the data structure (the Schechter function), as in the generating code:

```
data {
zeros<-obsn-obsn
C<-10
}
model {
n[i] <- pow(10,lgn[i])
lgn[i] ~ dunif(-1,3)
# -log likelihood
phi[i] <- n[i]/10^2-(-2+1-1)*lgn[i]
zeros[i] ~ dpois(phi[i]+C)
}.
```

As before, observed `obsn[i]` values are Poisson distributed around their true values:

```
obsn[i] ~dpois(n[i]).
```

The stacked `obslgMm[i]` values scatter around a linear relation having slope `alpha` and intercept `beta` which is to be estimated

```
obslgMm[i] ~ dnorm(lgM[i],pow(err[i],-2))
lgM[i] <- alpha*(lgn[i]-1.5)+beta.
```

For these two parameters we adopt weak priors, a uniform distribution on the angle and a Gaussian with $\sigma = 3$ for the intercept (which is fairly flat in the range of plausible values, (13, 14.6):

```
alpha ~ dt(0,1,1)
beta ~ dnorm(14.4,pow(3,-2)).
```

Therefore, the full analysis code reads:

```
data {
zeros<-obsn-obsn
C<-10
}
model {
for (i in 1:length(obsn)) {
obslgMm[i] ~ dnorm(lgM[i],pow(err[i],-2))
lgM[i] <- alpha*(lgn[i]-1.5)+beta
obsn[i] ~dpois(n[i])
```

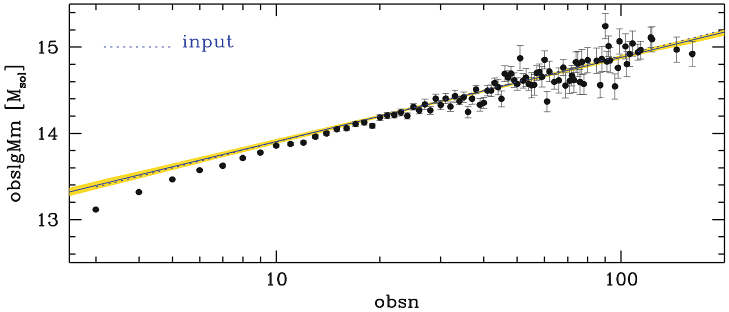

Fig. 8.23 Richness-mass scaling for our simulated data with an important data structure. The *solid line* marks the mean fitted regression line of lgMm on log(n). The *shaded region* marks the 68 % probability interval for the regression. The input relation is also plotted (*dotted*), but hardly distinguishable from the mean (recovered) model. The data deviates from the input relation because of the population gradient (Malmquist bias)

```
n[i] <- pow(10,lgn[i])
lgn[i] ~ dunif(-1,3)
# -log likelihood
phi[i] <- n[i]/10^2-(-2+1-1)*lgn[i]
zeros[i] ~ dpois(phi[i]+C)
}
alpha ~ dt(0,1,1)
beta ~ dnorm(14.4,pow(3,-2))
}.
```

The fit of the model to the data is depicted in Fig. 8.23, and gives a slope of 0.98±0.04 (input was 1) and an intercept of 14.39±0.01 (input was 14.40). Clearly, it recovers the input values. The figure also clearly shows that the model does not pass through the data points at low obsn. This is indeed correct, because the trend is between true values, and obsn values are biased estimates of n at small values: because of the population gradient (Malmquist bias), clusters with low richness obsn are most likely of lower richness n<obsn because of the scatter. They thus have lower masses than their obsn apparently tell. Since the bias depends on the $\sqrt{\mathtt{obsn}/\mathtt{n}}$, it is minor at large obsn and a tilt on the slope is implied. It is equally obvious that methods not modeling the data structure and that pass through the data would recover a steeper, biased high, trend (by five times the quoted uncertainty for the published work).

Of course, this Malmquist-like bias affects the slope, and this feature is a general one, pertaining to all quantities that have abundance gradients, no matter how they are named: mass, richness, velocity dispersion, luminosity, color, parallax, etc.

8.8 Fit with a Non-ignorable Data Collection[4]

Data collection effects cannot be ignored when easier-to-detect objects enter with a higher probability than the ones which are more difficult to detect. As shown in Andreon and Bergé (2012), this is the typical case of weak-lensing detected clusters.

[4] This section draws material from Andreon and Bergé (2012).

Weak-lensing is a (coherent) deformation of the images of background sources induced by the deformation of the space-time manifold induced by the gravitational potential of a cluster of galaxies on the line of sight between us and the background sources. As a result of the deformation, background galaxies will appear slightly elongated in the direction tangential to the line connecting them to the cluster's center. The deformation is minor, and in addition, we do not know the true shape of any of the background galaxies. In practice, clusters are hard to detect with the weak-lensing effect, and even when detected their detection has a small signal-to-noise ratio. In spite of the difficulty of this measurement, weak-lensing offers a direct measurement of mass, which is useful for calibrating the relation with mass proxies, such as richness, and from these we are able to infer about cosmological parameters. However, as shown here following Andreon and Bergé (2012), the determination of the scaling between mass and richness has to account for the weak-lensing selection function; not accounting for the selection function would return a wrong calibration.

We consider here the case of the most ambitious cosmological-devoted space survey, Euclid. In order to show the impact of the non-random data collection, we use simulated data (from Andreon and Bergé 2012) because the satellite is under construction at the time of the writing of this book. The sample is formed simulating 10714 clusters with $S/N > 5$. The (simulated) data consists of measured values of richness `obsn[i]`, mass `obslgM[i]` (with error `errlgM[i]`), and an estimate of the cluster's distance (photometric redshift) `obsz[i]` (see the bottom line of the graph in Fig. 8.24). We use these values to determine the relation between richness and mass and its evolution, parameterized with a five-parameter function whose parameters are given in the five left-most upper boxes in Fig. 8.24.

Figure 8.25 depicts the observed values individually (points). The lack of points in the bottom-right corner of the figure is due to non-random data collection: these clusters are undetectable with Euclid's data (as faint radiogalaxies were at the time of Sandage, see Fig. 8.1). Following Andreon and Bergé (2012), clusters enter in the sample if their mass is larger than

```
lgMtruc = 13.9891+1.04936 *obsz[i]+0.488881 *obsz[i]^2 .
```

Note that the figure plots observed values of mass, and therefore there are points below `lgMtrunc` (the diagonal curve in the figure).

Following Andreon and Bergé (2012), richness is Poisson distributed:

```
obsn[i] ~ dpois(pow(10, lgn[i])) ,
```

whereas masses and photometric redshifts have Gaussian errors:

```
obslg[i] ~ dnorm(lgM[i],pow(errlgM[i],-2))
obsz[i] ~ dnorm(z[i],pow(0.02,-2)) .
```

We use the posterior distributions measured from the local universe and given in Andreon and Bergé (2012) as the prior distributions for the slope `beta`, intercept `alpha`, and intrinsic scatter `intrscat` (remember, yesterday's prior is today's posterior):

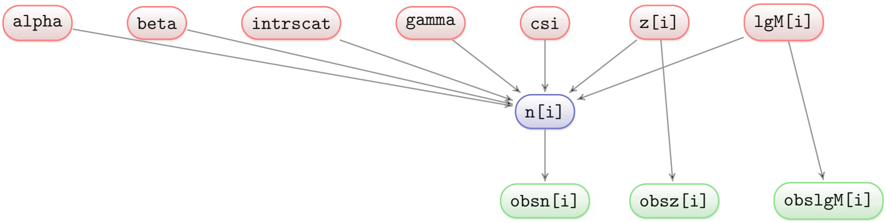

Fig. 8.24 Graph showing the stochastic relationship between quantities involved in estimating the parameters of the fit between richness and weak-lensing masses

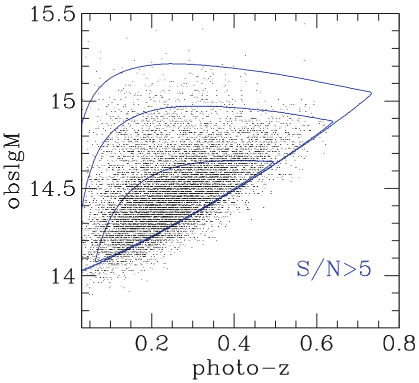

Fig. 8.25 Contours: number of clusters for which weak-lensing mass estimates can be obtained by a Euclid-like survey. From outer to inner contours, the lines represent isocontours of $S/N > 5$ weak-lensing detection of $1, 10$, and 100 clusters as a function of redshift and mass. Points: a Poisson realization of the above, with errors on mass and redshift (these move points outside the $N = 1$ contour). Reproduced from Andreon and Bergé (2012) with permission

```
intrscat ~ dnorm(0.25,pow(0.03,-2))
alpha ~ dnorm(0.08,pow(0.04,-2))
beta ~ dnorm(0.47,pow(0.12,-2) .
```

Following Andreon and Bergé (2012), we assume that the scatter and the intercept may both change with redshift

```
lgnm[i] <- alpha+1.5 +beta*(lgM[i]-14.5)+ gamma*(log(1+z[i]))
lgn[i] ~ dnorm(lgnm[i], pow(intrscat.z[i],-2))
intrscat.z[i] <- sqrt(pow(intrscat,2)-1+(1+z[i])^(2*csi)) ,
```

and we adopt weak priors for the newly introduced parameters. First we take a Student-t distribution centered on zero with 1 degree of freedom for `gamma` and `csi` slopes, similar to other slopes in this book, to make our choice independent of astronomer's rules for measuring angles:

```
gamma ~dt(0,1,1)
csi ~dt(0,1,1) .
```

We assume a uniform distribution for the redshift prior

```
z[i]~dnorm(0,1).
```

We now specify the prior on `lgM`. We cannot ignore the fact that the mass function is steep and that the data collection is non-random, ignoring them would lead to a biased fit: the recovered slope would be much shallower than the input one, due to a Malmquist-like bias. In fact, mass errors tend to make the mass' distribution broader at both low mass values because of the non-random data collection (alias the weak-lensing detection), and large mass values because of the steepness of the mass function. Since high-mass values are overestimated and low-mass values are underestimated, any quantity that is fitted against these (biased) values, neglecting the selection function, would return a shallower relation. From theory (following a path quite similar to the one adopted in Sect. 8.7), Andreon and Bergé (2012) found that the prior (mass function) at a given redshift is well approximated by a Schechter (1976) function with slope -1, and characteristic mass given by

```
lgM* = 12.6- (obsz[i]-0.3)
```

truncated at `lgMtruc`, computed as before (see Andreon and Bergé 2012 for details about this computation).

The coding in JAGS for the above model requires the zero trick to sample distributions not in JAGS. Since we are studying a sample with a non-random selection, we also need an integral (see Chap. 7.1) for each `obsz[i]`, which we numerically computed and found to be well approximated by

```
lg10tot.norm <-0.386165-3.92996*obsz-0.247050*obsz^2-
               2.55814*obsz^3-5.26633*obsz^4 .
```

To sum up, our model reads:

```
data{
# normaliz
lg10tot.norm <-0.386165-3.92996*obsz-0.247050*obsz^2-2.55814*
               obsz^3-5.26633*obsz^4
# dummy variable for zero-trick, to sample from a distribution
# not available in JAGS
for (i in 1:length(obslgM)) {
dummy[i] <-0
}
C<-2
}
model{
intrscat ~ dnorm(0.25,pow(0.03,-2))
alpha ~ dnorm(0.08,pow(0.04,-2))
beta ~ dnorm(0.47,pow(0.12,-2))
gamma ~dt(0,1,1)
csi ~dt(0,1,1)
for (i in 1:length(obsn)) {
 # modeling lgM
 # dummy prior, requested by JAGS, to be modified later
 lgM[i] ~ dunif(13.9891+1.04936*obsz[i]+0.488881*
                obsz[i]^2,16)
 # modeling a truncated schechter
 lnnumerator[i] <- -(10^(0.4*(lgM[i]-12.6+(obsz[i]-0.3))))
 # its integral, from the starting point of the integration
 # (S/N=5)
 loglike[i] <- -lnnumerator[i]+lg10tot.norm[i]*log(10)+C
 # sampling from an unavailable distribution
 dummy[i] ~ dpois(loglike[i])
 obslg[i] ~ dnorm(lgM[i],pow(errlgM[i],-2))
 # modeling n, z and relations
 obsn[i] ~ dpois(pow(10, lgn[i]))
 obsz[i] ~ dnorm(z[i],pow(0.02,-2))
 z[i]~dnorm(0,1)
 # modeling mass -n relation allowing evolution
 lgnm[i] <- alpha+1.5 +beta*(lgM[i]-14.5)+ gamma*(log(1+z[i]))
 lgn[i] ~ dnorm(lgnm[i], pow(intrscat.z[i],-2))
 intrscat.z[i] <- sqrt(pow(intrscat,2)-1+(1+z[i])^(2*csi))
 }
}.
```

Fitting the 10714 masses, richness and photometric redshift with this model returns parameters whose (posterior) probability distributions are depicted in Fig. 8.26. Figure 8.26 is the desired quantity: how well we will be able to measure the richness-mass scaling with Euclid data. This is an indispensable ingredient for estimating how well we will be able to measure cosmological parameters using clusters of galaxies. Marginal probabilities are shown on the diagonal panels, whereas the off-diagonal panels show the joint probability distributions, i.e., the covariance between pairs of parameters in the posterior distribution. Some contours are somewhat noisy because of the finite length of the MCMC chain. The diagonal panels also show the input values (vertical lines). They are all within 1.5 posterior standard deviations from the recovered value.[5] By fitting the observed data, we recover, with good precision and without bias (i.e., high accuracy), the five (input) parameters describing the mass-richness scaling. Accounting for the non-random data collection (the weak-lensing selection) is compulsory, not accounting for it would lead to a fitted slope $\gg 5\sigma$ away from the input one (0.47).

In addition to the input values, the diagonal panels show the prior, i.e., the current low-redshift calibration of the richness-mass scaling (dashed green line).

There is a strong covariance between the evolution and the $z = 0$ value of the intercept (see the off-diagonal panels in Fig. 8.26). It can be easily understood by noting that $z = 0$ is outside the range of sampled redshifts. The covariance between intrinsic scatter and its evolution has a similar origin: the intrinsic scatter is defined at an unobserved redshift, $z = 0$, instead of a redshift where it is well observed.

Figure 8.27 compares the model fit (solid line) to the true input relation in stacks of 201 clusters per point. There is a good agreement between the model fit and the (unobserved and unused in the analysis) noise-less data, indicating that the fit to the noisy data captures well the real trend of the noise-less true data.

[5] There is only a 10 % probability that in a 5-parameter fit, all fitted values are found within 1 σ from the input values and 50 % that they are all within 1.5 σ.

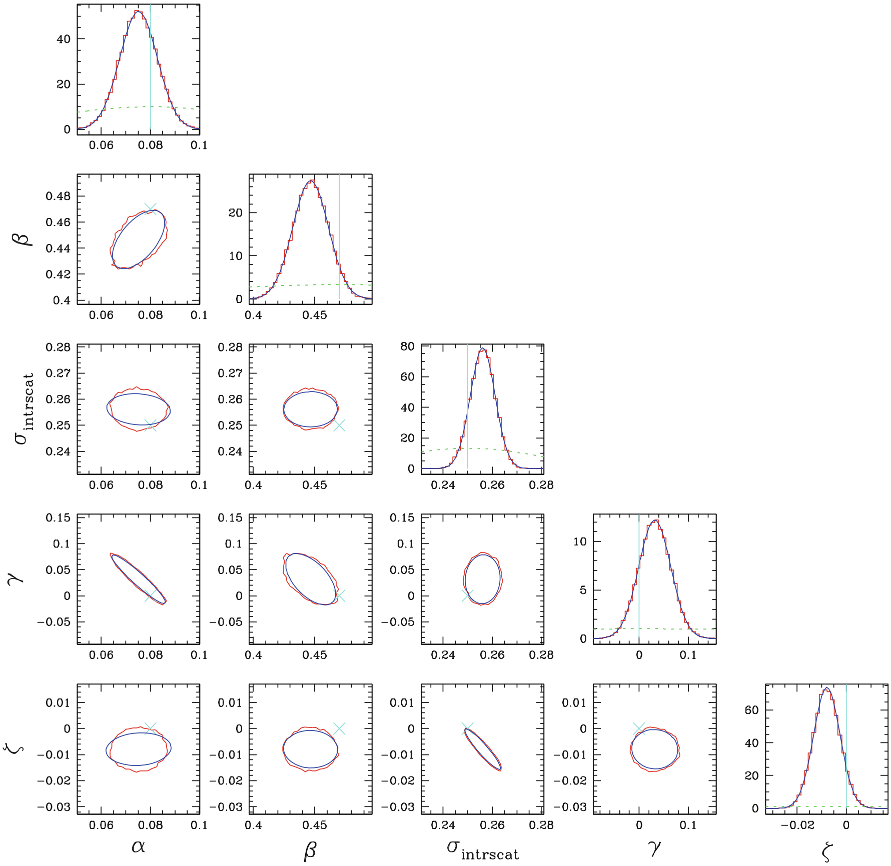

Fig. 8.26 Marginal (panels on the diagonal) and joint (panels on the off-diagonals) posterior probability distributions of the mass-richness scaling derived from simulated Euclid data. Red jagged contours and histograms refer to probabilities computed from the MCMC sampling, whereas the blue smooth contours refer to a Gaussian approximation of the numerically found posterior. Contours are a 68 % probability level. *Vertical (cyan) lines and crosses* indicate the values used to generate the data, whereas the *dashed (green) lines* show the prior probability distributions (the current low-redshift calibration of the richness-mass scaling). Reproduced from Andreon and Bergé (2012) with permission

8.9 Fit Without Anxiety About Non-random Data Collection

Most samples available in astronomy are a non-random subsampling of what exists in the Universe, because what is available is a either a collection of objects with an unknown selection function (i.e., without a known function that gives the probability that an object enters in the sample as a function of object parameters) or because the

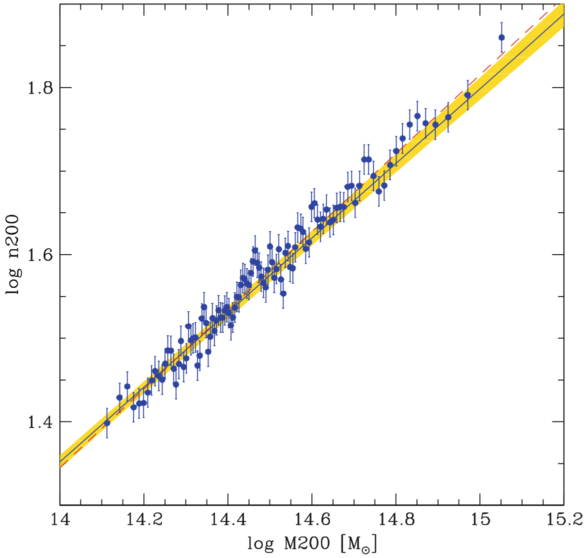

Fig. 8.27 Richness-mass scaling for the simulated Euclid data. The *solid line* marks the regression line fitted on observed data. The *shaded region* marks the 68 % probability interval for the regression. The *red dashed line* indicates the input relation. The data points are stacks of true data in bins of 201 clusters each, which was not used in the fitting. Reproduced from Andreon and Bergé (2012) with permission

sample has a known selection function (as in surveys), but easier-to-observe objects are over-represented compared to difficult-to-observe objects.

While it must not be ignored that available samples are not a random sampling of what exists (unless one is happy with providing biased estimates of target quantities), the importance of the non-random data collection should not be overemphasized because not all collections of non-random data display biases. We illustrate this case by considering an X-ray selected cluster sample. In spite of an obvious non-random data collection (brighter X-ray clusters are over-represented in the sample), the non-random data collection is ignorable for deriving the scaling between richness and mass in the local Universe for the sample considered in Andreon and Bergé (2012) and also in this section.

The sample selection function (i.e., the probability that a cluster is in the sample) is fully computed in Andreon and Bergé (2012), but, in short, it follows the same path already mentioned a couple of times in this chapter: random samples are drawn from the cluster's mass function, then "observed" and kept in the collected sample if they satisfy the conditions used for inclusion in the real sample. For the sample of interest here, listed in Table 8.5 and taken from Andreon and Hurn (2010), Andreon and Bergé (2012) derived:

```
lgM[i] ~ dnorm(14.5,pow(0.33,-2)) .
```

Table 8.5 Data for Sects. 8.9 and 8.10, from Andreon and Hurn (2010)

ID	obstot[i]	obsbkg[i]	C[i]	obslgM[i]	errlgM[i]
A0160	29	13	2.951	13.99	0.1928
A0602	23	37	10.77	14.79	0.04194
A0671	36	20	5.443	14.66	0.1192
A0779	19	0	0.4303	14.41	0.06695
A0957	26	20	7.947	14.35	0.2096
A0954	28	168	41.05	13.93	0.242
A0971	50	127	19.57	14.80	0.07902
RXCJ1022.0+3830	26	28	10.27	14.19	0.2592
A1066	65	41	6.421	14.83	0.04492
RXJ1053.7+5450	40	70	13.77	14.69	0.04863
A1142	15	1	1.606	14.5	0.1129
A1173	27	110	30.45	14.13	0.1429
A1190	63	88	9.896	14.53	0.1693
A1205	42	67	11.84	14.47	0.1286
RXCJ1115.5+5426	45	50	10.48	14.58	0.05061
SHK352	32	24	6.125	14.59	0.03659
A1314	33	5	1.832	14.39	0.09893
A1377	50	48	5.913	14.27	0.231
A1424	39	45	13.47	14.60	0.1142
A1436	64	51	8.021	14.09	0.09893
MKW4	19	1	0.5456	14.38	0.0421
RXCJ1210.3+0523	36	67	19.22	14.02	0.2159
Zw1215.1+0400	90	62	7.965	14.55	0.2723
A1552	78	113	11.33	14.41	0.202
A1663	55	86	12.23	14.76	0.1322
MS1306	19	104	44.83	13.92	0.08167
A1728	22	135	26.7	14.62	0.1945
RXJ1326.2+0013	12	118	57.11	13.91	0.2173
MKW11	9	8	4.284	13.69	0.1406
A1750	71	86	12.32	14.36	0.2772
A1767	59	35	6.624	14.94	0.1434
A1773	49	90	15.08	14.45	0.2922
RXCJ1351.7+4622	29	31	13.96	13.73	0.07343
A1809	67	121	11.66	14.28	0.1996
A1885	21	74	50.58	14.81	0.1057
MKW8	17	8	3.39	13.8	0.2704
A2064	22	47	21.44	14.37	0.1405
A2061	85	80	7.381	14.88	0.03937
A2067	28	128	24.3	13.77	0.2001
A2110	32	176	34.32	14.40	0.0796
A2124	48	29	6.036	14.94	0.06879
A2142	186	115	6.141	14.68	0.1478
NGC6107	22	10	4.034	14.18	0.1691
A2175	71	77	14.64	14.05	0.1277
A2197	63	3	0.8029	14.02	0.1936
A2199	88	0	0.239	14.53	0.1213
A2245	88	80	8.376	14.88	0.04412
A2244	99	112	11.75	14.79	0.1607
A2255	173	60	3.514	15.03	0.07326
NGC6338	16	2	1.068	14.34	0.2298
A2399	56	48	7.135	14.16	0.1029
A2428	33	154	25.2	14.31	0.1383
A2670	109	41	4.442	14.43	0.1882

Roughly speaking, massive clusters are rare, those not massive enough have a low chance of entering in the sample because they are hard to observe. Therefore the selection function should go to zero at both ends, as the Gaussian above.

The modeling of richness in the presence of a background has been already presented a number of times in this book (see Fig. 8.28):

```
obsbkg[i] ~ dpois(nbkg[i])
obstot[i] ~ dpois(nbkg[i]/C[i]+10^lgn[i])
nbkg[i] ~ dunif(0,3000) .
```

The variables `n[i]` and `nbkg[i]` represent the true richness and the true background galaxy counts in the studied solid angles, whereas we add a prefix `obs` to indicate the observed values. We allow Gaussian errors on mass:

```
obslgM[i] ~ dnorm(lgM[i],pow(errlgM[i],-2)) .
```

Following Andreon and Bergé (2012), we assume a power-law relation between mass and richness `n` with intercept `alpha`+1.5, slope `beta` and intrinsic scatter `intrscat`:

```
lgnm[i] <- alpha+1.5 +beta*(lgM[i]-14.5)
lgn[i] ~ dnorm(lgnm[i], pow(intrscat,-2)) .
```

As usual, we centered quantities for improving the MCMC efficiency used to fit the model and to reduce the covariance between parameter estimates.

Priors are taken to be quite flat:

```
intrscat <- 1/sqrt(prec.intrscat)
prec.intrscat ~ dgamma(1.0E-5,1.0E-5)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
nbkg[i] ~ dunif(0,3000) .
```

To summarize, the model reads:

```
model {
intrscat <- 1/sqrt(prec.intrscat)
prec.intrscat ~ dgamma(1.0E-5,1.0E-5)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
for (i in 1:length(obstot)) {
 obsbkg[i] ~ dpois(nbkg[i])
 obstot[i] ~ dpois(nbkg[i]/C[i]+10^lgn[i])
 nbkg[i] ~ dunif(0,3000)
 # modeling data selection
 lgM[i] ~ dnorm(14.5,pow(0.33,-2))
 # decomment for ignoring data selection
 # lgM[i] ~dunif(13,16)

 obslgM[i] ~ dnorm(lgM[i],pow(errlgM[i],-2))
 # modeling mass -n relation
 lgnm[i] <- alpha+1.5 +beta*(lgM[i]-14.5)
 lgn[i] ~ dnorm(lgnm[i], pow(intrscat,-2))
}
}.
```

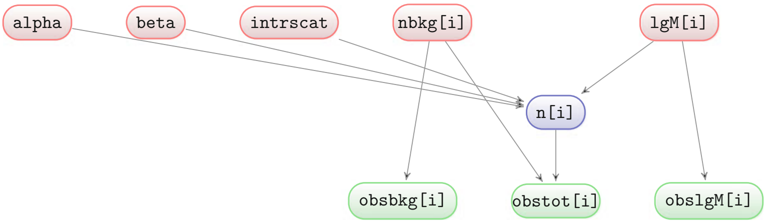

Fig. 8.28 Graph showing the stochastic relationship between quantities involved in estimating the parameters of the fit between richness and masses for a non-random sample, but with an ignorable data collection

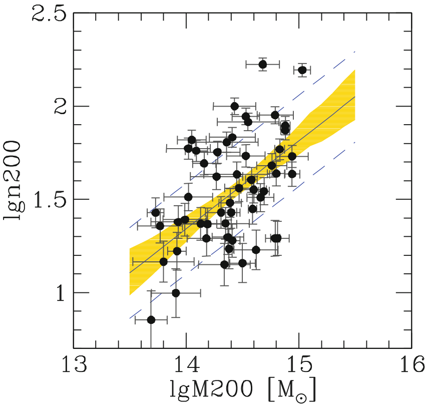

Fig. 8.29 Richness-mass scaling. The *solid line* marks the mean fitted regression line of $log(n200)$ on $lgM200$, while the *dashed lines* show this mean plus or minus the intrinsic scatter σ_{scat}. The *shaded region* marks the 68 % probability interval for the regression. Error bars on the data points represent observed errors for both variables. The distance between the data and the regression line is due in part to the measurement error and in part to the intrinsic scatter. Reproduced from Andreon and Bergé (2012) with permission

Fitting the sample of 52 clusters listed in Table 8.5 with the model above yields:

$$\texttt{lgn} = (0.47 \pm 0.12)\,(\texttt{lgM} - 14.5) + 1.58 \pm 0.04 \quad . \tag{8.7}$$

Figure 8.29 shows the scaling between richness and mass, the observed data, the mean scaling (solid line), and its 68 % uncertainty (shaded yellow region) and the mean intrinsic scatter (dashed lines) around the mean relation.

Figure 8.30 shows the posterior marginals for the key parameters, i.e., for the intercept, slope, and intrinsic scatter `intrscat`. The intrinsic mass scatter at a given richness, `intrscat`$= \sigma_{lgM|\log n}$, is small, 0.25 ± 0.03 dex.

These posterior probability distributions do not depend, to an appreciable level, on the adopted prior because the posterior is dominated by the data. In fact, we refit the same data by using a uniform prior on `lgM[i]` (a manifestly wrong population structure, which ignores both the rarity of massive clusters and the under-

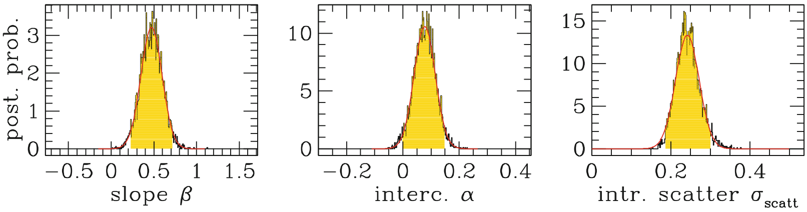

Fig. 8.30 Posterior probability distribution for the parameters of the richness-mass scaling computed from the real data. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation to the posterior. The *shaded (yellow) range* shows the 95 % probability interval. Reproduced from Andreon and Bergé (2012) with permission

representation in the sample of less massive clusters) which returns almost indistinguishable results. This occurs because the prior is broad (has a sigma of 0.33 dex) compared to `errlgM[i]` (typically 0.1 dex). This example shows that a (specific) sample whose data collection selection is not random may give a sample whose data collection is ignorable for the studied trend. Of course, the benign effect of the selection function cannot be taken for granted, and must be checked case by case.

8.10 Prediction[6]

In previous sections, we largely addressed the parameter estimation problem, i.e., we were interested in the parameters quantifying the (sometimes non-linear) trend between two or more quantities. As mentioned, this is one of the few reasons why quantities are regressed. Another reason is prediction.

There are plenty of quantities in astronomy that are too observationally expensive to measure for large samples, for example mass. In such cases, astronomers use proxies, which are far less expensive to acquire: from a typically small sample of objects, the researchers measure the target quantity, y and the (mass) proxy, x. Then, they regress x vs. y and infer y for those objects having only x. This is the most common way galaxy-cluster masses are estimated, for example using the X-ray luminosity, X-ray temperature, the cluster's richness, the total optical luminosity, or other proxies as well.

Let us restate the above in more mathematical terms. Suppose we want to estimate the value of a quantity not yet measured (e.g., the mass of a not yet weighed cluster). Before data y are collected (or even considered), the distribution of the predicted values $\tilde{y}$ can be expressed as

$$p(\tilde{y}) = \int p(\tilde{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \qquad (8.8)$$

[6] Part of the material of this section has been drawn from Andreon and Hurn (2010).

These two equalities result from the application of probability definitions (see Chap. 2), the first equality is simply that a marginal distribution results from integrating over a joint distribution (Eq. 2.5), the second one is simply an application of the conditional probability (Sect. 2.4). This result is called the prior-predictive distribution.

If some data y have been already collected for similar objects, we can use these data to improve our prediction for $\tilde{y}$. For example, if mass and richness in clusters are highly correlated, one may better predict the cluster's mass knowing its richness than in the absence of such information, simply because mass shows a lower scatter at a given richness than when clusters of all richnesses are considered (except if the relationship has slope exactly equal to $\tan k\pi/2$, with $k = 0, 1, 2, 3$).

By explicitly acknowledging the presence of such data, y, we rewrite Eq. (8.8) by conditioning on y:

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta. \tag{8.9}$$

The conditioning on y in the first term in the integral simplifies out because y and $\tilde{y}$ are considered conditionally independent given θ, so that this term becomes simply $p(\tilde{y}|\theta)$. The left-hand side of the equation is called the posterior predictive distribution for a new, unobserved $\tilde{y}$ given observed data y and marginalized with respect to the posterior distribution of our model parameters θ. Its width is a measure of the uncertainty of the predicted value $\tilde{y}$, a narrower distribution indicates a more precise prediction. The quantity computed in Eq. (8.9) is named the posterior predictive distribution.

How does this work in practice? Let us first consider a simple example. Suppose we do not know the mass, $\widetilde{lgM}$, of a given cluster and we are interested in predicting from it our knowledge of its richness. In this didactic example we assume for simplicity that

a) all probability distributions are Gaussian,
b) that previous data lgM for clusters of the same richness allowed us to determine that clusters of that richness have on average a mass of $lgM = 13.3 \pm 0.1$, i.e., $p(\theta|lgM) = \mathcal{N}(13.3, 0.1^2)$,
c) that the scatter between the individual and the average mass of the clusters is 0.5 dex, i.e., $p(\widetilde{lgM}|\theta) = \mathcal{N}(\theta, 0.5^2)$.

Then, Eq. (8.9) can be written in closed form and gives the intuitive solution that $p(\widetilde{lgM}|lgM)$ is a Gaussian centered on $lgM = 13.3$ and with a σ given by the sum in quadrature of 0.1 and 0.5 (= 0.51 dex). Therefore, a not-yet weighed cluster of the considered richness has a predicted mass of 13.3 with an uncertainty of 0.51 dex.

This, didactic example, shows that one may analytically predict a quantity, and its uncertainty. Realistic cases are also addressed by Eq. (8.9), but its evaluation will likely be a numerical one. The latter is offered by JAGS as a standard feature. In practice, the integral in Eq. (8.9) is computed quite simply using sampling:

1. Randomly draw θ from the posterior $p(\theta|y)$.
2. For each of the randomly drawn θ, randomly select values of $\tilde{y}$ from $p(\tilde{y}|\theta, y)$.

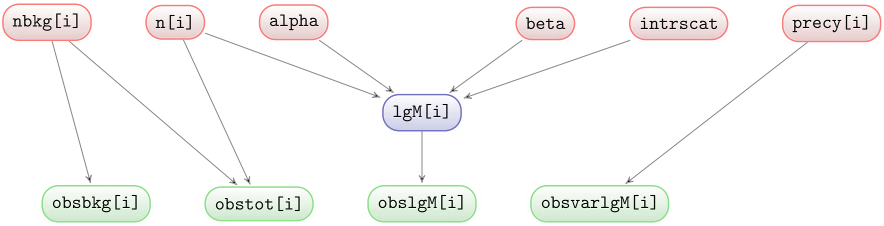

Fig. 8.31 Graph showing the stochastic relationship between quantities involved in predicting masses from richnesses

The width of the distribution of these $\tilde{y}$ values gives the uncertainty of the predicted value. Therefore, the quoted performance accounts for all terms entering into the modeling of the predictor and target, which includes the uncertainty of the predictor, the uncertainty on the parameters describing the regression between predictor and target (slope, intercept, intrinsic scatter, and their covariance), as well as other modeled terms. Some factors are automatically accounted for without any additional input, for example, where data are scarce, or when we are sampling near or outside the given range, predictions are noisier (because the regression is poorly determined in this region). We already used Eq. (8.9) in the upper limit example (Sect. 8.6), it allowed us to improve the measurements for which we only have upper limits.

Once the above, somewhat technical, text is digested (or at least believed to be true), the only thing to remember is that in order to predict $\tilde{y}$ for a new object, we have to put the x values for which we want the prediction in the data file, and "NA" (for not available) as their y values. JAGS will return a sampling of the posterior predictive distribution, that we will use as all other posterior samplings. If the object is already in the list, the user has nothing to do, the posterior distribution is computed by default.

Andreon and Hurn (2013) and March et al. (2011) show that the Bayesian approach to prediction is useful: it is more precise than any other method and lacks the large systematic biases of other approaches (see also Chap. 10). This means, in the case of masses, that more precise masses can be derived for the same input data, i.e., at the same telescope-time cost.

Let us now consider a real case, the mass-richness scaling discussed in Andreon and Hurn (2010). Mass estimates are one of the holy grails of astronomy. Since these are observationally expensive to measure, or even unmeasurable with existing facilities, astronomers use mass proxies, which are far less expensive to acquire and are used to estimate masses. Here we consider the cluster's richness, i.e., the number of member galaxies. The model graph is shown in Fig. 8.31.

As usual, we do not measure the true value of the cluster's richness, but only the total number of galaxies in the cluster's line-of-sight and in a control-field region, the cluster's richness being the difference of the true values. We already encountered this problem many times, and by now we should know how to model it:

```
obsbkg[i] ~ dpois(nbkg[i])
obstot[i] ~ dpois(nbkg[i]/C[i]+n[i]) .
```

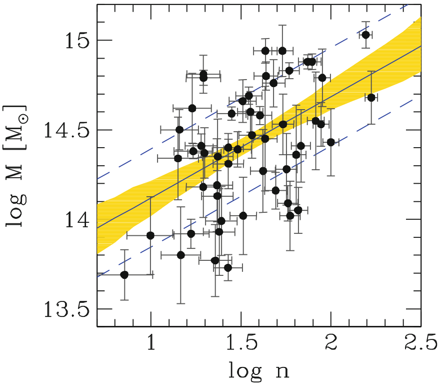

Fig. 8.32 Richness-mass scaling. The *solid line* marks the mean fitted regression line of `lgM` on `log(n)`, while the *dashed line* shows this mean plus or minus the intrinsic scatter `intrscat`. The *shaded region* marks the 68 % probability interval for the regression. Error bars on the data points represent observed errors for both variables. Adapted from Andreon and Hurn (2010), reproduced with permission

The variables `n[i]` and `nbkg[i]` represent the true richness and the true background galaxy counts in the studied solid angles, whereas we add a prefix `obs` to indicate the observed values. The data are listed in Table 8.5 and shown in Fig. 8.32. Upper limits to `n` are automatically accounted for in our fit (see Sect. 8.6).

The relation between mass, M, and proxy, n, (richness) is usually parametrized as a power-law with slope `beta` and intercept `alpha`+14.5:

```
z[i] <- alpha+14.5+beta*(log(n200[i])/2.30258-1.5) .
```

There is, as usual, a Gaussian intrinsic scatter `intrscat`

```
lgM[i] ~ dnorm(z[i], pow(intrscat,-2)) ,
```

and Gaussian errors on log mass,

```
obslgM[i] ~ dnorm(lgM[i],precy[i]) .
```

In this specific example we have an additional complication: mass errors are difficult to measure, and therefore we only know their approximate size. Errors (or anything we measure, indeed) are measured with a finite degree of precision. We assume that the measured error, `obserrlgM200[i]`, is not biased (i.e., it is not systematically larger or smaller than the true error, `sigma[i]`, but somewhat noisy). If a χ^2 distribution is adopted, it satisfies both our request of unbiasedness and noisiness. In formulae:

$$obserrlgM_i^2 \sim \sigma_i^2 \chi_\nu^2 \tag{8.10}$$

or, in JAGS,

```
obsvarlgM200[i] ~ dgamma(0.5*nu,0.5*nu*precy[i])
```

where we used the property that the χ^2 distribution is a particular form of the gamma distribution.

The parameter ν regulates the width of the distribution, i.e., how precise measured errors are. Following Andreon and Hurn (2010), we are 95 % confident that quoted errors are correct up to a factor of 2,

```
nu <-6 .
```

The remaining part of the model is the same already seen several times in this book. The data used in this investigation have a high enough quality that all parameters can be estimated, except `precy[i]`, to a sufficient degree of accuracy that we should not care too much about priors and we can safely take weak (almost uniform) priors, zeroed for un-physical values of parameters (to avoid, for example, negative richnesses). The exception is given by the prior on the errors (i.e., σ_i), for which there is only one measurement per datum, for which we take a Gamma prior:

```
precy[i] ~ dgamma(1.0E-5,1.0E-5)
```

following the authors (Andreon and Hurn 2010). The same Gamma prior is also used for the intrinsic scatter term, although any weak prior would return the same result, because this term is well determined by the data.

We do not need to model the data collection selection: errors on mass are smaller than the correct mass' prior (mentioned in the previous section, a Gaussian of 0.3 dex), no selection is applied on richness (the probability of being included in the sample is richness-independent) and errors on richness are negligible compared to the variation of the richness' prior (richness function). Note however that if lower richnesses were used, as in Sect. 8.9, modeling the data structure would be compelling. Therefore, the population structure cannot be overlooked in general.

The whole model thus reads:

```
data
{
nu <-6
obsvarlgM <- pow(errlgM,2)
}
model
{
for (i in 1:length(obstot)) {
 obsbkg[i] ~ dpois(nbkg[i])
 obstot[i] ~ dpois(nbkg[i]/C[i]+n[i])
 n[i] ~ dunif(0,3000)
 nbkg[i] ~ dunif(0,3000)

 precy[i] ~ dgamma(1.0E-5,1.0E-5)
 obslgM[i] ~ dnorm(lgM[i],precy[i])
 obsvarlgM[i] ~ dgamma(0.5*nu,0.5*nu*precy[i])

 z[i] <- alpha+14.5+beta*(log(n[i])/2.30258-1.5)
 lgM[i] ~ dnorm(z[i], pow(intrscat,-2))
 }
intrscat <- 1/sqrt(prec.intrscat)
```

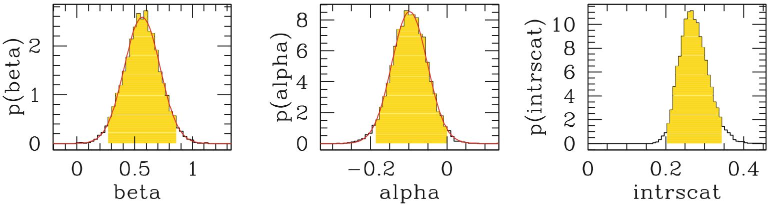

Fig. 8.33 Posterior probability distribution for the parameters of the richness-mass scaling. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation to it. The *shaded (yellow) range* shows the 95 % probability interval

```
prec.intrscat ~ dgamma(1.0E-5,1.0E-5)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
}.
```

Figure 8.32 shows the scaling between richness and mass, observed data, the mean scaling (solid line) and its 68 % uncertainty (shaded yellow region) and the mean intrinsic scatter (dashed lines) around the mean relation. The ±1 intrinsic scatter band is not expected to contain 68 % of the data points, because of the presence of measurement errors.

Using the model above, we found for our sample of 52 clusters:

$$\mathtt{lgM} = (0.57 \pm 0.15)\ (\log \mathtt{n} - 1.5) + 14.40 \pm 0.05\ . \qquad (8.11)$$

Figure 8.33 shows the posterior marginals for the key parameters: for the intercept, slope and intrinsic scatter σ_{scat}. These marginals are reasonably well approximated by Gaussians. The intrinsic mass scatter at a given richness, `intrscat`$=\sigma_{lgM|\log n}$, is small, 0.27 ± 0.03. The small scatter and its small uncertainty is promising from the point of view of using richness for estimating the cluster mass.

As mentioned, to predict the mass of a cluster we only need to list in the data file the values of `obstot`, `obsbkg` and `C`, the string 'NA' (not available) in place of the (not available) cluster mass (and any number in the mass error column). JAGS will compute `lgM` in the form of sampling for that cluster. Predicted masses will have errors whose size do account for the proxy error, the scatter between proxy and mass, and by the uncertainty of the mean relation at the richness of the considered cluster. Therefore, a cluster with a noisier richness, or a richness not well sampled by the clusters used to measure the mass-richness regression will have a larger error. For example, cluster # 38 in Table 8.7 has a larger error on the predicted mass than other clusters in the table because in the calibrating sample only a few clusters are as rich as it is. Cluster #9 also has large mass errors because of the extrapolation in going from the range where the richness-mass is well calibrated, from seven galaxies on, to its richness, about four galaxies.

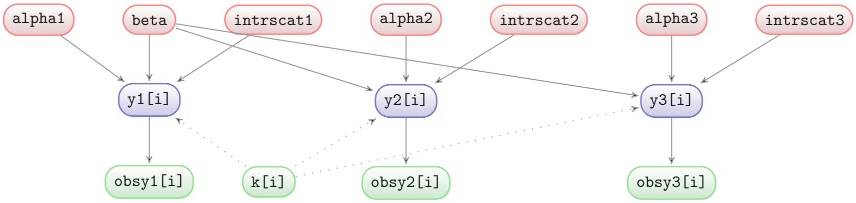

Fig. 8.34 Graph showing the relationship between quantities involved in the meta-analysis of Sect. 8.11

This example offers us the possibility of directly showing what we illustrated at the start of the chapter (Sect. 8.1.3): the fit of the inverse relation is not the inverse of the fit of the direct relation. We fitted twice the elements of Table 8.5, we fitted `n` as a function of `lgM` in Sect. 8.9, and the opposite here. Here we found a slope of 0.57 ± 0.15. There we found 0.47 ± 0.12 and, beyond any doubt $(0.57 \pm 0.15) \neq \frac{1}{0.47 \pm 0.12}$. As mentioned, there is no reason why two regressions should have reciprocal slopes when, as in this case, they have different goals (and priors), even if they are both based on the same data.

8.11 A Meta-Analysis: Combined Fit of Regressions with Different Intrinsic Scatter

Suppose that now you have to combine the results of two fits that share some similarities. For example, you may have fitted two data sets with models sharing only one parameter but not the others. How can one combine such different analyses? We now come back to the example of the physical constant taking perhaps different values in our laboratories and at the location of QSO 0347-383 (or at the time its light was emitted). This object has been independently observed for the very same reason by two other teams: Ubachs et al. (2007) and Thompson et al. (2009) and these data are listed in Table 8.6.

Figure 8.35 shows the data of these two teams, which are independently analyzed, using the model presented in Sect. 8.3 for the Wendt and Molaro (2011) data. The figures show the scaling between $\Delta\lambda/\lambda$ and `k`, the observed data, the mean scaling (solid line), its 68 % uncertainty (shaded yellow region), and the mean intrinsic scatter (dashed lines) around the mean relation. We found a negative slope with a large intrinsic scatter for the data set of Thompson et al. (2009), and a positive slope for the data set of Ubachs et al. (2007). The intrinsic scatter differs from data set to data set. This is not surprising because we are witnessing systematic biases (systematics for short) caused by the different instrument settings at the times of the observations.

We can enhance this analysis by combining the three data sets. The model is straightforward, we just need to write three regressions models, one per data set, with same slope `beta` (physical constants take the same values independently of

Table 8.6 Data for the physical constant determination, from Thompson et al. (2009) (upper part of the table) and Ubachs et al. (2007) (lower part)

k[i]	obsy2[i]	err.y2[i]	k[i]	obsy2[i]	err.y2[i]	k[i]	obsy2[i]	err.y2[i]
0.05147	0.0918	0.0082	0.00487	0.1047	0.0078	0.01997	0.1032	0.0050
0.04625	0.0944	0.0063	0.03753	0.0908	0.0027	0.01346	0.0985	0.0038
0.02149	0.1018	0.0068	0.03475	0.1061	0.0053	0.01051	0.1035	0.0048
0.02149	0.1061	0.0053	0.03408	0.0958	0.0034	0.01099	0.1004	0.0030
0.02097	0.0930	0.0041	0.03408	0.0808	0.0075	0.01001	0.0985	0.0041
0.04821	0.1018	0.0035	0.00525	0.0987	0.0070	0.00953	0.0949	0.0052
0.04772	0.0973	0.0063	0.00710	0.0997	0.0051	0.00719	0.0911	0.0029
0.01396	0.1068	0.0043	0.00710	0.0861	0.0110	0.00184	0.0975	0.0086
0.01272	0.1049	0.0068	0.03027	0.0980	0.0058	0.00115	0.1018	0.0088
0.03682	0.0994	0.0067	0.02460	0.0925	0.0036	0.00143	0.1018	0.0036
0.01088	0.1061	0.0032	0.02033	0.0911	0.0097	0.00259	0.1025	0.0025
0.00487	0.1049	0.0035	0.02064	0.1059	0.0063			
k[i]	obsy3[i]	err.y3[i]	k[i]	obsy3[i]	err.y3[i]	k[i]	obsy3[i]	err.y3[i]
0.04625	0.1051	0.0031	0.03689	0.1004	0.0012	0.01997	0.0975	0.0018
0.02149	0.1009	0.0025	0.03408	0.0994	0.0010	0.01647	0.1040	0.0026
0.02097	0.0975	0.0026	0.02849	0.0975	0.0020	0.01556	0.1020	0.0025
0.04821	0.0994	0.0027	0.00525	0.0956	0.0032	0.01346	0.0963	0.0040
0.04297	0.0992	0.0030	0.00710	0.0958	0.0026	0.01051	0.1013	0.0011
0.01396	0.0947	0.0015	0.03027	0.1006	0.0019	0.01099	0.0954	0.0021
0.01272	0.0951	0.0023	0.02454	0.0963	0.0069	0.01001	0.0975	0.0016
0.03682	0.0968	0.0023	0.02324	0.0958	0.0024	0.00953	0.1016	0.0027
0.01088	0.0987	0.0018	0.02214	0.0989	0.0066	0.00719	0.0968	0.0024
0.04005	0.1013	0.0019	0.02033	0.1040	0.0017	0.00504	0.0975	0.0025
0.00368	0.1009	0.0038	0.02149	0.0963	0.0021	0.00115	0.0954	0.0027
0.03753	0.1004	0.0019	0.02064	0.1044	0.0016	0.00143	0.0939	0.0016
0.00259	0.0992	0.0030						

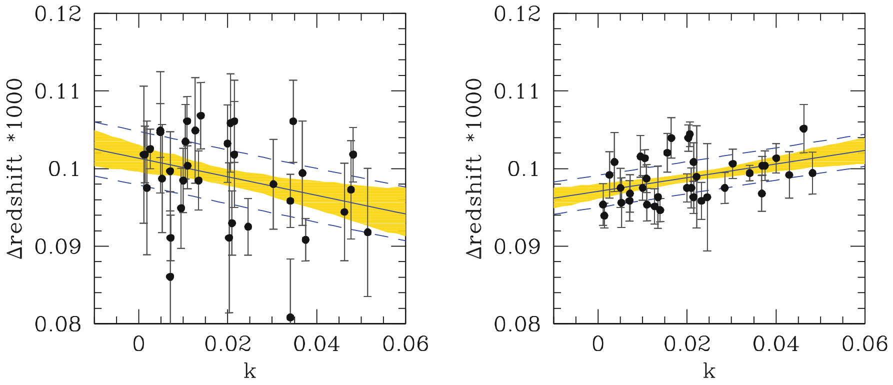

Fig. 8.35 Scaling between $\Delta\lambda/\lambda$ and k, the observed data, the mean scaling (*solid line*), its 68 % uncertainty (*shaded yellow region*), and the mean intrinsic scatter (*dashed lines*) around the mean relation. Data from Thompson et al. (2009) (*left panel*) and Ubachs et al. (2007) (*right panel*). The distance between the data and the regression line is due in part to the data's observational error and its intrinsic scatter

which instrument acquires the data) and different intrinsic scatters, `intrscat[1]`, `intrscat[2]`, `intrscat[3]`, these account for different systematics in the three experiments. We allow the intercepts, `alpha[1]`, `alpha[2]`, and

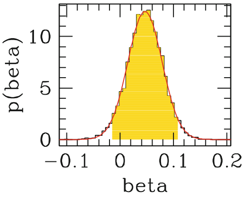

Fig. 8.36 Posterior distribution of the slope β of the meta-analysis of the varying constant problem

alpha[3], to be possibly different, in order to allow possible systematic shifts between calibrations of the three used instruments. The graph is illustrated in Fig. 8.34.

Here is the model where most of the lines are repeated three times, one per experiment (a more compact form exists):

```
model {
intrscat[1] ~ dunif(0,3)
intrscat[2] ~ dunif(0,3)
intrscat[3] ~ dunif(0,3)

alpha[1] ~ dnorm(0.0,1.0E-4)
alpha[2] ~ dnorm(0.0,1.0E-4)
alpha[3] ~ dnorm(0.0,1.0E-4)

beta ~ dt(0,1,1)
for (i in 1:length(x1)) {
 # modeling y
 obsy1[i] ~ dnorm(y1[i],pow(err.obsy1[i],-2))
 y1[i] ~ dnorm(z1[i],pow(intrscat[1],-2))
 # modeling y-x
 z1[i] <- alpha[1]+0.1+beta*(x1[i]-0.03)
}
for (i in 1:length(x2)) {
 # modeling y
 obsy2[i] ~ dnorm(y2[i],pow(err.obsy2[i],-2))
 y2[i] ~ dnorm(z2[i],pow(intrscat[2],-2))
 # modeling y-x
 z2[i] <- alpha[2]+0.1+beta*(x2[i]-0.03)
}
for (i in 1:length(x3)) {
 # modeling y
 obsy3[i] ~ dnorm(y3[i],pow(err.obsy3[i],-2))
 y3[i] ~ dnorm(z3[i],pow(intrscat[3],-2))
 # modeling y-x
 z3[i] <- alpha[3]+0.1+beta*(x3[i]-0.03)
}
}.
```

In the combined analysis of the three data sets, we found a slope of $4.6 \pm 3.2\ 10^{-5}$ when converted in the usual astronomical units. Figure 8.36 shows that $\beta = 0$ has a probability of about 1/6th of the most probable slope, insufficient to claim that physical constants are variable. As mentioned in previous sections, the model with a variable β has an advantage: it has one more degree of freedom, that can be adjusted to increase the quality of the fit. Its better performances (larger value of the maximal posterior) are therefore partially due to the larger flexibility of the model. To summarize, the data carry little or no evidence for varying physical constants.

8.12 Advanced Analysis

8.12.1 Cosmological Parameters from SNIa[7]

Supernovae (SNIa) are very bright objects with very similar luminosities. The luminosities spread can be made even smaller by accounting for the correlation with the supernova's color and stretch parameter (the latter is a measurement of how slowly SNIa fade), as illustrated in Fig. 8.37 for the sample in Kessler et al. (2009). These features make SNIa very useful for cosmology: they can be observed at large distances and the relation between flux (basically the rate of photons received) and luminosity (the rate of photons emitted) is modulated by the luminosity distance (to the square), which in turn is a function of the cosmological parameters. Therefore, measuring SNIa fluxes (and redshift) allows us to put constraints on cosmological parameters. The only minor complication is that the SNIa luminosity is a function of its color and stretch parameters, and these parameters have an intrinsic scatter too, which in turn has to be determined from the data at the same time as the other parameters.

March et al. (2011) show that the Bayesian approach, also followed in this section, delivers tighter statistical constraints on the cosmological parameters over 90 % of the time, that it reduces the statistical bias typically by a factor $\sim 2-3$, and that it has better coverage properties than the usual chi-squared approaches.

We observe SNIa magnitudes `obsm[i]` $(= -2.5log(flux)+c)$ with Gaussian errors `errmag[i]`):

```
obsm[i] ~ dnorm(m[i],pow(errmag[i],-2)) .
```

These `m[i]` are related to the distance modulus `distmod[i]`, with a Gaussian intrinsic scatter `intrscatM` via

```
m[i] ~ dnorm(Mm+distmod[i]- alpha* x[i] + beta*c[i],
             pow(intrscatM,-2)),
```

where `Mm` is the (unknown) mean absolute magnitude of SNIa, and `alpha` and `beta` allows one to reduce the SNIa luminosity scatter by accounting for the correlation with the stretch and color parameters.

Similarly to March et al. (2011), the `Mm`, `alpha`, `beta`, and log `intrscatM` priors are taken to be uniform across a wide range:

[7] This section has been drawn from Andreon (2012b).

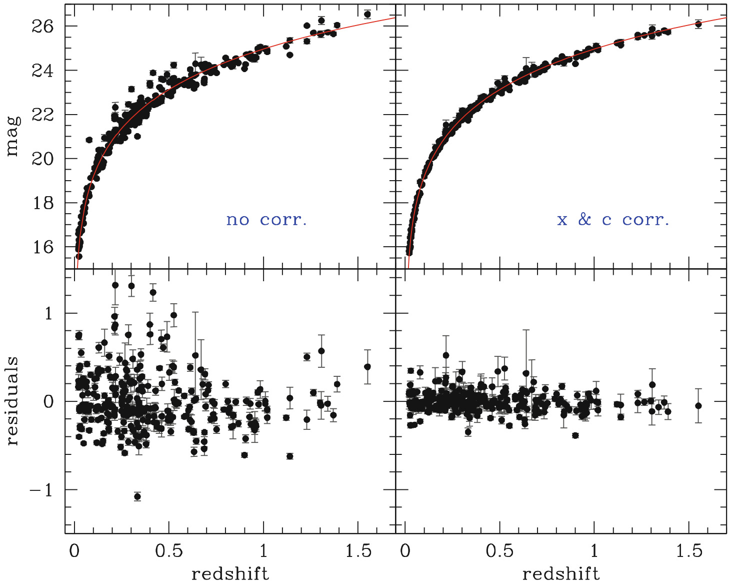

Fig. 8.37 Apparent magnitude vs. redshift of the SNIa sample (*upper panels*), and their residuals from a $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$ cosmological model (*bottom panels*) before (*left panels*) and after (*right panels*) correcting for stretching and the color parameter. Reproduced from Andreon (2012b) with permission

```
Mm~ dunif(-20.3, -18.3)
alpha ~ dunif(-2,2.0)
beta ~ dunif(-4,4.0)
intrscatM <- pow(10,lgintrscatM)
lgintrscatM ~ dunif(-3,0) .
```

x[i] and c[i] are the true value of the stretch and color parameters, respectively, of which we observe obsx[i] and obsc[i] with Gaussian errors errobsx[i] and errobsc[i]

```
obsc[i] ~ dnorm(c[i], pow(errobsc[i],-2))
obsx[i] ~ dnorm(x[i], pow(errobsx[i],-2)) .
```

The key point of the modeling is that the obsx[i] and obsc[i] values scatter more than their errors, but not immensely so, see Fig. 8.38. Therefore, the x[i] and c[i] distributions are themselves non-uniform. This non-uniformity induces a Malmquist-like bias (large |obsx[i]| are likely low |x[i]| values scattered to large values, more than the other way around because of the larger abundance of low |x[i]| values). Therefore, we model, as March et al. (2011) do, the individual x[i] and c[i] as drawn from independent normal distributions centered on xm and cm with standard deviation intrscatx and intrscatR. In formulae:

```
c[i] ~ dnorm(cm,pow(intrscatC,-2))
x[i] ~ dnorm(xm,pow(intrscatx,-2)) .
```

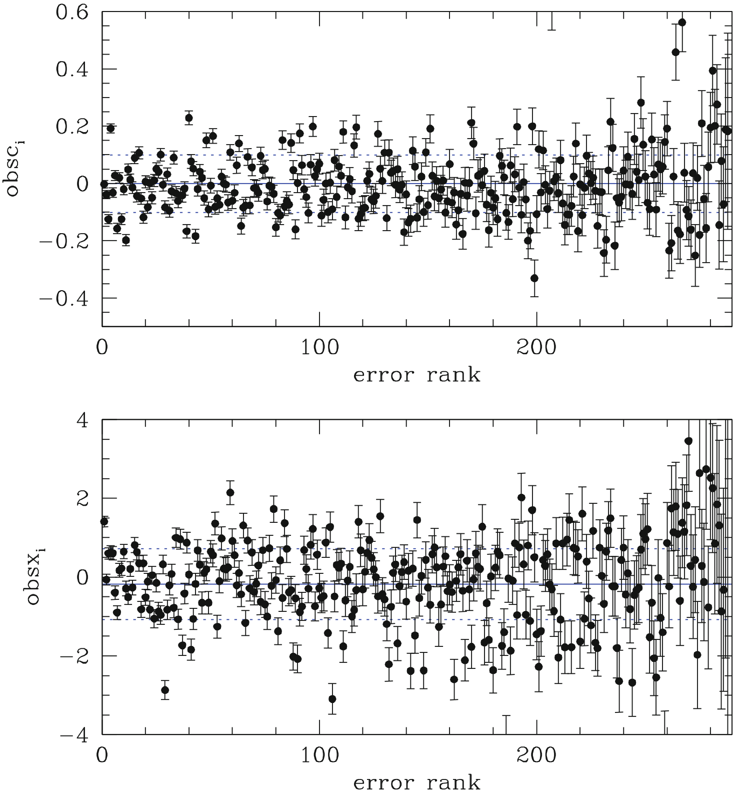

Fig. 8.38 Observed values of the stretch parameters, `obsx[i]`, and of the color parameter, `obsc[i]` ranked by error. Points scatter more than the error bars (see the left side of the figure). The *dashed lines* indicate the size of the intrinsic scatter as determined by our analysis. Reproduced from Andreon (2012b) with permission

We take uniform priors for `xm` and `cm`, and uniform priors on log `intrscatx` and on log `intrscatC`, between the indicated boundaries:

```
cm ~ dunif(-3,3)
xm ~ dunif(-10,10)
intrscatx <- pow(10,lgintrscatx)
lgintrscatx ~ dunif(-5,2)
intrscatC <- pow(10,lgintrscatC)
lgintrscatC ~ dunif(-5,2) .
```

To complete the model, we need to remember the definition of distance modulus:

```
distmod[i] <- 25 + 5/2.3026 * log(dl[i]) -5/2.3026*
              log(H0/300000) ,
```

where the luminosity distance, `dl` is a complicated expression, involving integrals, of the redshift `z[i]` and the cosmological parameters $\Omega_\Lambda, \Omega_M, w, H_0$ (see any recent cosmology textbook for the mathematical expression), named in the JAGS code as `omegal`, `omegam`, `w`, and `H0`.

Redshift, in the considered sample, has a heteroscedastic (i.e., non-constant variance) Gaussian error `errz[i]`):

```
obsz[i] ~ dnorm(z[i],pow(errz[i],-2)) .
```

The redshift's prior is taken to be uniform,

```
z[i] ~ dunif(0,2).
```

Supernovae alone do not allow one to determine all cosmological parameters, so we need an external prior on them, notably on H_0, taken from Freedman et al. (2001) to be (again this is the beauty of Bayesian methods, it gives us a framework to incorporate accumulated scientific knowledge)

```
H0 ~ dnorm(72,pow(8,-2)).
```

At this point, we may decide which sets of cosmological models we want to investigate using SNIa, for example a flat universe with a possible $w \neq 0$ with the following priors:

```
omegam~dunif(0,1)
omegak <-0
w ~ dunif(-4,0) ,
```

or a curved universe with $w = -1$:

```
omegam~dunif(0,1)
omegak~dunif(-1,1)
w <- -1 ,
```

or any other set of parameters.

Both considered cosmologies have:

$$\Omega_k = 1 - \Omega_m - \Omega_\Lambda \tag{8.12}$$

i.e.,

```
omegal<-1-omegam-omegak .
```

Finally, one may want to use some data. As shortly mentioned, we use[8] the compilation of 288 SNIa in Kessler et al. (2009).

Most of the distributions above are Gaussian and the posterior distribution can be almost completely analytically computed (March et al. 2011). However, numerical evaluations of the stochastic part of the model on an (obsolete) laptop take about one minute, therefore there is no need for speed up. Instead, the evaluation of the luminosity distance is CPU intensive (it takes $\approx 10^3$ more time, unless approximate analytic formulae for the luminosity distance are used), because an integral has to

[8] The table is given in electronic format at the link http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/SNdata.dat.

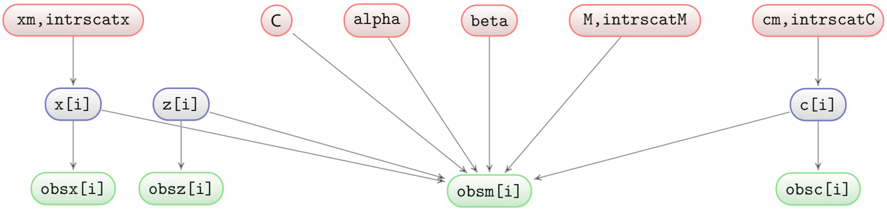

Fig. 8.39 Graph showing the relationship between quantities involved in the determination of cosmological parameters using SN1A

be evaluated a number of times equal to the product of the number of supernovae and the number of target posterior samplings, i.e., about four millions times in our numerical computation. The JAGS implementation of the luminosity distance integral is implemented in the code below as a sum over a tightly packed grid on redshift values.

To summarize, the fitted model, whose graph is shown in Fig. 8.39, reads:

```
data {
# grid for distance modulus integral evaluation
 for (k in 1:1500){
  grid.z[k] <- (k-0.5)/1000.
 }
 step.grid.z <-grid.z[2]-grid.z[1]
}
model {
for (i in 1:length(obsz)) {
obsm[i] ~ dnorm(m[i],pow(errmag[i],-2))
m[i] ~ dnorm(Mm+distmod[i]- alpha* x[i] + beta*c[i],
       pow(intrscatM,-2))
obsc[i] ~ dnorm(c[i], pow(errobsc[i],-2))
c[i] ~ dnorm(cm,pow(intrscatC,-2))
obsx[i] ~ dnorm(x[i], pow(errobsx[i],-2))
x[i] ~ dnorm(xm,pow(intrscatx,-2))
# distmod definition & H0 term
distmod[i] <- 25 + 5/2.3026 * log(dl[i]) -5/2.3026* log(H0/300000)
z[i] ~ dunif(0,2)
obsz[i] ~ dnorm(z[i],pow(errz[i],-2))
######### dl computation (slow and tedious)
tmp2[i] <- sum(step(z[i]-grid.z) * (1+w) / (1+grid.z)) *
               step.grid.z
omegade[i] <- omegal * exp(3 * tmp2[i])
xx[i] <- sum(pow((1+grid.z)^3*omegam + omegade[i] +
         (1+grid.z)^2*omegak,-0.5)*
         *step.grid.z * step(z[i]-grid.z))
# implementing if, to avoid diving by 0 added 1e-7 to omegak
zz[1,i] <- sin(xx[i]*sqrt(abs(omegak))) *
           (1+z[i])/sqrt(abs(omegak+1e-7))
zz[2,i] <- xx[i] * (1+z[i])
zz[3,i] <- (exp(xx[i]*sqrt(abs(omegak)))-exp(-xx[i]*
            sqrt(abs(omegak))))/2 *
            *(1+z[i])/sqrt(abs(omegak+1e-7))
```

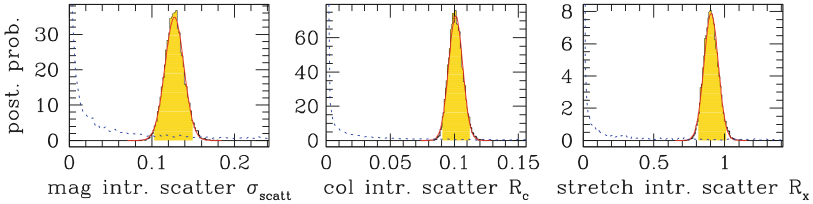

Fig. 8.40 Prior and posterior probability distribution for the three intrinsic scatter terms in the SNIa problem. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation to it. The *shaded (yellow) range* shows the 95 % probability interval. The adopted priors are indicated by the *blue dotted curve*. Reproduced from Andreon (2012b) with permission

```
dl[i] <- zz[b,i]
}
b <- 1 + (omegak==0) + 2*(omegak > 0)
########## end dl computation

# priors
Mm~ dunif(-20.3, -18.3)
alpha ~ dunif(-2,2.0)
beta ~ dunif(-4,4.0)
cm ~ dunif(-3,3)
xm ~ dunif(-10,10)
# uniform prior on logged quantities
intrscatM <- pow(10,lgintrscatM)
lgintrscatM ~ dunif(-3,0)
intrscatx <- pow(10,lgintrscatx)
lgintrscatx ~ dunif(-5,2)
intrscatC <- pow(10,lgintrscatC)
lgintrscatC ~ dunif(-5,2)
#cosmo priors
H0 ~ dnorm(72,pow(8,-2))
omegal<-1-omegam-omegak
# cosmo priors 1st set LCDM
#omegam~dunif(0,1)
#omegak~dunif(-1,1)
#w <- -1
# cosmo priors 2nd set: wCDM
omegam~dunif(0,1)
omegak <-0
w ~ dunif(-4,0)
}.
```

Figure 8.40 shows the prior (dashed-blue line) and posterior (histogram) probability distributions for the three intrinsic scatter terms present in the cosmological parameter estimation: the scatter in absolute luminosity after color and stretch corrections, (`intrscatM`), and the intrinsic scatter in the distribution of the color and stretch terms (`intrscatx` and `intrscatC`). The (posterior) probability at intrinsic scatters near zero is approximately zero and thus the three intrinsic scatter terms are necessary parameters for the modeling of SNIa, and not useless complications.

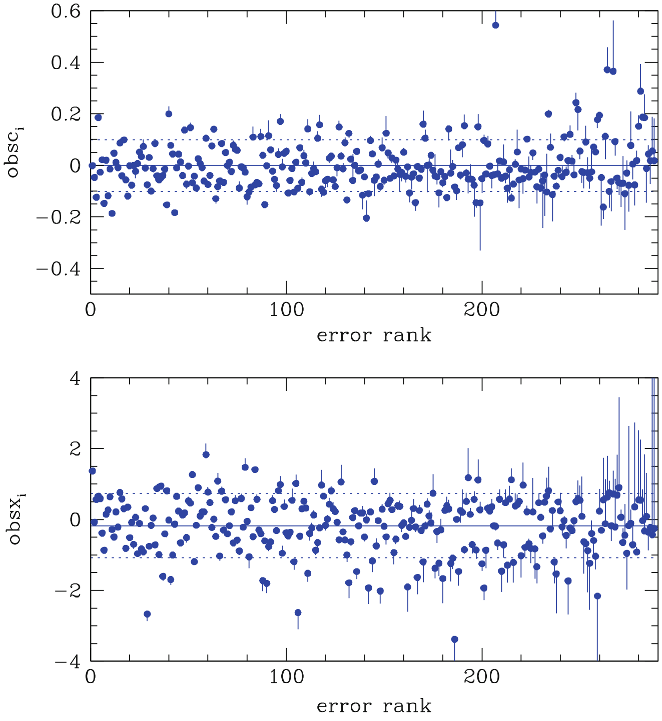

Fig. 8.41 Effect of the population structure for the stretch and color parameters. Each tick goes from the observed value to the posterior mean. The population modeling attempts to counterbalance the increased spread (Malmquist-like), especially those with larger error (on the right, in the figures), pulling values toward the mean. Reproduced from Andreon (2012b) with permission

The three posteriors are dominated by the data, the prior is quite flat in the range where the posterior is appreciably not zero (Fig. 8.40). Therefore, any other choice for the prior, as long as it is smooth and shallow over the shown parameter range, would have returned indistinguishable results.

Not only do SNIa have luminosities that depend on color and stretch terms, but these in turn have their own probability distribution (taken to be Gaussian for simplicity) with a well-determined width. Figure 8.41 depicts the Malmquist-like bias one should induce if the spread of the distribution of color and stretch parameters is ignored: it reports the observed values (as in Fig. 8.38), `obsx[i]` and `obsc[i]` as well as the posterior means of the true values `x[i]` and `c[i]`. The effect of their modeling is to pull values toward the mean, and more so those with large errors, to compensate for the systematic shift (Malmquist-like bias) towards larger observed values.

Figure 8.42 shows the probability distribution of the two color and stretch slopes: `alpha`$= 0.12 \pm 0.02$ and `beta`$= 2.70 \pm 0.14$, respectively. As for the intrinsic scatter terms, the posterior distribution is dominated by the data and therefore any other prior, which should be smooth and shallow where the likelihood is greater than zero, would have returned indistinguishable results.

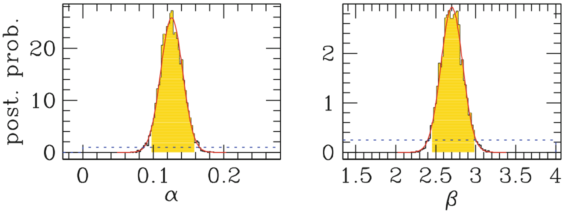

Fig. 8.42 Prior and posterior probability distribution for the color and stretch slopes of the SNIa problem. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation of the posterior. The *shaded (yellow) range* shows the 95 % probability interval. The adopted (uniform) priors are indicated by the *blue dotted curve*. Reproduced from Andreon (2012b) with permission

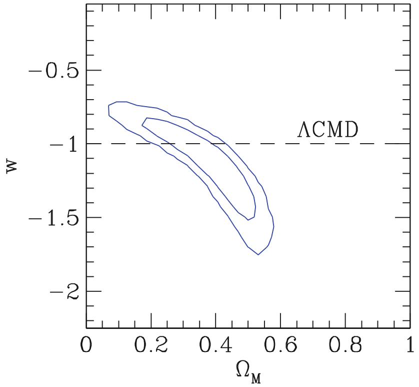

Fig. 8.43 Constraints on the cosmological parameters Ω_M and w. The two contours delimit 68 % and 95 % constraints. Reproduced from Andreon (2012b) with permission

Finally, Fig. 8.43 reports perhaps the astronomically most interesting result: contours of equal probability for the cosmological parameters Ω_M and w. The banana shape of the contours is due to the known correlation between these parameters for the considered experiment. For one dimensional marginals, we found: omegam= $\Omega_M = 0.40 \pm 0.10$ and w= $w = -1.2 \pm 0.2$, but with non-Gaussian probability distributions.

8.12.2 The Enrichment History of the ICM[9]

Measuring the evolution of the metallicity of the gas that fills the potential well of galaxy clusters is key information in understanding when metals are produced and to know the time of stellar feedback on the intracluster medium (i.e., on the gas).

[9] The material of this section has been drawn from Andreon (2012a).

The determination of the evolution of intracluster medium metal content is made difficult by

a) upper limits,
b) the low signal-to-noise ratio of the abundance measurements,
c) possible selection effects,
d) boundaries in the parameter space,
e) non-Gaussian errors,
f) the intrinsic variety of the objects studied,
e) abundance systematics.

We simultaneously address all these issues, thus allowing cross-talk (covariance), following Andreon (2012a). The data contains information on 130 clusters, taken from Andreon (2012a),[10] that combines data observed with different instruments (Chandra and XMM), a point which requires accounting for. Furthermore, the studied sample, being a heterogeneous collection without any known selection function, requires special attention.

8.12.2.1 Enrichment History

Abundance by definition is a positive quantity. Therefore, Andreon (2012a) choose to fit abundance measurements by a function that can never cross the physical boundary $Ab = 0$. It also seems plausible that the Fe abundance at the time of the big bang is zero. Therefore, following Andreon (2012a), we adopt for the enrichment history an exponential function, with characteristic time `tau`, parametrized as:

```
ft[i] <- Abz02*(1-exp(-t[i]/tau))/(1-exp(-11/tau)) ,
```

where `t[i]` is the Universe's age at the redshift of cluster i. The denominator has been parameterized to have as its second parameter the Fe abundance at $z = 0.2$, `f[i]`=`Abz02`, at a redshift well sampled by the data used. This choice simplifies the interpretation of the results. The τ parameter regulates when enrichment occurs: early in the Universe's history (small `tau`'s) or gradually over time (large `tau`'s). More complex histories may be taken in the future when observations will reach the epoch of first enrichment, for example one with two characteristic times to account for an initial enrichment by core collapse supernovae followed by a metal production spread on longer time scales.

8.12.2.2 Intrinsic Scatter

Galaxy clusters show a spread of Fe abundance. The presence of an intrinsic scatter implies that the information content of a single measurement is lower than indicated by the error, especially when the latter is comparable to, or is smaller than,

[10] The table in electronic format is available at the link http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/Andreon12_abundance.dat.

the intrinsic scatter. The intrinsic scatter acts as a floor: the information content of a measurement is not better than the intrinsic scatter, no matter how precise the measurement is. Following Andreon (2012a) we model the distribution of Fe abundances as a log-normal process of unknown intrinsic scatter,

```
Ab[i] ~ dlnorm(log(ft[i]), pow(intrscat,-2)) .
```

Expressed in words, the Fe abundance of cluster *i*, `Ab[i]` shows a log-normal intrinsic scatter `intrscat`, around the median value, `ft[i]`. Of course, a Gaussian scatter is precluded by the positive nature of the Fe abundance.

8.12.2.3 Controlling for Temperature T

The Fe abundance might depend on cluster mass. If neglected, this dependence induces a bias in determining the evolution of the Fe abundance unless the studied sample is a random, redshift-independent sampling of the cluster's mass function. For example, if the average mass of clusters in the sample increases with redshift and the Fe abundance increases with temperature T, one may observe a spurious Fe abundance tilt (increase) with redshift. Other combinations of dependences are potential sources of a bias, such as a decreasing metal abundance with increasing T. Among these combinations, we should also consider those that include variations in the mass range sampled at a given redshift (e.g., lower redshifts sampling a wider cluster mass range). To summarize, given the uncontrolled nature of the available samples, one must at the very least control for T (i.e., a mass proxy) in order to avoid the risk of mistaking a mass dependency for an Fe abundance evolution. Even if data are unable to unambiguously determine a T trend, controlling for T allows a trend to be there as much as allowed by the data and not to overstate the statistical significance of a redshift trend.

Following Andreon (2012a) we control for mass, by allowing the Fe abundance, `Ab`, to depend on `T` with a power-law dependence.

8.12.2.4 Abundances Systematics

Metal abundances may show some systematic differences when derived by two different teams using data taken with two different X-ray telescopes (Chandra and XMM) analyzed with similar, but not identical, procedures. In particular, Andreon (2012a) found for 13 common clusters that XMM abundances are 0.77 ± 0.065 times those measured with Chandra, i.e.,

```
factor~dnorm(0.77-1,pow(0.065,-2)) .
```

We account for these systematics by allowing metal abundances measured by different telescopes to differ by a multiplicative factor as large as allowed by the data. Observationally, the factor is constrained by measurements of both telescopes having to agree after the multiplicative scaling. Expressed mathematically, we only

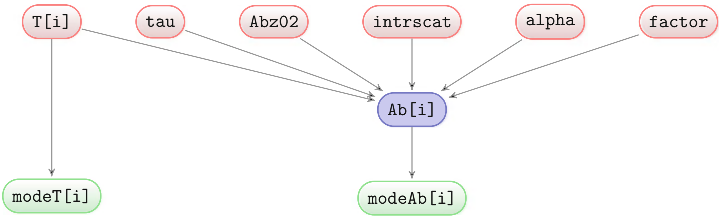

Fig. 8.44 Graph showing the relationship between quantities involved in the enrichment history of the ICM determination

need to introduce a quantity, `tid`, that takes the value of zero for the Chandra data, and one for the XMM data (a convention, but one may choose to do the reverse), and multiply metal abundances by the factor, $1 + \texttt{factor} * \texttt{tid}$, to bring all measurements on a common scale (Chandra, with our convention). For controlling for temperature and abundance systematic we just need to write:

```
Abcor[i] <-Ab[i]*pow(T[i]/5,alpha)*(1+factor*tid[i]) .
```

As usual, we center values near the data's average (5 keV).

8.12.2.5 T and Fe Abundance Likelihood

X-ray temperature is modeled as log-normal (see Andreon 2012a for data that support this statement):

```
modeT[i] ~ dlnorm(log(T[i]),pow(sigmaT[i],-2)).
```

Following Andreon (2012a), abundance has a Gaussian likelihood:

```
modeAb[i] ~ dnorm(Ab[i],pow(sigmaAb[i],-2)) .
```

Note that astronomers do not report `modeT[i]` and `sigmaT[i]` in their papers (and similarly for abundance), but other numbers closely related to these (see Andreon 2012a for details).

The sample includes a few upper limits to the metal abundance. These data do not require any special treatment, their likelihood is a Gaussian centered at (or below) zero. Therefore, upper limits to the abundance are automatically accounted for.

8.12.2.6 Priors

We assume we know little about the parameters, i.e., we adopt weak priors, that assign zero mass for impossible values of the parameters. We adopt almost uniform priors for all the parameters over a wide range of values, including those of the observed values, i.e., we do not prefer a parameter value over another one a priori. Following Andreon (2012a) we take:

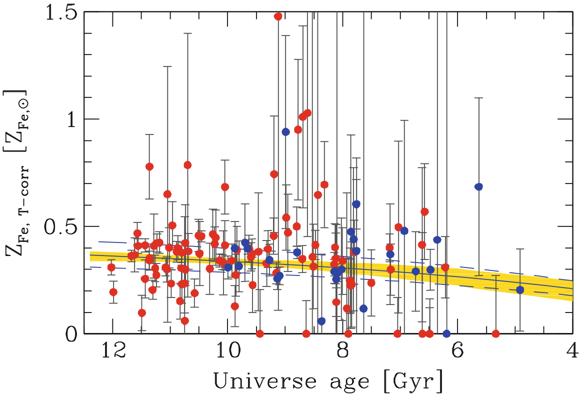

Fig. 8.45 Metal abundance, on the Chandra scale, vs. Universe age. Observed values of metal abundance and errors are corrected for the *T* dependence and for the Chandra vs. XMM systematic as determined by a simultaneous fit of all parameters. *Red (blue) circles* refer to Chandra (XMM) measurements. Points indicate the maximum a posteriori, whereas error bars represent the 68 % probability intervals. The *solid line* marks the mean fitted relation between metal abundance and redshift, while the *dashed line* shows this mean plus or minus the intrinsic scatter `intrscat`. The *shaded region* marks the 68 % probability interval for the regression. Reproduced from Andreon (2012a) with permission

```
Abz02~dunif(0,1)
tau ~ dunif(1,100)
alpha ~dt(0,1,1)
intrscat ~ dunif(0,1)
T[i] ~ dunif(1,20) .
```

To sum up, the fitted model, whose graph is shown in Fig. 8.44 is:

```
model {
for (i in 1:length(modeT)) {
 ft[i] <- Abz02*(1-exp(-t[i]/tau))/(1-exp(-11/tau))
 Ab[i] ~ dlnorm(log(ft[i]), pow(intrscat,-2))
 Abcor[i] <-Ab[i]*pow(T[i]/5,alpha)*(1+factor*tid[i])
 modeAb[i] ~ dnorm(Abcor[i],pow(sigmaAb[i],-2))
 T[i] ~ dunif(1,20)
 modeT[i] ~ dlnorm(log(T[i]),pow(sigmaT[i],-2))
 # for p-value computation
 Ab.rep[i] ~ dlnorm(log(ft[i]), pow(intrscat,-2))
 Abcor.rep[i] <-Ab.rep[i]*pow(T[i]/5,alpha)*(1+factor*tid[i])
 modeAb.rep[i] ~ dnorm(Abcor.rep[i],pow(sigmaAb[i],-2))
}
Abz02~dunif(0,1)
tau ~ dunif(1,100)
alpha ~dt(0,1,1)
intrscat ~ dunif(0,1)
factor~dnorm(0.77-1,pow(0.065,-2))
}.
```

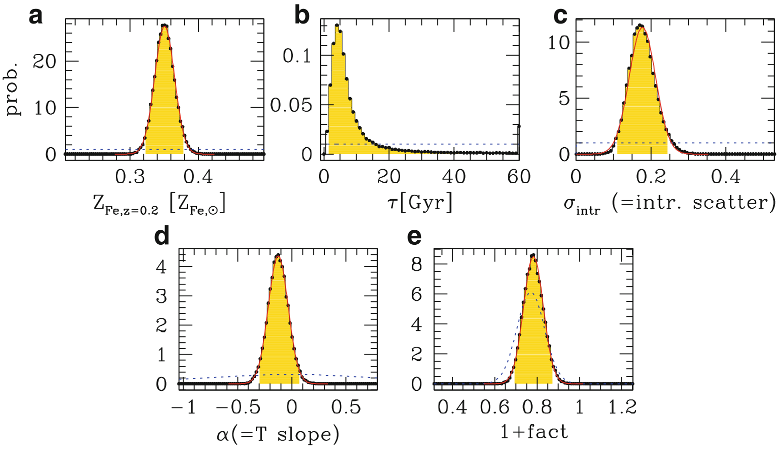

Fig. 8.46 Probability distribution for the parameters of the metal abundance vs. redshift fit. The *solid circles* show the posterior probability distribution as computed by MCMC, marginalized over the other parameters. The *red curve* (when present) shows a Gaussian approximation to the posterior. The *dashed curve* displays the adopted prior. The *shaded (yellow) range* shows the 95 % probability interval. Reproduced from Andreon (2012a) with permission

8.12.2.7 Results

The result of the fit of abundance and temperature values for the sample of 130 clusters is summarized in Figs. 8.45 and 8.46.

Figure 8.45 shows the data, corrected for the T dependence and for the abundance systematics, as determined by a simultaneous fit of all parameters. The solid line marks the mean fitted relation between abundance and Universe age, while the dashed line shows this mean plus or minus the intrinsic scatter `intrscat`, that turns out to be 0.18 ± 0.03, i.e., 18 % of the Fe abundance value. The shaded region marks the 68 % probability interval of the fit.

Figure 8.46 shows how our knowledge about the parameters changes from before (prior, dotted line) to after (posterior, solid line) observing and incorporating the data. The posterior distribution of all parameters is much more concentrated than the prior, i.e., data are highly informative about these parameters and conclusions on these parameters are not prior dependent.

A final note of caution: while the analysis accounts for T (mass) dependent selection effects, there could be mass-independent selection effects. The latter are not accounted for by the analysis.

8.12.3 The Enrichment History After Binning by Redshift

Some researchers are not happy with adopting a parametric estimate of the studied quantities, and prefer "to let the data speak for themselves," which typically involves binning the data to reduce the noise until a trend appears. This situation is quite common in the context of measuring the enrichment history. In fact, in the analysis of the previous section some authors might prefer not to impose a specific time dependency, and prefer to just inspect the enrichment history found by binning cluster data in redshift bins. This can be done easily using JAGS, it is just a matter of removing the code line describing the time dependency and allowing metal abundances in different redshift bins to be independent of each other. The logical link between intervening quantities becomes

```
 modeAb[i] ~ dnorm(Ab[i],pow(sigmaAb[i],-2)) T(0,)
 modeT[i] ~ dlnorm(log(T[i]),pow(sigmaT[i],-2))
 Ab[i] ~ dlnorm(log(meanAb), pow(intrscat,-2))
 Abcor[i] <- Ab[i]*pow(T[i]/5,alpha)*(1+factor*tid[i])
 T[i] ~ dunif(1,20) ,
```

to which we have to add the prior for the newly introduced quantity, the mean abundance in the redshift bin, `meanAb` taken to be a priori in a large range with no preference of a value over the other:

```
meanAb ~ dunif(0.,1) .
```

For the other parameters we adopt the posteriors determined in the previous section as priors. These are, as shown in the previous section, well described by Gaussian distributions:

```
alpha ~dnorm(-0.12,pow(0.09,-2))
intrscat ~ dnorm(0.18,pow(0.03,-2))
factor~dnorm(-0.22,pow(0.045,-2))I(-1,) ,
```

The whole model therefore reads:

```
model {
for (i in 1:length(modeT)) {
 modeAb[i] ~ dnorm(Abcor[i],pow(sigmaAb[i],-2))
 modeT[i] ~ dlnorm(log(T[i]),pow(sigmaT[i],-2))
 Ab[i] ~ dlnorm(log(meanAb), pow(intrscat,-2))
 Abcor[i] <- Ab[i]*pow(T[i]/5,alpha)*(1+factor*tid[i])
 T[i] ~ dunif(1,20)
}
meanAb ~ dunif(0.,1)
alpha ~dnorm(-0.12,pow(0.09,-2))
intrscat ~ dnorm(0.18,pow(0.03,-2))
factor~dnorm(-0.22,pow(0.045,-2))I(-1,)
}.
```

We emphasize that, strictly speaking, this analysis is using the data twice: once to derive the posterior of `alpha`, `intrscat`, and `factor` (in previous section), and

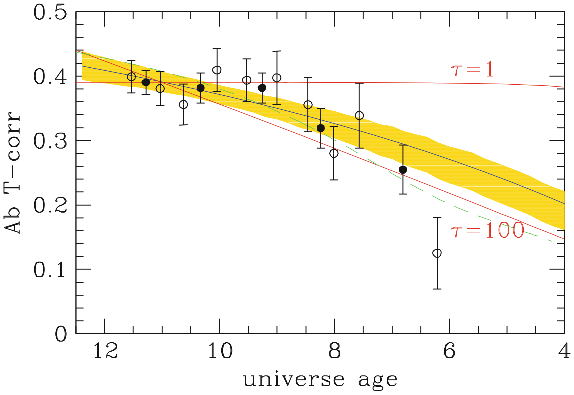

Fig. 8.47 Metal abundance, on the Chandra scale, vs. Universe age, where the individual abundances are replaced by binned values (*solid/open points* refer to 5/10 redshift bins). The *solid line and shading* are as in Fig. 8.46. Two extreme enrichment histories (*red lines*) are also plotted. Reproduced from Andreon (2012a) with permission

once to infer the meanAb (in this section). This double use of the data is conceptually wrong, deprecate, and in general returns underestimated errors. Practically, the information on the alpha, intrscat, and factor, derived in the previous section, is almost independent of meanAb derived here, and therefore errors are very close to the correct value. Readers unsatisfied by this two-step approach should rely only on the analysis given in the previous section.

Figure 8.47 shows the result of this binning exercise, after distributing clusters in 5 (solid dots) or 10 (open dots) redshift bins of equal cardinality. The Fe abundance stays constant for a long time, 4 Gyr or so, after which it decreases, in agreement with the rigorous derivation of the previous section.

8.12.4 With An Over-Poissons Spread

In this section we are interested in knowing how much the X-ray luminosities of clusters of a given richness may differ, whether X-ray luminosity and richness are tightly or only broadly related to each other. The tightness of the relation has implications on the cluster's astrophysics (the tighter it is, the more homogeneous the clusters are and their formation histories), and on the X-ray telescope cost of following-up cluster candidates of a given richness (the larger the spread, more time should be spent on each target to be sure to acquire a minimal number of photons, or a given signal to noise).

We now want to find the trend between two quantities we already investigated: richness, which is given by the difference of two Poisson quantities (cluster and field counts, already addressed several times in this book), and X-ray luminosity, which, in first approximation is also given by the difference of two Poisson quantities

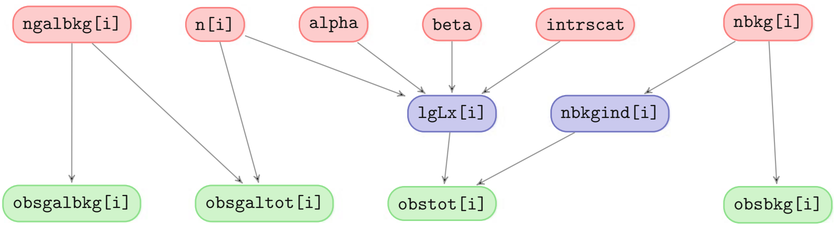

Fig. 8.48 Graph showing the relationship between quantities involved in the X-ray luminosity-richness scaling

(the X-ray flux in the cluster and field directions). There is an additional complication, however: background counts fluctuate more than allowed by a Poisson distribution, an issue modeled in Sect. 6.3.1.

The current regression is, therefore, a merging of the two mentioned models. Its graph is depicted in Fig. 8.48. This high modularity is one of the advantages of the Bayesian approach: once the full problem is broken down into smaller parts and these parts are modeled, the full problem is largely solved too.

As mentioned, richness is modeled as we have done many times (here we change variable names to avoid any confusion):

```
obsgalbkg[i] ~ dpois(ngalbkg[i])
obsgaltot[i] ~ dpois(ngalbkg[i]/Cgal[i]+n[i])
n[i] ~ dunif(1,3000)
ngalbkg[i] ~ dunif(0,3000) .
```

X-ray luminosity is modeled accounting for the over-Poisson background fluctuations as in Sect. 6.3.1:

```
# modelling X-ray photons
obstot[i] ~ dpois(nclus[i]+nbkgind[i]/nbox[i])
nbkgind[i] ~ dlnorm(log(nbkg[i]),1/0.2/0.2)
obsbkg[i] ~ dpois(nbkg[i])
nbkg[i] ~ dunif(0,10000)
# convert nclus in Lx
nclus[i] <- exp(2.30258*(lgLx[i]-C[i])) ,
```

where the last line is the usual conversion from $\log_{10}$ X-ray luminosities into counts.

To complete the model, we only need to write the link between the abscissa and ordinate (the linear regression), and specify the prior of the newly inserted parameters (slope, intercept, and intrinsic scatter):

```
# modeling Lx -n relation
z[i] <- alpha+44+beta*(log(n[i])/2.30258-1.8)
lgLx[i] ~ dnorm(z[i], prec.intrscat)
# prior
intrscat <- 1/sqrt(prec.intrscat)
prec.intrscat ~ dgamma(1.0E-5,1.0E-5)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1) .
```

Table 8.7 Data for the X-ray luminosity-richness scaling, from Andreon and Moretti (2011)

ID	obsgaltot[i]	obsgalbkg[i]	Cgal[i]	obstot[i]	obsbkg[i]	nbox[i]	C[i]
1	25	61	9.95	259	4079	33	40.41
2	60	29	2.52	35	269	22	41.49
3	115	73	3.78	457	356	51	41.73
6	68	29	4.00	156	325	52	41.77
7	68	18	1.70	53	256	25	41.38
9	4	19	20.33	415	3827	10	39.67
10	72	62	3.85	59	190	41	41.80
11	7	39	34.20	231	4174	35	40.26
12	13	120	28.95	176	4509	63	40.63
13	41	104	12.52	150	4503	61	40.66
16	81	29	1.75	204	330	13	41.04
17	64	22	3.62	296	171	29	41.68
19	53	45	4.73	277	2326	34	40.59
20	48	11	2.01	567	188	10	41.30
23	89	46	3.60	101	562	58	41.88
24	67	21	2.31	179	239	27	41.84
25	76	32	3.48	59	400	43	41.74
26	114	33	2.01	223	337	19	41.20
27	48	65	7.41	190	340	65	41.97
28	46	38	3.36	310	190	14	41.43
29	95	41	3.10	101	262	48	41.90
30	27	18	5.24	434	840	17	40.94
31	59	39	5.51	691	396	36	41.60
32	47	50	7.04	49	724	52	41.29
33	23	172	28.64	380	4650	61	40.80
34	45	43	8.30	38	468	77	41.72
35	15	80	38.45	144	2539	58	40.76
36	8	10	8.05	79	828	15	40.59
37	32	27	7.13	53	199	25	41.63
39	21	101	19.48	412	3302	66	40.87
40	17	39	11.73	50	1098	28	40.91
42	104	87	5.50	33	387	98	42.13
43	139	124	5.37	96	198	96	42.51

The full model, whose graph is shown in Fig. 8.48 then reads:

```
model{
for (i in 1:length(obstot)) {
 # modelling X-ray photons
 obstot[i] ~ dpois(nclus[i]+nbkgind[i]/nbox[i])
 nbkgind[i] ~ dlnorm(log(nbkg[i]),1/0.2/0.2)
 obsbkg[i] ~ dpois(nbkg[i])
 nbkg[i] ~ dunif(0,10000)
 # convert nclus in Lx
 nclus[i] <- exp(2.30258*(lgLx[i]-C[i]))
 # modelling galaxy counts
 # n200 term
 obsgalbkg[i] ~ dpois(ngalbkg[i])
 obsgaltot[i] ~ dpois(ngalbkg[i]/Cgal[i]+n[i])
 n[i] ~ dunif(1,3000)
 ngalbkg[i] ~ dunif(0,3000)
 # modeling Lx -n200 relation
 z[i] <- alpha+44+beta*(log(n[i])/2.30258-1.8)
 lgLx[i] ~ dnorm(z[i], prec.intrscat)
 }
```

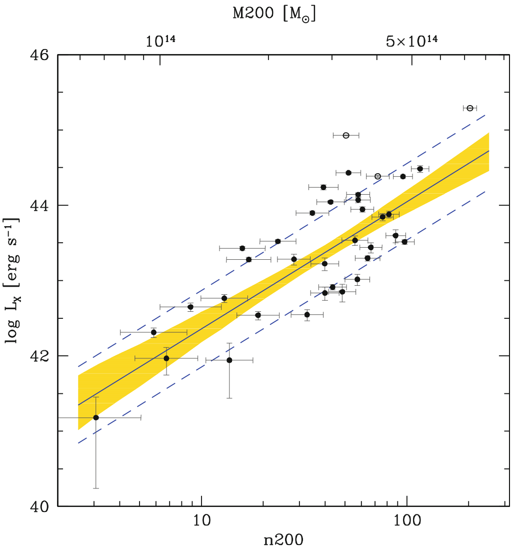

Fig. 8.49 X-ray luminosity-richness scaling. The *solid line* indicates the mean fitted regression line of $\log L_X$, on $\log n$, while the *dashed line* shows this mean plus or minus the intrinsic scatter σ_{scat}. The *shaded region* marks the 68 % probability interval for the regression. The *upper abscissa* indicates the cluster mass. Reproduced from Andreon and Moretti (2011) with permission

```
intrscat <- 1/sqrt(prec.intrscat)
prec.intrscat ~ dgamma(1.0E-5,1.0E-5)
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
}.
```

Using this model, we found for the samples of 33 color-selected clusters in Andreon and Moretti (2011) listed in Table 8.7:

$$\texttt{lgLx} = (1.69 \pm 0.30)\ (\log \texttt{n} - 1.8) + 43.71 \pm 0.11. \tag{8.13}$$

Figure 8.49 shows the scaling between richness and X-ray luminosity, the observed data, the mean scaling (solid line) and its 68 % uncertainty (shaded yellow region) and the mean intrinsic scatter (dashed lines) around the mean relation. The ±`intrscat` band is not expected to contain 68 % of the data points because of the measurement errors. All points are, however, within twice the intrinsic scatter.

Figure 8.50 shows the posterior probability distribution of the intercept, slope, and intrinsic scatter σ_{scat}. These probability distributions are reasonably well approximated by Gaussians. The intrinsic L_X scatter at a given richness, $\sigma_{scat} = \sigma_{lgL_X|\log n200}$, is very large, 0.51 ± 0.08 dex. In other words, a whole 1 dex in L_X is needed to bracket 68 % of all clusters of a given richness.

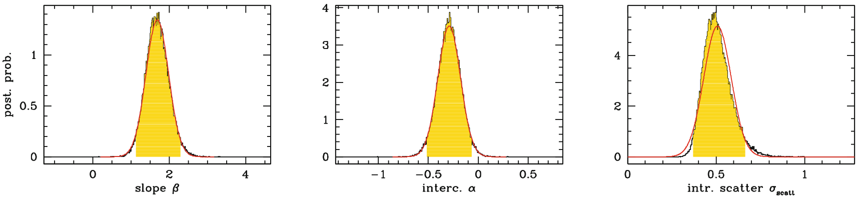

Fig. 8.50 Posterior probability distribution for the parameters of the X-ray luminosity-richness scaling. The *black jagged histogram* shows the posterior as computed by MCMC, marginalized over the other parameters. The *red curve* is a Gaussian approximation of it. The *shaded (yellow) range* shows the 95 % probability interval. The jagged nature of the histogram is caused by the finite sampling of the posterior. Reproduced from Andreon and Moretti (2011) with permission

8.13 Exercises

Exercise 1

Compare analytically and numerically computed posteriors. D'Agostini (2005) analytically computed the posterior probability distribution of a linear fit with normal errors and a Gaussian intrinsic scatter (shown in Sect. 8.4), known in the statistical literature as the errors-in-variable Bayesian regression (Dellaportas and Stephens 1995). After having analytically derived the posterior (perhaps with the help of D'Agostini 2005), generate some fake data and fit them. Then compare the numerically and analytically derived posteriors.

Exercise 2

Recycle. The model introduced in Sect 8.4, a trend between two quantities, *both* subject to errors and with an intrinsic scatter, is quite general and may find many applications. Use it for computing the scaling between stellar mass and cluster mass using the data[11] in Andreon (2010a). Then, compute the scaling between the stellar mass *fraction* vs. cluster mass. Note that cluster mass enters in both abscissa and ordinate (stellar mass fraction is stellar mass over cluster mass). There is a way to account for the error covariance that requires no work at all, fitting stellar mass vs. cluster mass! (Answer:

$$lgM_\star = (0.45 \pm 0.08)\ (\log M - 14.5) + 12.68 \pm 0.03, \tag{8.14}$$

[11] The table is given in electronic format at the link http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/fstar.dat.

and, as a consequence,

$$\log f_{\star} = (-1 + 0.45 \pm 0.08)\ (\log M - 14.5) + 12.68 \pm 0.03. \qquad (8.15)$$

)

Exercise 3

Recycle. The model for estimating efficiency and completeness, illustrated in Sect. 8.2, is perfectly suited for describing the run of the fraction of galaxies of early-type morphology with distance from the cluster's center. Using the data,[12] in Andreon et al. (1996) and Andreon et al. (1997), compute the radial profile of the fraction of early-type galaxies in the Coma cluster.

Exercise 4

Practice writing your own JAGS code. Consider the simple data set:

`obsY`	-7.821	-1.494	-15.444	-10.807	-13.735	-14.442	-15.892	-18.326
`obsX`	5	6	7	8	9	10	11	12

For this data assume the simple linear regression model

```
Z[i] = a + b * obsX[i]
obsY[i] ~ dnorm(Z[i],prec).
```

Experts familiar with the process that generated this data expect that `a` $\in [1,10]$, `b` $\in [-1,3]$, and `prec` $\in [.03,4]$. Use this information to formulate your own prior distributions and then obtain the posterior distributions of `a`, `b`, and `prec`.

Exercise 5

Return to the previous exercise (i.e., Exercise 4). Using the posterior distribution from Exercise 4, calculate:

Exercise 5a

the posterior distribution of $\sigma = 1/\sqrt{\texttt{prec}}$

[12] The table is given in electronic format at the link http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/Coma_earlytype.dat.R.

Exercise 5b

the posterior distribution of $\mathtt{a}+\mathtt{b}*13$

Exercise 5c

the prior predictive distribution of $\mathtt{a}+\mathtt{b}*13$

Exercise 5d

the posterior predicted distribution of $\mathtt{a}+\mathtt{b}*13$
In particular, compare the distributions from 5b and 5c and the distributions from 5c and 5d.

Exercise 6

Recreating history. The following data set is the same data set that Edwin Hubble used to show that galaxies are either moving away from us or towards us. Using this data and your own priors, fit the model

```
obsV[i] ~ dnorm(b * obsD[i],prec),
```

where `obsV` is the observed velocity (in units of km/s), and `obsD` is the observed distance (in units of 106 parsecs).

obsD	obsV	obsD	obsV	obsD	obsV	obsD	obsV	obsD	obsV
0.032	170	0.275	−220	0.8	300	1	920	2	500
0.034	290	0.45	200	0.9	−30	1.1	450	2	850
0.214	−130	0.5	290	0.9	650	1.1	500	2	800
0.263	−70	0.5	270	0.9	150	1.4	500	2	1090
0.275	−185	0.63	200	0.9	500	1.7	960		

Exercise 7

Regression using indicator variables. Consider the data set given in the table below. Using this data, fit the model

```
obsY[i] ~ dnorm(a+b * obsX[i]+r*Z[i],prec).
```

Prior information that can be used for the parameters are a $\in (0,2)$, b $\in (0,16)$, r $\in (0,2)$ and prec $\in (0.577,1)$.

obsY	10.79	18.72	23.89	31.53	39.77	50.73	57.86	69.76	76.93	83.02
obsX	1	2	3	4	5	6	7	8	9	10
Z	0	0	0	0	0	1	1	1	1	1

References

S. Andreon and M.A. Hurn. Measurement errors and scaling relations in astrophysics: a review. *Statistical Analysis and Data Mining*, 6:15–33, 2013.

A. Sandage. The redshift-distance relation. III. Photometry and the Hubble diagram for radio sources and the possible turn-on time for QSOs. *The Astrophysical Journal*, 178:25–44, 1972.

F. Pacaud, M. Pierre, C. Adami, B. Altieri, S. Andreon, L. Chiappetti, and et al. The XMM-LSS survey: the class 1 cluster sample over the initial 5 deg2 and its cosmological modelling. *Monthly Notices of the Royal Astronomical Society*, 382 (3):1289–1308, 2007.

S. Andreon and A. Moretti. Do X-ray dark, or underluminous, galaxy clusters exist? *Astronomy & Astrophysics*, 536(A37), 2011.

M. Wendt and P. Molaro. Robust limit on a varying proton-to-electron mass ratio from a single H_2 system. *Astronomy & Astrophysics*, 526:A96, 2011.

S. Tremaine, K. Gebhardt, R. Bender, G. Bower, A. Dressler, S. M. Faber, and et al. The slope of the black hole mass versus velocity dispersion correlation. *The Astrophysical Journal*, 574:740–753, 2002.

A. Raichoor and S. Andreon. Star formation and environment in clusters up to $z \sim 2.2$. *Astronomy & Astrophysics*, 537:A88, 2012.

S. Andreon and M.A. Hurn. The scaling relation between richness and mass of galaxy clusters: a Bayesian approach. *Monthly Notices of the Royal Astronomical Society*, 404(4):1922–1937, 2010.

P. Schechter. An analytic expression for the luminosity function for galaxies. *The Astrophysical Journal*, 203:297–306, 1976.

A. Jenkins, C. S. Frenk, S. D. M. White, J. M. Colberg, S. Cole, A. E. Evrard, and et al. The mass function of dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 321:372–384, 2001.

S. Andreon and J. Bergé. Richness-mass relation self-calibration for galaxy clusters. *Astronomy & Astrophysics*, 547:A117, 2012.

M. C. March, R. Trotta, P. Berkes, G. D. Starkman, and P. M. Vaudrevange. Improved constraints on cosmological parameters from Type Ia supernova data. *Monthly Notices of the Royal Astronomical Society*, 418:2308–2329, 2011.

R. I. Thompson, J. Bechtold, J. H. Black, D. Eisenstein, X. Fan, R. C. Kennicutt, and et al. An observational determination of the proton to electron mass ratio in the early Universe. *The Astrophysical Journal*, 703:1648–1662, 2009.

W. Ubachs, R. Buning, K. S. E. Eikema, and E. Reinhold. On a possible variation of the proton-to-electron mass ratio: H_2 spectra in the line of sight of high-redshift quasars and in the laboratory. *Journal of Molecular Spectroscopy*, 241:155–179, 2007.

S. Andreon. Understanding better (some) astronomical data using bayesian methods. *Astrostatistical Challenges for the New Astronomy, edited by J. Hilbe.*, Publisher: Springer Series on Astrostatistics:41–62, 2012b.

R. Kessler, A. C. Becker, D. Cinabro, J. Vanderplas, J. A. Frieman, J. Marriner, and et al. First-year Sloan Digital Sky Survey-II supernova results: Hubble diagram and cosmological parameters. *The Astrophysical Journal Supplement*, 185:32–84, 2009.

W. L. Freedman, B. F. Madore, B. K. Gibson, L. Ferrarese, D. D. Kelson, S. Sakai, and et al. Final results from the Hubble Space Telescope Key Project to measure the Hubble constant. *The Astrophysical Journal*, 553:47–72, 2001.

S. Andreon. The enrichment history of the intracluster medium: a Bayesian approach. *Astronomy & Astrophysics*, 546:A6, 2012a.

S. Andreon and A. Moretti. Do X-ray dark, or underluminous, galaxy clusters exist? *Astronomy & Astrophysics*, 536(A37), 2011.

G. D'Agostini. Fits, and especially linear fits, with errors on both axes, extra variance of the data points and other complications. *Arxiv preprint physics/0511182*, 2005.

P. Dellaportas and D.A. Stephens. Bayesian analysis of errors-in-variables regression models. *Biometrics*, 51(3):1085–1095, 1995.

S. Andreon. The stellar mass fraction and baryon content of galaxy clusters and groups. *Monthly Notices of the Royal Astronomical Society*, 407:263–276, 2010a.

S. Andreon, E. Davoust, R. Michard, J.-L. Nieto, and P. Poulain. Morphological classification and structural parameters for early-type galaxies in the Coma cluster. *Astronomy & Astrophysics Supplement*, 116:429–445, 1996.

S. Andreon, E. Davoust, and P. Poulain. Morphological classification and structural parameters of galaxies in the Coma and Perseus clusters. *Astronomy & Astrophysics Supplement*, 126:67–72, 1997.

Chapter 9
Model Checking and Sensitivity Analysis

In this section we introduce two tasks:

1. model checking, i.e., assessing whether the considered model fits the data in some sense or needs to be revised;
2. sensitivity analysis, i.e., assessing the effects on conclusions caused by deviations from the assumptions of the statistical model.

The purpose of model checking is to understand whether the model accurately describes the data, and, in the case it does not, the ways in which the fitted model does not fit the data and how to improve it. Therefore, the purpose of model checking is related to, but different from, rejecting a model, or choosing between a collection of candidate models. The reader interested in these topics and these conceptual subtleties should consult Gelman et al. (1996), Bayarri and Castellanos (2007) and, most importantly, the discussion at the end of these papers.

Let us emphasize that all models that describe our world are wrong in the sense of being approximations to the "true" model (e.g., Newtonian gravitation is an approximation of general relativity), yet even approximated models are still useful in many circumstances (as the late George Box said, "All models are wrong, but some are useful"). For example, all cosmological simulations use Newtonian gravity instead of general relativity gravity (Frenk and White 2012). A model is hardly useful if it does not fit the data in hand (those being fitted). A poor model fit is, however, a starting point for building a better model. For example, the poor fit between redshift and luminosity distance of SNIa leads to the building of a better fitted model (the discovery of the dark energy, which awarded a Nobel prize, see Sect. 8.12.1).

Therefore, the purpose of this chapter is not to determine and declare that once and forever whether a fitted model is correct or wrong for centuries to come (it is certainly wrong, for sure), but if the considered model is at odds with the current available data (the fitted data), for example because it is over-simplified compared to some specific complexity pointed out by the data. In the latest case, it needs to be updated right now. In the former case, be prepared, gentle reader, to state (today) that the model fits the data and to discover, perhaps in a century from now, that

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_9

the model no longer fits the much better data that will be available in the future. The astronomer may recall, for example, the famous velocity vs. distance *linear* fit by Hubble (1925) and remember that the relation between these quantities is, today, obviously non-linear. The physicist may recall the *linear* composition of velocities and remember how to compose them at speeds near to those of light.

Sensitivity analysis is used to investigate how deviations from your statistical model and its assumptions affect your conclusions. Examples of such deviations are:

- deviations from the distributional assumptions (i.e., deviations from the assumptions regarding the stochastic elements of your model),
- alternative functional forms for the relationships in your model,
- alternative prior distributions.

In summary, model checking and sensitivity analysis are tools used to assess and understand the current adopted model. The purpose of this chapter is to provide some advice on how the careful scientist should perform model checking and sensitivity analysis. We begin with sensitivity analysis and conclude with model checking.

9.1 Sensitivity Analysis

Every research analysis makes assumptions (e.g., that counts fluctuate according to a Poisson likelihood, that the relation between two quantities has some relationship and not another). Some of these assumptions are likely well justified and unlikely to be wrong. Some other assumptions are taken for simplicity or for other reasons, but without any strong justification. Before deciding that the analysis is complete, start relaxing, and being proud of our own work, it is important to test the sensitivity of the results to the adopted assumptions, and, if any of these assumptions contain uncertainty, estimate the impact of its uncertainty on the results. It may be that the results are robust to the uncertainty assumption, or perhaps overly sensitive to it. It is important to know and inform the reader how much the results depend on the uncertain assumption.

There is ample literature on the subject of incorporating model uncertainty (i.e., we made a good assumption by picking the correct model) into the estimation of parameter uncertainties, Hoeting et al. (1999) is a good entry point. At the very least *we need to replace uncertain assumptions with other plausible assumptions and check how the results change.* These can be split into three cases (though these cases may not be exhaustive): check a different prior distribution, check an alternative link function, and check alternative distributional assumptions.

9.1.1 Check Alternative Prior Distributions

We devoted all of Chap. 5 to discussing the importance of checking the prior's influence on the posterior and we reiterated this point several times within this book (e.g.,

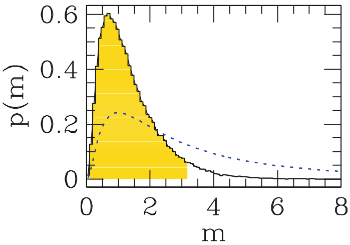

Fig. 9.1 Posterior probability distribution (*solid histogram*) of the neutrino mass for a lognormal prior (*dotted blue curve*). The *shaded region* marks the 95 % highest posterior credible interval

Sects. 6.2, 8.7, 8.9, 8.12.2), and we can only re-emphasize its importance here and illustrate it by adopting an alternative prior for the neutrino experiment of Sect. 4.1.

Recall we made a mass observation of −5.4 with the likelihood's width (σ) set at 3.3. We adopted a uniform prior between 0 and 33.

We now consider an alternative prior: we take a lognormal prior with both parameters equal to 1. It is a smooth density (gently decreasing at high and low values), has its median equal to 2.7, and contains 97 % of the probability between 0.33 and 16.5 (i.e., between 0.1 and roughly 5 times the instrument resolution). This prior is plotted as the dashed curve in Fig. 9.1. With the revised (lognormal) prior, the model reads:

```
model{
obsm ~ dnorm(m,prec)
m ~ dlnorm(1,1)
prec <- pow(3.3,-2)
}.
```

With the lognormal prior, the 95 % upper limit to the neutrino mass is 3.2 eV/c^2. With the previously adopted (uniform) prior, it was 3.8 eV/c^2. The two upper limits are not considerably different (remember, the instrument resolution is 3.3 eV/c^2).

In this example, conclusions (i.e., the upper limit) depend little on which one of the two priors is adopted.

9.1.2 Check Alternative Link Functions

Suppose you have modelled the y vs. x relation using a given function $y = f(x)$, and suppose you have chosen the link function f mostly for convenience, not because theory predicts that the form of the function f is the one you should assume. It may be (or may be not, this is the point to be checked) that your conclusions rely on the assumed form of the function f. For example, you may have taken f to be linear (for convenience) when instead it might be a non-linear function. Or you may have taken a constant (i.e., predictor-independent) intrinsic scatter (as everywhere in this book) when instead it might be a linear function of the predictor (e.g., the scatter increases as the predictor decreases).

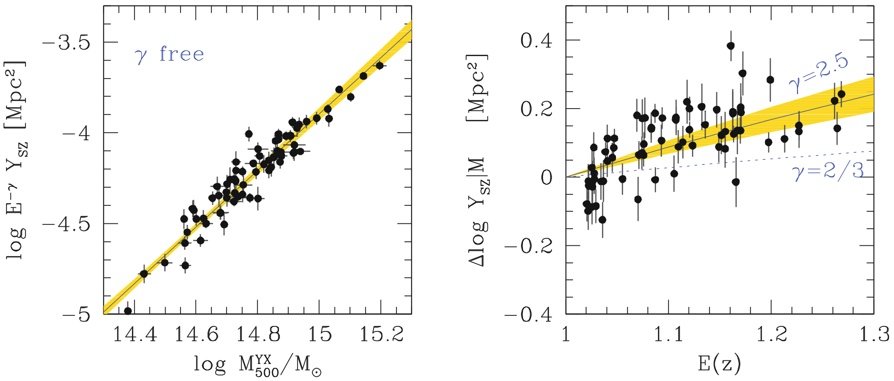

Fig. 9.2 Mass-Y_{SZ} scaling (left–hand panel) and residuals (observed minus expected) as a function of redshift (right–hand panel). The solid line marks the mean fitted regression line. The shaded region marks the 68% uncertainty for the regression. In the left–hand panel, measurements are corrected for evolution. In the right–hand panel the dotted line shows the $E^{2/3}$ dependency assumed by the Planck team. Reproduced from Andreon (2014) with permission

Uncertainties on the form of the link f function can be incorporated as usual in the prior-to-posterior inference: for example, the possibly non-constant scatter can be modelled as a linear function of the predictor quantity. Sometimes, in practice, this super-model often requires hard-to-know priors (e.g., on the slope of the linear coefficient) or on hard-to-know alternative links (the alternative link should be linear or an exponential?). In spite of these difficulties, alternative link functions must be explored, in particular when one suspects that conclusions may be badly impacted by an alternative form of the link function.

Let us consider the Sunayev-Zeldovich mass-observable fitted by the Planck team (Planck Collaboration et al. 2013) shown in Fig. 9.2. The authors have recently derived the value of the cosmological parameters using two probes, the cosmic microwave background and clusters of galaxies. While cosmological parameters are unique, the values found using the two probes differ from each other. Let us look in more detail at the galaxy cluster probe: the cosmological constraint comes from matching the galaxy cluster abundance (per unit observable, `YSZ`) to the cluster's mass function (per unit mass `M`), via a mass-observable relation. This relation f is taken by Planck Collaboration et al. (2013) to be

```
logYSZ=alpha + beta * log(M/Mref) +2/3 * log10 E(z) ,
```

where the coefficient 2/3 is the value suggested by (a simple) theory, and with `alpha` and `beta` to be determined from the data. `Mref` is the usual value for centering quantities near their average to speed up computations and to simplify the interpretation of the found results. `log10 E(z)` is the decimal log of a simple function of the redshift (distance) `z`. The key point to remember is that the derived value of the cosmological parameters depends on the value of the parameters `alpha` and `beta`, a point well known in the astronomical literature. Performing a sensitivity analysis on the form of the link function f is, therefore, paramount!

Andreon (2014) argues that the link function f may be slightly different, with the difference being that the $2/3$ (fixed) coefficient has to be replaced by a gamma parameter which is to be estimated from the data rather than assuming its value is known. In fact, as reported in Andreon (2014), there is no observational evidence that gamma$= 2/3$, and the estimated value points toward a different gamma, as shown below.

The full analysis of this data set requires us to properly account for the Malmquist bias and the sample selection function,[1] accounted for in Andreon (2014), but ignored here to focus on the main point. We instead fit the very same bias-corrected data[2] fitted by the Planck team with the errors-in-variable regression model introduced in Sect. 8.4 (and which has already been used a number of times in this book):

```
model {
for (i in 1:length(obslogM)) {
logM[i] ~ dunif(13,17)
obslogM[i] ~ dnorm(logM[i],pow(errlogM[i],-2))
logYm[i] <- alpha-4.3 +beta*(logM[i]-14.778)+gamma*
             log(Ez[i])/2.303
logYSZ[i] ~dnorm(logYm[i], prec.scat)
obslogYSZ[i] ~ dnorm(logYSZ[i],pow(errlogYSZ[i],-2))
}
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
prec.scat ~ dgamma(1.0E-2,1.0E-2)
intrscat <- 1/sqrt(prec.scat)
gamma <-2/3 # to hold gamma fixed.
}.
```

When fitted to the Planck data, this model returns beta$= 1.68 \pm 0.06$.

As argued by Andreon (2014), what if the considered link function

```
logYm[i] <- alpha-4.3 +beta*(logM[i]-14.778)+gamma*
             log(Ez[i])/2.303
```

with fixed gamma were replaced by another with a variable gamma? To test this hypothesis, we just need to replace the single line

```
gamma <-2/3
```

with

```
gamma ~ dt(0,1,1) .
```

With the revised model, beta$= 1.51 \pm 0.07$, fairly different from the Planck team's preferred value of beta$= 1.79 \pm 0.06$ which they obtained by fitting the very same data with gamma held fixed at $2/3$. The two fits differ by $\approx 3\sigma$, which is

[1] This full analysis was performed using a JAGS model whose building blocks are all listed in this book. The reader can therefore perform the full analysis, basically by cut & paste.

[2] The data are available at: http://www.brera.mi.astro.it/~andreon/BayesianMethodsForThePhysicalSciences/planckdata_biascor.dat.R.

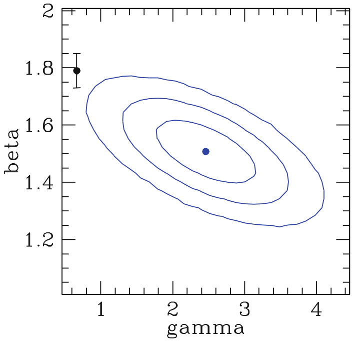

Fig. 9.3 Joint probability distribution of the mass-YSZ scaling. Contours are at 68 %, 95 % and 99.7 % probability level. The point with the error bar marks the value derived by the Planck team imposing gamma$= 2/3$. Reproduced from Andreon (2014) with permission

remarkable, because the very same data are being fitted.[3] Figure 9.3 shows the joint posterior probability distribution and the Planck team's preferred value (solid dot with error bar, derived with a non-Bayesian method fitting the very same data). The latter is close to the boundary of the 99.7 % probability contour, indicating that the data strongly favors a value of gamma lower than the one assumed by the Planck team. Expressed in different words, the data do not favor beta$= 2/3$, assumed by the Planck analysis. Back to our original purpose (sensitivity analysis, not cosmology), the slope beta strongly depends on the form of the linking function, in particular whether gamma should, or should not, be held fixed at $2/3$.

In summary, it is wise to check if and how conclusions depend on the adopted form of the linking function, sometimes an apparently minor difference, such as holding a parameter fixed or fitting it, may make a major difference.

9.1.3 Check Alternative Distributional Assumptions

As an example, let us consider the case of the Gaussian scatter adopted in the investigation of the enrichment history of the intracluster medium (Sect. 8.12.2). In this example, the existence of an intrinsic scatter is strongly motivated by the data's behavior. However, intrinsic scatter's nature, i.e., whether it is Gaussian, log-Normal, or some other distribution, is largely unknown. Section 8.12.2 adopts, following Andreon (2012a), a log-Normal scatter distribution. How strongly do the results depend on the (log-)Normality assumption on the scatter?

[3] This situation occurs because in the analyzed sample more distant clusters are also those with larger masses. Therefore, the slope beta and evolutionary term gamma are partially collinear, as shown in Fig. 9.3. The collinearity is partially broken by the spread in logY at a given z.

Following Andreon (2012a), we replace this assumption by another, a Student-t scatter, and check how the results change. In Sect. 8.12.2 we adopted:

```
Ab[i] ~ dlnorm(log(ft[i]), pow(intrscat,-2)) ,
```

which is now replaced by

```
Ab[i] ~ dt(log(ft[i]), pow(intrscat,-2), 10) .
```

We fit the new model to the same data used before. We found that values for the parameters and errors are quite similar to those in Sect. 8.12.2. In particular, we found `intrscat`$= 0.16 \pm 0.03$. A Student-t distribution has a standard deviation (parameter that we measured for the Gaussian) $\sqrt{10/8}$ times larger, i.e., 0.18 ± 0.04 to be compared with the (Gaussian) scatter 0.18 ± 0.03 derived in Sect. 8.12.2.

To summarize, for the fit of the enrichment history of the intracluster medium, the adoption of a Student-t or log-normal scatter makes no difference on the results.

9.1.4 Prior Sensitivity Summary

1. If your prior is uncertain, check the sensitivity to the prior, for example, by adopting a different distribution for the prior or by checking whether the prior is approximately constant where the posterior is non-null (see Chap. 5).
2. If your link function is adopted for simplicity, explore alternative functional forms for the link function.
3. If some adopted distribution is not strongly motivated by theoretical arguments (as often the intrinsic scatter is), explore alternative distributions.

If the results do depend on the stated assumptions, then: first, note that the Bayesian approach is offering you an easy way to discover that your conclusions depend on assumptions (and how), a dependence that may have escaped you otherwise. Second, we suggest to, at the very least, specify how the results depend on the stated assumptions. Third, consider incorporating the model uncertainty into the uncertainty of the final result, i.e., average results obtained with different assumptions following, e.g., Hoeting et al. (1999).

9.2 Model Checking

9.2.1 Overview

As mentioned, the work of the careful researcher does not end by finding the parameter set that best describes the data, (s)he must also check whether the adopted model is an acceptable description of the data, or if it is misspecified, i.e., in tension with the fitted data. Whenever feasible, graphical assessment should be done because plots allow us to directly inspect the magnitude of the various quantities and identify problems that are not visible otherwise. A number of such plots have been shown throughout this book: in the average of incompatible measurements (Sect. 6.2.2) a simple plot of the data (Fig. 6.19) shows that the data is incompatible with a single Gaussian model. Similarly, when modeling the distribution of globular cluster metallicities (Sect. 6.2.1), the histogram of the data values (Fig. 6.15) shows a bimodal distribution that cannot be modeled by a unimodal distribution such as a Gaussian. Recall in Chap. 8 we inspected many figures and noted that data scatter more than the error, and thus an intrinsic term is needed to describe the data.

Data summaries and visualizations (such as a histogram) are useful tools for building a model that fits the data. Moreover, these early investigations are a common way to discover that a preliminary model does not fit. Visual inspection is very useful for spotting anomalies (differences between the behavior of the data in hand and of the simulated data) which should carefully be investigated later.

After inspecting the data and fitting a model, it is useful to inspect the residuals. Inspecting the residuals is one way of looking for anomalies or discrepancies (unexpected behavior of the residuals) in the model fit. For example, if we fit a linear model to our data and our residuals (the difference between our data and the mean of the model) are not symmetrically scattered about zero, then we know that our model is not adequately describing the data.

If a possible discrepancy is detected or is suspected, or if one is just willing to assess a feature of the model, then how can we quantitatively determine the statistical significance of the discrepancy? In the non-Bayesian paradigm (i.e., frequentist statistics), the statistical significance of the discrepancy is measured by computing a p-value, i.e., the probability of obtaining new data that are more extreme (or discrepant) than the data in hand once parameters are taken at the best fit value. The Bayesian version of this concept (e.g., Gelman et al. 2004) acknowledges that parameters are not perfectly known, and therefore one should also explore, in addition to best fit values, other values of the parameters. In practice, discrepancy becomes a vector, instead of a scalar, of dimension j, that measures the distance between the data and j potential models, one per set of parameters considered. Of course, more probable models should occur more frequently in the list, quantifying that discrepancy from an unlikely model is less detrimental than discrepancy from a likely model. Therefore, if parameters are explored by sampling (as in this book), it is just a matter of computing the discrepancy of the data in hand for each of j sets of parameters stored in the chain, instead of relying on one single set of parameter values (say, those that maximize the likelihood). Once one has acquired the discrepancy vector for the data, one then repeats the computation for fake data generated from the model, and counts how many times the simulated data are more extreme than the data on hand, as illustrated in detail in the next sections.

9.2.2 Start Simple: Visual Inspection of Real and Simulated Data and of Their Summaries

In Sect. 8.2 we computed the trigger efficiency of a CERN experiment by fitting a phi (error) function to experimental data. The fitting model is quite simple, there are only very few features that we may get wrong and therefore there is little to test. As mentioned, in order to test whether the model fits the data, we need simulated (replicated) data from the model. These can be generated by using

```
nrec.rep[i] ~ dbin(eff[i],ninj[i]) ,
```

which draws for each energy bin `[i]` a value `nrec.rep[i]` from a binomial distribution having the same parameters as those from the data fit. By this operation, we are generating simulated data drawn from the model which is assumed to be a correct description of the data. In practice, we have taken the existing code line corresponding to the part of the model that we want to test and replicate it by adding a `.rep` to the variable name.

The left panel of Fig. 9.4 shows the first data set generated. Qualitatively speaking, it does not look much different from the real data set, as the few other simulated data sets that we have inspected. Changing from a qualitative to a quantitative statement about the similarities between the true and simulated data, we compute (and plot) the 68 % interval of replicated data and compare them to the observed data. This comparison is shown in the right panel of Fig. 9.4. Figure 9.4 shows a good agreement, and suggests that a more attentive exploration of possible model discrepancies is probably not worth the time it takes (but those of you with spare time or with a real experiment where a precise modeling of the energy dependency of efficiency is an important issue should follow the instructions in the next section).

To summarize, an attentive visual inspection of the data, of simulated data, and of summaries of simulated data (as the 68 % intervals) is the first step of model checking. We invite the reader to plot the data and the model in all the ways that physical intuition suggests, looking for discrepancies of different kinds (offsets between

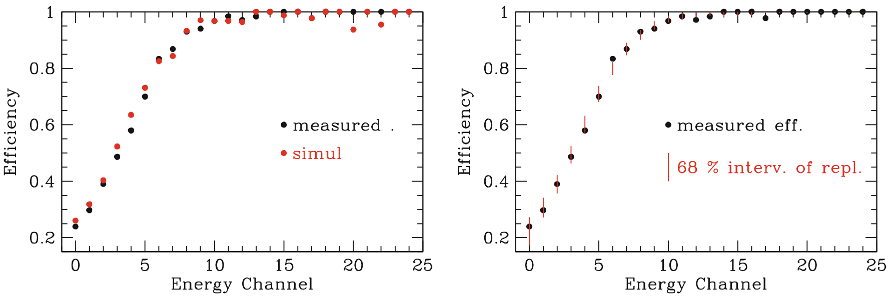

Fig. 9.4 *Left panel:* Real (*black*) and simulated (*red*) data as a function of the energy bin. *Right panel:* 68 % range of simulated data (*red intervals*) vs. observed data (*point*)

model and data, too much/little scatter around the mean model, localized systematic differences, etc.). If one is found, then change the model until discrepancies disappear.

9.2.3 A Deeper Exploration: Using Measures of Discrepancy

Let us now consider the cosmological estimate using SNIa of Sect. 8.12.1. We assumed there a linear relation with an intrinsic scatter between the absolute mag, color and stretch parameters (see Sect. 8.12.1 for details):

```
mcor[i] <- M+distmod[i]- alpha * x[i] + beta * c[i]
m[i] ~ dnorm(mcor[i], pow(intrscatM,-2) .
```

Suppose that we wish to test this assumption. To this purpose we simulate `m.fake[i]` magnitudes and, from these, we also simulate observations with the same errors as our true data:

```
m.fake[i] ~ dnorm(mcor[i], pow(intrscatM,-2)
obsm.fake[i] ~ dnorm(m.fake[i], precmag[i]) .
```

In practice, we added the following three JAGS lines:

```
mcor[i]<-Mm+distmod[i]- alpha* x[i] + beta*c[i]
m.fake[i] ~ dnorm(mcor, precM)
obsm.fake[i] ~ dnorm(m.fake[i],precmag[i])
```

and we replaced one line by the simpler

```
m[i] ~ dnorm(mcor[i], precM) .
```

We performed 15,000 simulations,[4] each one generating 288 simulated measurements of SNIa (one per real measurement).

Visual inspection of the data does not indicate any obvious discrepancies. We therefore begin our exploration by adopting a modified χ^2 to quantify discrepancy (or its contrary, agreement). For the real data set we have:

$$\chi^2_{real}[j] = \sum_i \frac{(\texttt{obsm[i]} - \texttt{mcor[i,j]})^2}{\texttt{errmag[i]}^2 + E^2(\texttt{intrscatM})}, \qquad (9.1)$$

where the summation is over the data and j refers to the index in the sampling chain. Apart from the j index, Eq. (9.1) is just the usual χ^2 test statistic, the difference between observed, `obsm[i]`, and true `mcor[i]`, values, weighted by the expected variance, computed as a quadrature sum of errors, `errmag[i]`, and

[4] Skilled readers may note that we are dealing with Gaussian distributions, and may attempt an analytic computation.

supernovae mag intrinsic scatter `intrscatM`. Similarly, the χ^2 of the jth fake data set, $\chi^2_{fake}[j]$ is:

$$\chi^2_{fake}[j] = \sum_i \frac{(\texttt{obsm.fake[i,j]} - \texttt{mcor[i,j]})^2}{\texttt{errmag[i]}^2 + E^2(\texttt{intrscatM})}, \tag{9.2}$$

We now compute which fraction of the simulations has $\chi^2_{fake}[j] > \chi^2_{real}[j]$ and quote the result (this may also be coded inside JAGS, asking it to return a chain of 0 or 1 depending on whether the inequality above is satisfied). If the modeling is appropriate, then the computed fraction (p-value) is not extreme (far from zero or one). If the modeling is not appropriate, our statistical model needs to be revised because the data are in disagreement with the model.

For the data used in Sect. 8.12.1, we found a p-value of 45 %, i.e., that the discrepancy of the data in hand is quite typical (similar to the discrepancy of the fake data). Therefore, according to the χ^2 measure of discrepancy, the real data is adequately described by our statistical model (at least the aspect of the model that was tested).

In our exploration of possible model misfits, we adopted a χ^2 measure of discrepancy. This is just one of the many measures of discrepancy one may adopt, and, furthermore, it does not detect all types of discrepancies. It is useful to visually inspect several data summaries to guide the choice of which discrepancy measure one should adopt (e.g., Eq. 9.1 or something else), and, if the adopted model turns out to be inadequate, to suggest how the tested aspect of the model should be revised. For example, a possible (and common) data summary is the distribution of normalized residuals, that for `errobsx[i]` reads:

$$\texttt{stdresobsx[i]} = \frac{\texttt{obsx[i]} - E(\texttt{x[i]})}{\sqrt{\texttt{errobsx[i]}^2 + E^2(\texttt{intrscatx})}} \tag{9.3}$$

i.e., observed minus expected value of `x[i]` divided by their expected spread (the sum in quadrature of errors and intrinsic spread). A similar summary may be built for `obsc[i]` too. At the very least, standardized residuals should be distributed as a Gaussian with a standard deviation of one (by construction). Figure 9.5 shows the distribution of normalized residuals of `obsx[i]` and `obsc[i]`, with a Gaussian distribution superimposed and centered at 0 with a standard deviation equal to one (in blue). Both distributions show a possible hint of asymmetry. At this point, the careful researcher may want to use a discrepancy measure that is useful for detecting asymmetries, for example the skewness index, in addition to the χ^2 during model testing. While leaving the actual computation to the reader, we emphasize that if an extreme Bayesian p-value is found (on `obsx[i]` for exposing simplicity), then one may update the model (the JAGS line describing its distribution) with a distribution allowing a non-zero asymmetry which is easily done using a Bayesian approach, and easily implemented in JAGS, it is just a matter of replacing the adopted Gaussian with a suitable asymmetric distribution. If instead the data exploration gives a hint of a double-bumped distribution (again on `obsx[i]` for exposing simplicity), and an extreme Bayesian p-value is found when a measure of discrepancy sensitive to

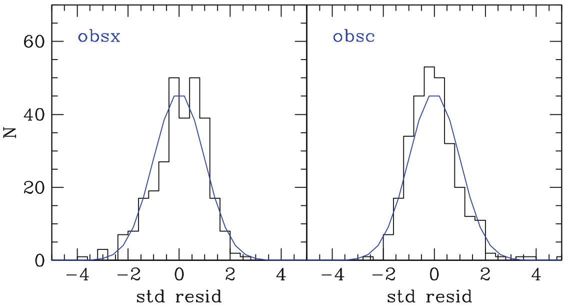

Fig. 9.5 Standardized residuals (*histogram*) and their expected distribution (*blue curve*). Reproduced from Andreon (2012b) with permission

double-bumped distributions is adopted, then one may adopt a mixture of Gaussians (see Sect. 6.2 about how this is implemented). Even more simple is the (hypothetical) case of possible distributions (again of `obsx[i]` for exposing simplicity) with fat tails: one may adopt a Student-t distribution. In such a case, this is implemented in JAGS by simply replacing a `dnorm` with `dt`.

Once the certain aspect of the model has had a thorough investigation, the careful researcher should then move to other aspects of the model, whose detailed exploration is left as an exercise.

9.2.4 Another Deep Exploration

Let us now check our modeling of the history of metal abundance described in Sect. 8.12.2. To this aim, we simulate data using the fitted model and we count which fraction of the simulations are more discrepant than the real data. Following the approach in Andreon (2012a) we use a modified χ^2 statistic as a measure of discrepancy:

$$\chi^2 = \sum_i \frac{(\mathtt{modeAb[i]} - E(\mathtt{Ab[i]}))^2}{\mathtt{sigmaAb[i]}^2 + E^2(\mathtt{intrscat})} \tag{9.4}$$

where the summation is taken over the data. This simulation is performed in JAGS, by generating replicated data:

```
Ab.rep[i] ~ dlnorm(log(Abcor[i]), pow(intrscat,-2))
modeAb.rep[i] ~ dnorm(Ab.rep[i],pow(sigmaAb[i],-2)) .
```

As illustrated in previous examples, we duplicated lines describing the part of the model that we want to test by adding a `.rep` to the variable name. The simulation automatically accounts for all modeled sources of uncertainties (intrinsic scatter, measurement errors, and their non-gaussian nature).

In practice, we performed 30,000 simulations, each one generating 144 simulated measurements of Fe abundance (one per actual observation). The first set of

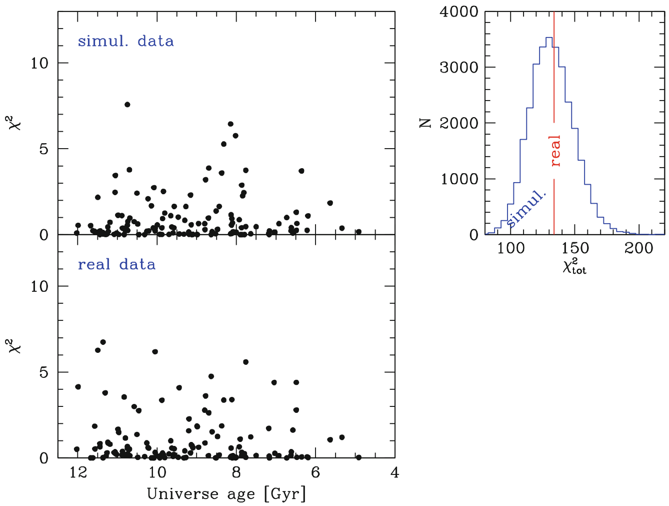

Fig. 9.6 *Bottom-left panel:* χ_i^2 values of the real data used for the metal abundance determination. *Top-left panel:* χ_i^2 values of simulated data. *Top-right panel:* Total χ^2 of real data (*vertical red line*) and distribution of total χ^2 of simulated data. Reproduced from Andreon (2012a) with permission

generated data is shown in Fig. 9.6 (top-left panel). We found that the χ^2 statistic of the real data is quite typical: 52 % of the simulated data sets show larger χ^2 (top-right panel) values. Therefore, it is quite reasonable to assume that our real data could have been generated by the fitted model and our model shows no evidence of misspecification.

Generating simulated data from the model and displaying them is useful because it allows the researcher to understand how a model can be improved. This is exemplified in Fig. 9.7, which shows a previous modeling attempt to the evolution of Fe content of clusters. The real data (bottom-left panel) look quite different than the simulated data (top-left panel) because of the presence of several large-χ^2 values missing in the simulated data. This occurs because this older modeling lacks an intrinsic scatter term. More specifically, in the plot we are using data, model and best fit from Balestra et al. (2007). The top-right panel shows that such a disagreement is almost never found using simulated data drawn from the fitted model. This misfit prompted Andreon (2012a) to revise the model, allowing an intrinsic scatter in the metal abundance.

9.3 Summary

In summary, model checking consists of continuously evaluating a proposed model and updating it until it produces data similar to those in hand. One should start by carefully and attentively inspecting the data and data summaries, plotting the data, evaluating residuals from the model fit, which includes displaying a histogram of the

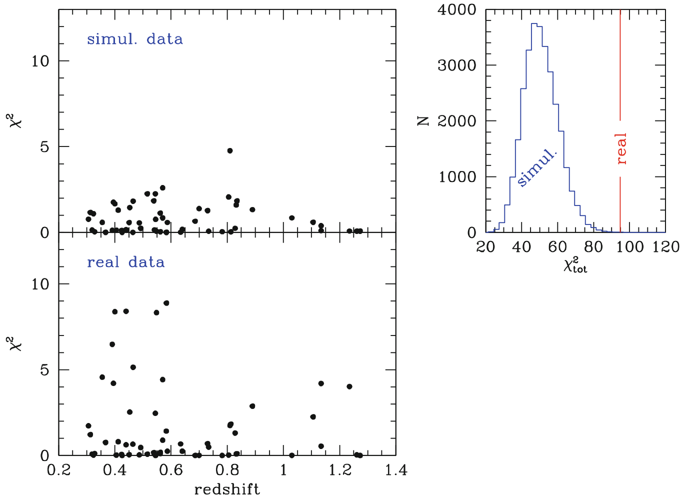

Fig. 9.7 Same description as Fig. 9.6 but for the sample and model in Balestra et al. (2007). Reproduced from Andreon (2012a) with permission

residuals, etc. This inspection should suggest some sort of measure of discrepancy which can be used to quantify the model misfit and, if one is found, to guide the model updating. The procedure should be iterated until the model fully reproduces the data. Indeed, the good agreement we found for the model in the last section is the result of failed attempts at modeling the history of metal abundance with simpler models. In a similar vein, the misfit of the (now old) model relating luminosity and redshift of SNIa (Sect.8.12.1) leads to the revision of the above relation and the introduction (discovery) of dark energy to allow the model to fit the data.

Sensitivity analysis consists of replacing uncertain assumptions with other plausible assumptions and check how the results change. Reporting whether results change, or not, is paramount to understanding the limits of the performed analysis and, in the case it does, to identify the appropriate observations needed to remove the found limits.

References

A. Gelman, X.L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):773–796, 1996.

M.J. Bayarri and M.E. Castellanos. Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22(3):322–343, 2007.

C. S. Frenk and S. D. M. White. Dark matter and cosmic structure. *Annalen der Physik*, 524:507–534, 2012.

E. P. Hubble. Cepheids in spiral nebulae. *The Observatory*, 48:139–142, 1925.

J.A Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

S. Andreon. The important role of evolution in the Planck Y_{SZ}-mass calibration. *Astronomy & Astrophysics*, 570:L10, 2014.

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, and et al. Planck 2013 results. XX. Cosmology from Sunyaev-Zeldovich cluster counts. *ArXiv e-prints*, page arXiv:1303.5080, 2013.

S. Andreon. The enrichment history of the intracluster medium: a Bayesian approach. *Astronomy & Astrophysics*, 546:A6, 2012a.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Champan and Hall/CRC, 2004.

S. Andreon. Understanding better (some) astronomical data using bayesian methods. *Astrostatistical Challenges for the New Astronomy, edited by J. Hilbe.*, Publisher: Springer Series on Astrostatistics:41–62, 2012b.

I. Balestra, P. Tozzi, S. Ettori, P. Rosati, S. Borgani, V. Mainieri, and et al. Tracing the evolution in the iron content of the intra-cluster medium. *Astronomy & Astrophysics*, 462:429–442, 2007.

Chapter 10
Bayesian vs Simple Methods

In this chapter, we illustrate the advantages and shortcoming of Bayesian methods against simple, non-Bayesian alternatives commonly used in astronomy and physics. We will not make an exhaustive comparison of all methods (there are too many alternate methods) and those methods we illustrate are selected because of their simplicity, and not necessarily the "best choice." Very good non-Bayesian approaches exist, and some of them are even mathematically equivalent (i.e., the same equations are solved) to their corresponding Bayesian methods (e.g., ridge regression). However, these are rarely used in astronomical and physics papers, and thus are not considered here. This chapter is not a review of Bayesian vs non-Bayesian methods, it is instead the view of a researcher faced with the day-to-day choice of using one of the models included in this book or a simple, non-Bayesian modeling approach commonly used. To be fair in our comparisons of Bayesian methods with their competitors, we will only focus on simple problems. This choice penalizes Bayesian methods because applying them to simple problems does not reveal how easily they are adapted to more complex situations (in our opinion), and will give, if overlooked, the impression that Bayesian methods should only be applied to simple problems.

In this chapter we first give a very short description of the conceptual differences of Bayesian and non-Bayesian methods and we then illustrate how the maximum likelihood estimate can fail. We consider, in particular, two cases: mixtures and small samples. We then compare the performances of one of the most used statistical methods in astronomy to infer the location (mean) and scale (dispersion) of a sample, with a simple Bayesian estimate, showing that the Bayesian estimate has a lower bias, is fairer (in some sense) and has less noisy errors. We then move to compare Bayesian and some non-Bayesian linear regression (fitting) models in various situations. We do not consider non-Bayesian non-linear models because these are rarely used in Astronomy and Physics. We will finally conclude with a summary and discuss the experience of the first author, trained as a non-Bayesian astronomer that discovered the Bayesian literature.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5_10

10.1 Conceptual Differences

In the Bayesian paradigm, the quantities we want to measure have a probability distribution (e.g., the prior) which describes how well we know these quantities. As Bayesian researchers, we may say "with 0.90 probability, the Higgs boson's mass is in this range." Repeated experimentation (if done correctly), hopefully, reduces the width of their probability distribution. In the Bayesian paradigm, the data are fixed (i.e., they are given), we have some data and we do not have some other data. Interestingly, datum (plural: data) is a Latin word which means "given," and therefore "data are data" (given, fixed). In short, in the Bayesian approach, the data are fixed and parameters have a probability distributions which are updated by experiments.

It is difficult to collectively describe non-Bayesian methods, because a large variety exists. Often, data are not data (given), but random: the data we have are one of the many possible sets of data that we may possibly observe. Typically, parameters are fixed (but unknown), without any associated probability distribution, and therefore these non-Bayesian researchers cannot say "the Hubble constant has a 90 % probability to be in this range." These methods, generally speaking, do not return how probable that a parameter is in a given range (a concept that simply does not exist: parameters have no probabilities associated with them), and offer something else as an alternative to it (e.g., the confidence interval).

Some methods define an event's probability as the limit of its relative frequency in a large number of trials. What is, for these methods, the probability of unique events, such as "tomorrow it is raining in New York," "there is life on Mars" or "the James Webb telescope will see the first stars"? The Bayesian approach solves this issue by associating probability with all quantities, including unique events.

10.2 Maximum Likelihood[1]

One of the most widely used statistical tools in astronomy and physics is maximum likelihood. We showed many times that the observed value, i.e., the maximum likelihood estimate, is a biased estimator in the presence of an abundance gradient (the Malmquist-Eddington bias, Sect. 5.1.1). In Sect. 10.4.1 we will remember the well-known fact that for regressions with errors on the predictor quantity the maximum likelihood estimate is not unique, but an infinite number of them exists. In the next two sections we will show two more examples of where the maximum likelihood methods are not adequate.

The first example shows that the maximum of the likelihood (the best fit) may not be a good estimate of the true value: averaging the likelihood is preferable to maximizing it. The second example shows that when the sample size is small, even simple operations on data, such as performing an average (i.e., taking the maximum

[1] Some of the material of this section has been drawn from Andreon (2010b).

likelihood estimate for the mean of a Gaussian) or fitting it with a function using maximum likelihood, is a potentially risky operation.

These two examples are quite simple (and we chose them to be) compared to true problems researchers are faced with, to make obvious certain unpleasant effects caused by using the best fit or observed value. Of course, in more complex problems these unpleasant effects could still be there (why should these more complicated applications not suffer the same problems?) but it would be hard to notice them in absence of a more appropriate analysis.

10.2.1 Average vs. Maximum Likelihood

Maximum likelihood estimates (called "best fit" by astronomers and physicists) are one of the most widely used tools and it is taken for granted that maximizing the likelihood (or minimizing the $\chi^2 = -2\ln\mathscr{L}$) will *always* give the correct result.

As repeatedly illustrated in this book, mixture distributions naturally arise in astronomy and physics when data comes from two populations, such as:

a) a signal superposed on a background;
b) there are interlopers in the sample;
c) there are two distinct populations;
d) there are two sources in an image;
e) there are two (stellar) populations in an object spectrum.

To focus this idea, let us consider the simple case of a mixture (sum) of two Gaussians:

$$p(y_i|\mu_1,\sigma_1,\mu_2,\sigma_2,\lambda) = \lambda\mathscr{N}(y_i|\mu_1,\sigma_1^2) + (1-\lambda)\mathscr{N}(y_i|\mu_2,\sigma_2^2) \qquad (10.1)$$

where (μ_j,σ_j) $j = 1,2$ are the location (or center) and scale (or width) parameters of the two Gaussians with densities

$$\mathscr{N}(y_i|\mu_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right],$$

λ and $1-\lambda$ are the proportions of the two components and y_i is datum i. Figure 10.1 shows a two dimensional example. Even by eye, one should be able to determine, with some accuracy, both the centers and the width of the two Gaussians. We now show that, instead, the maximum likelihood estimators of these parameters fail miserably to find reasonable estimates, in spite of the problem's simplicity.

As repeatedly mentioned, the likelihood of independently and identically distributed data (such as the data considered in this example) is given by the product, over the data y_i, of the terms in Eq. (10.1):

$$p(y|\mu_1,\sigma_1,\mu_2,\sigma_2,\lambda) = \prod_i p(y_i|\mu_1,\sigma_1,\mu_2,\sigma_2,\lambda) \qquad (10.2)$$

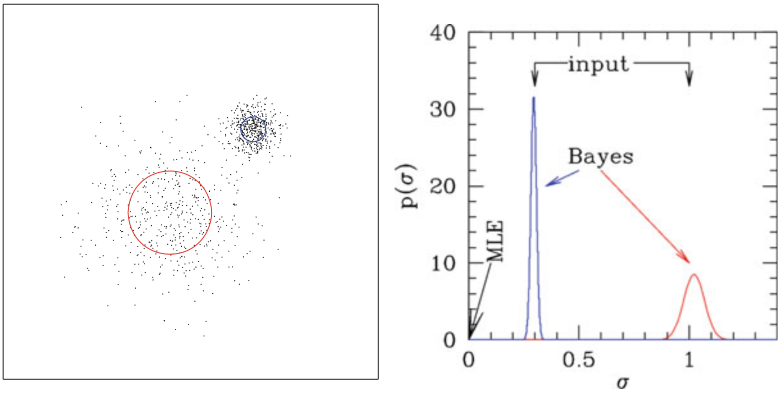

Fig. 10.1 *Left Panel*: An example of a mixture distribution in two dimensions. Readers may think that we displayed the spatial distribution of photons coming from two astronomical extended sources as observed by a perfect instrument (no background, perfect angular resolution, etc.), or the distribution of two galaxy populations in a two dimensional parameter space, or something similar. *Right panel*: true (input) values, maximum likelihood estimate (MLE), and posterior (Bayes) probability distributions for the example in the left panel. Reproduced from Andreon (2010b) with permission

The parameters that maximize the likelihood above (the "best" fit) are not near their true values (e.g., those used to draw the points in Fig. 10.1), but occur when $\mu_j = y_i$ and $\sigma_j \to 0$. In fact, the likelihood $\to \infty$ as $\sigma_j \to 0$, because the i^{th} term of Eq. (10.1) diverges and the remaining ones take a finite, non-zero, value. The maximum likelihood estimates of the parameters therefore badly fail to recover parameters for a very simple problem, simple enough that the solution may be guessed by visual inspection of the data.

The failure of maximum likelihood estimates is a general characteristic of mixtures of probability distributions of non-fixed variance (not only of Gaussians) and does not disappear "in the long run," i.e., by disposing of a sufficiently large sample (that we don't have, Time Allocation Committees are reluctant to allocate and perhaps does not exist). On the contrary, our chances of failure increase with sample size, because there is an increasing number of values for which the likelihood goes to infinity (one per datum).

Therefore, maximizing the likelihood, even for unlimited data, does not return the size of two astronomical sources, or the velocity dispersion of a cluster in the presence of interlopers, or many other quantities in the presence of two populations or signals (or an interesting and an uninteresting population or signal). Even worse, there is no "warning bell," i.e., something that signals that something is going wrong, until an infinity is found when maximizing the likelihood. In some real applications nothing as bad as infinity appears, and thus there is no "warning bell" signaling that something has gone wrong.

The Bayesian approach is not affected by such issues: it never instructs us to maximize any unknown parameters because the sum axiom of probability tells us to sum (or integrate) over unknown quantities, so that their effect is averaged over all plausible values. From a mathematical point of view, the (marginal!) posterior probability distribution of, say, σ_1 is given by the integral over μ_1, μ_2, and σ_2. In formulae,

$$p(\sigma|y) = \int\int\int p(\sigma_1, \sigma_2, \mu_1, \mu_2|y) d\mu_1 d\mu_2 d\sigma_2.$$

This can be decomposed as the sum of the integrals over $\mu_1 = y_i$ and over $\mu_1 \neq y_i$. The former integral is zero because the support has zero length. The integrand of the latter has the form $\frac{1}{\sigma}\exp\left(-c/\sigma^2\right)$ with $c \neq 0$ (assuming uniform priors), and therefore goes to zero when $\sigma \to 0$ because the exponential function converges to 0 faster than the function $\frac{1}{\sigma}$ diverges to infinity.[2]

More simply, the posterior probability distribution derived following Sect. 6.2 is shown in the right panel of Fig. 10.1 (we adopted a uniform prior over a large interval). The posterior is well behaved and it is centered on the values used to generate the data. The likelihood, instead, has hundreds of points of singularity, one per datum, all at $\sigma = 0$. Indeed, with a minimal knowledge of the statistical literature, the failure of the maximum likelihood estimate could have been anticipated: the just found maximum is not a maximum as intended by mathematicians: to have a maximum a function should be at least twice differentiable, have zero first derivative, and negative second derivative. Manifestly, $\sigma = 0$ cannot be a maximum because the first derivative is not zero (the likelihood is diverging, not flat, like in a maximum!). How many of you, gentle readers, have noted the above (before our comment) and regularly check the regularity conditions used for the application of maximum likelihood methods?

There are many such cases where the maximum likelihood approach fails, for example: when the physicist-maximum falls close to the boundary of the parameter space, when one measures zero (galaxies of a given type, photons in a given band, decays compatible with some criteria) out of n trials, when the maximum likelihood estimate is close to the boundary (e.g., the best estimate of the intensity of an *emission* line is near to zero), or in the presence of upper limits (also in such a case the likelihood shows a non math-maximum). In mixture settings (i.e., when something is modeled as a sum of two contributions), one may have the same problem if the maximum likelihood estimate of the proportion of the two components is zero or poorly determined and near to one of the extremes (0 or 1). Finally, as detailed below, in fitting a regression on data with errors in the presence of an intrinsic scatter, maximum likelihood estimates are non-unique (because the number of variables to be fitted is larger than the number of data) and, in such conditions the likelihood is far from being Gaussian, and will never become as such even with increasing the sample size.

[2] We thank Gianluigi Filippelli for help with this computation.

10.2.2 Small Samples

Astronomers and physicists are often faced with computing an average by combining a small number of estimates, or fit a trend or a function disposing of just a few data points. We almost always start with a few measurements of an interesting quantity, say the rate at which the galaxy's mass increases. In most of these cases, the measurand may be parametrized in several ways. For example, if the aim is to measure the relative evolution of luminous (L) and faint (F) red galaxies, a central topic on galaxy evolutionary studies, should we study L/F, F/L, $L/(L+F)$ or perhaps the selected parametrization is not relevant? All of these parametrization have been adopted in recent astronomical papers (and the first author, in Andreon 2008, took one more different parametrization!). Specific star formation rates (sSFR) and e-folding times, τ, are approximately reciprocal measures (long e-folding times correspond to small sSFR). To the authors' knowledge, none of the parameterizations have a special status: there is not a physical reason behind the choice of any of them.

Unfortunately, when the sample size is small, results obtained using commonly used formulae (e.g., weighted average, best fit, etc.) do depend on the adopted parameteric form. For example, an average value, computed by a weighted sum, or a fit performed by minimizing the χ^2, has a special meaning, because the result depends on which parametrization is being adopted. In the Bayesian approach this situation does not occur.

As an example, let us consider two data points for the sake of clarity, $(f/l)_1 = 3 \pm 0.9$ and $(f/l)_2 = 0.3333 \pm 0.1$. The small sample is taken to allow the reader to easily follow the algebra, but the result is general and valid for larger sample sizes. The error weighted average, which is the maximum likelihood estimate of the averages, $\langle f/l \rangle$ is 0.37. The reciprocal values ($(l/f)_i = 1/(f/l)_i$; 0.3333 ± 0.1 and 3 ± 0.9) have error weighted averages equal again to 0.37, fairly different from the reciprocal of $\langle f/l \rangle$, $1/0.37 = 2.7$. Therefore $\langle f/l \rangle \neq 1/\langle l/f \rangle$, and they differ by much more than their error (obviously, the error on the mean is smaller than the error on any data point). At first sight, by choosing the variable parametrization, the researcher may select the result he/she wants, a situation surely not recommended by the scientific method. Similar problems are present with two data points differing by just 1σ, or in general with small samples.

Why does such a strange situation occur? It comes from the freedom, in non-Bayesian settings, of choosing an estimator of the measurand ($\langle f/l \rangle$ or $1/\langle l/f \rangle$ for example). Most estimators (satisfying certain conditions) will asymptotically converge on the true value of the parameters, but without any assurance, however, that such a regime is reached with the sample size in hand. Until this regime is reached, different estimators will return different numbers. Bayesian methods do not present this shortcoming, because they already hold with $n = 2$ and do not pass through the intermediate and non-unique step of building an estimator of the measurand. On the other hand, they require a prior for the quantity being measured, and this may, or may be not, be easily available (or the exact form of the prior may be disputed amongst a group of researchers).

One may argue that in this example, the number of points is so small that no one will likely make an average of them. However, one almost always starts by averaging two or three values. Furthermore, the same problem occurs when one fits a function using a number of data points only slightly exceeding the number of parameters. Often, few points are the result of a large observational effort (and obtained through analyzing thousands of galaxies, as in the example above) and it is very hard (when not impossible) to assemble a larger sample, and thus the average of a few numbers is almost all we can do. For example, how many estimations of SFR at $z \sim 8$ or of the Higgs mass exist?[3] Should we not then combine the very few observations available in some way to utilize the information within them? Small sample problems are often "hidden" in large samples: even the largest astronomical surveys, including hundreds of thousand of galaxies, estimate galaxy densities using sub-samples equivalent to just one to ten galaxies.

To summarize, when dealing with small samples, it is not guaranteed that the asymptotic convergence of the maximum likelihood estimate has occurred with the sample in hand. Honestly, how many of you, gentle researchers, check, before performing a fit or an average, if the sample size is large enough to perform inference based on the fit or average? How many of you regularly check that the regularity conditions requested for the use of maximum likelihood estimates? Would it not be better to adopt a method that is almost *always* safe, even when one has a small sample?

10.3 Robust Estimates of Location and Scale

The estimation of location and scale (spread) parameters of a distribution, contaminated by some unwanted events, is a common astronomical exercise. In this book we addressed this case a number of times: in the case of globular cluster metallicities (Sect. 6.2.1), in the average of incompatible measurements (Sect. 6.2.2) and in a few examples in Chap. 7. When faced with this problem, many astronomers adopt the non-Bayesian method introduced in astronomy by Beers et al. (1991) and used/cited in over 500 peer-reviewed articles. Beers et al. (1991) present estimators (e.g., of the standard deviation), classified as "robust" because they are minimally affected by the presence of a contaminating population. Basically, the robust estimator down weights data in the tails of the observed distribution to estimate location (center) and scale (sigma).

We now compare the performances of the "robust" method and the Bayesian approach, largely drawing ideas from Andreon et al. (2008) and Andreon (2010b), and where comparisons in more realistic cases are presented. To focus this argument, we will consider simulated data for use in the determination of the velocity dispersion of galaxy clusters (addressed in Sect.7.3).

[3] If the answer is "many," you are reading the book well after it was written. Choose then a different redshift, where only a very few SFR estimates exist.

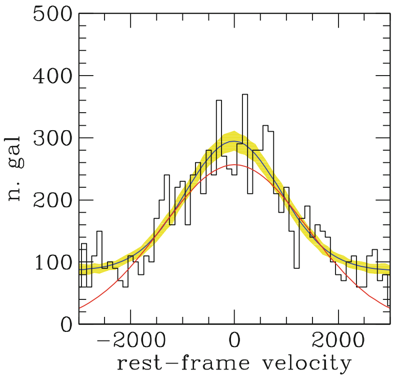

Fig. 10.2 Velocity histogram for a simulated data set. The *red curve* is the cluster's velocity estimate derived from robust statistics, whereas the *blue curve* and the *shaded yellow region* show the Bayesian estimate: $\sigma_v = 940 \pm 85$ km/s. Reproduced from Andreon et al. (2008) with permission

An ideal estimate of the velocity dispersion

- would scatter around the true value (i.e., show no bias)
- would scatter by about the quoted error (i.e., return fair error estimates)
- would return an estimate of the error with little scatter (i.e., return errors with little noise). This property is important every time these numbers are used. For example, if one combines two estimates by any sort of weighted average, with weight given by the errors (squared), it is important that no data point overdominates the average only because it has an overly optimistic estimated error as a result of the error's noisiness.

10.3.1 *Bayes Has a Lower Bias*

To quantify the amplitude of the bias, we consider a large sample (small errors help to see biases). Let us consider a simulated, Gaussian distributed sample of 500 (cluster) galaxy velocities having $\sigma = 1000$ km/s superposed over a background of 500 uniformly distributed galaxy velocities within ±3000 km/s. Note that within 1000 km/s from the cluster's center there are on average $500 \cdot 0.68 = 340$ members and $500/3 = 83$ background galaxies, i.e., the contamination here is just 20 %, and the "robust estimator" handles such cases.

The data are generated by:

```
model{
vcl ~ dnorm(10000,pow(1000,-2))
vbkg ~ dunif(7000,13000)
},
```

where we set the cluster velocity to 10000 km/s (but results are independent of this assumption). The data are displayed in the histogram in the left panel of Fig. 10.2.

Applying the methods of Beers et al. (1991) yields $\widehat{\sigma_v} = 1400$ km/s with a negligible error estimate, when the input velocity dispersion is $\sigma = 1000$ km/s.

Using the model described in Sect. 7.3, which fits a sum of two Gaussian to the data, we find that the posterior mean of the velocity dispersion is 940 ± 85 km/s, close to, and in agreement with, the value used to generate the data. This simulation shows the presence of a systematic bias in the Beers et al. estimator, even for a large sample. Actually, the bias is independent of the sample size, and its amplitude depends on the relative fraction of clusters and interlopers. While the robust estimate is less biased than what it may be by fitting a Gaussian distribution to the Gaussian+Uniform distribution, it is more biased than making the right assumption, i.e., that we are observing two populations, a mixture, which is what the Bayesian approach does. The Bayesian approach models the mixture instead of down weighting the data in the tails (as the robust approach does) and for this reason it shows a much lower bias (if any at all).

10.3.2 Bayes Is Fairer and Has Less Noisy Errors

The purpose of this section is to assess which method returns more realistic and less noisy errors.

We begin by simulating 1000 (virtual) clusters of 25 members drawn from a Gaussian with $\sigma_v = 1000$ km/s. We allow no background in this numerical experiment, otherwise the Robust estimate would be biased (this choice, of course, penalizes the Bayesian approach, not suffering from this limitation).

For each data set, the robust method outputs its estimate of the velocity dispersion error. In contrast, the Bayesian method returns probability distributions (i.e., the posterior), and we choose as our error estimate the posterior standard deviation.

These two estimates of the errors are displayed as histograms in Fig. 10.3. One may immediately note that the width of the histogram of the Bayesian's quoted error is narrower than the corresponding Robust histogram: the quoted error is noisier in the Robust approach than in the Bayesian (70 km/s vs. 18 km/s).

Of course we would like to have less noisy errors, but their (i.e., the less noisy errors) utility may be completely offset if the quoted errors are wrong (imagine if a program always returns "42," i.e., with no scatter, as an error estimate: the quoted error would be noiseless, but also useless). This is not the case for the Bayesian estimate: the scatter, from data set to data set, of the posterior mean is 185 km/s vs. a mean quoted error of 163 km/s. Instead, the robust estimator returns velocity dispersions which scatter, from data set to data set by 220 km/s, but claim his errors are smaller, on average, than the mean quoted error most of the time (see Fig. 10.3). In other words, Robust errors are noisier and underestimated.

A researcher using the Robust approach will believe that his measurement is accurate, when instead the error is larger than what is quoted most of the time, and, at the same time, will not take full profit of the data (as if he had used the more precise Bayesian estimate).

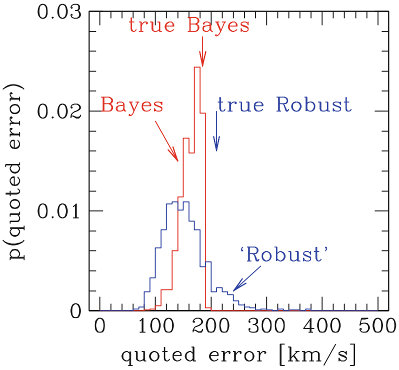

Fig. 10.3 Comparison between the distributions of the quoted error (histograms) by "Robust"' (biweight estimator of scale, *in blue*) and by our Bayesian method (*in red*) for 1,000 simulations of 25 galaxies uncontaminated by background. The "Robust" error estimate is noisier (the histogram is wider) and somewhat biased, because the histogram is mostly on the left of the true error (given by the standard deviation of returned velocity dispersions). Bayesian errors (posterior standard deviation) are less biased and show a lower scatter. Reproduced from Andreon (2010b) with permission

10.4 Comparison of Fitting Methods[4]

We now illustrate model fitting in astronomical or physics journals. As mentioned earlier, we will continue our trend of "fairness" by considering another unrealistically simple example (Gaussian homoscedastic with no errors, no mixtures, no selection effects, only linear relations, sometimes no intrinsic scatter) in order to use non-Bayesian methods in a situation where they are most valid, hence penalizing a priori Bayesian methods. If you are now smiling because nothing that simple exists in your domain of research, well, the same is true in ours but nevertheless, even the simplest of applications can still teach us valuable lessons (however, if you prefer to just opt out for the Bayesian method, you can abandon reading the rest of this chapter)!

10.4.1 Fitting Methods Generalities

In astronomy and physics, the methods most widely used for fitting a function to data points are (see, e.g., Andreon and Hurn 2013):

- Ordinary Least Square (OLS) methods minimize the sum of squared residuals (i.e., the difference between the observed data and the values fitted to each point by the model).
- Weighted least square (χ^2) methods minimize a weighted sum of squared residuals, with weights chosen to give a larger weight to less noisy data.

[4] Some of the material of this section has been drawn from Andreon and Hurn (2013).

- Maximum likelihood estimation maximize the likelihood of the data. Rigorously speaking, as soon as data have errors, the number of unknown (parameters) to be fitted exceeds the number of data: for example, a linear regression model with errors on the N `x[i]` values has $N+2$ unknown (the N true `x[i]` values, plus slope and intercept), which implies that the maximum likelihood estimate is not unique. Furthermore, in such conditions, the appealing asymptotical properties of maximum likelihood estimates (those used to derive errors) do not hold. Some assumptions need to be taken to use maximum likelihood estimates, and here we marginalize the likelihood over the nuisance `x[i]`, i.e., we give them a Bayesian flavor. In the presence of an intrinsic scatter, the likelihood cannot be written as, and is different from, $e^{-\chi^2/2}$. Therefore, the maximum likelihood estimates will differ, in general, from the minimal χ^2 estimate.
- Bivariate Correlated Errors and intrinsic Scatter (BCES, Akritas and Bershady 1996) is an extension of OLS to the case of data with Gaussian errors and a linear regression with intrinsic scatter. It only deals with a single predictor, i.e., it cannot fit multiple variables, for example $z = a*x + b*y + c$.
- Bayesian methods described in Chap. 8.

Survival or reliability methods are also sometimes used, however we note that they are, most of the time, misused because these deal with hard censoring (e.g., data below a threshold are all available but with missing value) but are used to fit data with truncated (e.g., data below a threshold are completely missing, including how many datum points are truncated) and often a soft (probabilistic) truncation. In this chapter, we do not want to misuse these methods, and are therefore not considered.

10.4.2 Regressions Without Intrinsic Scatter

Let us consider the following simple case: we generate true values x from a uniform distribution between −3 and 3 and we assume y=x. Measurement errors are Gaussian, with $\sigma^2 = 1$. For the time being, we assume no intrinsic scatter between x and y.

To generate these observed and true values we use the model:

```
model {
x ~ dunif(-3,3)
y <- x + 0.
obsx ~ dnorm(x,1)
obsy ~ dnorm(y,1)
}
```

and we save 10000 element of the chain (to have a large sample that speaks by itself), part of which is used for fitting the relation, and the other part estimates the performances of the fitted models.

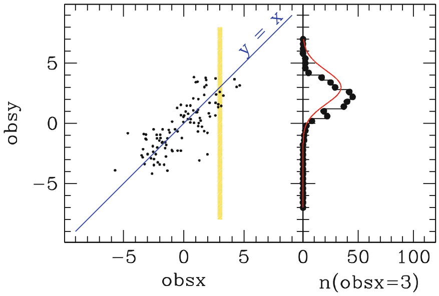

Fig. 10.4 *Left panel:* 100 points drawn from a y=x regression with no intrinsic scatter. The *yellow vertical stripe* captures those obsy for which x is close to 3. *Right panel:* Distribution of the obsy values for obsx values in a narrow band of x centered on 3, as shaded in the left panel. The *red curve* is a Gaussian centered on 3

10.4.2.1 Preamble: Restating the Obvious

We have already shown, in Sect. 8.1.3.1 that the naive prediction is an under-optimal estimate of the predicted value. We repeat here the argument once more for data drawn from the current model, because it is very important to understand the difference between the line most useful for describing the trend and the one most useful for prediction.

The left panel of Fig. 10.4 shows obsy vs obsx, after centimating it, i.e., removing 99 points out of 100 to limit crowding. The plotted line has slope 1 and intercept zero, it is the line used to generate the data. This line seems, by eye, to be an acceptable description of the plotted data: it does not look odd for any reason (especially because *we* simulated data from such a line; in research, Nature generates the data and we cannot be similarly reassured). If we consider a narrow range in obsx, say obsx$\sim$ 3 for explanational purposes only, the histogram of the obsy is not centered on 3, i.e., on the value predicted by y=obsx when obsx= 3. Instead, the histogram of obsy is clearly shifted towards a lower value. This shift is present for every obsx$\neq$ 0 value. Of course, to reduce uncertainty, the histograms are computed using all the points near obsx$\sim$ 3, not just the few plotted in the left panel.

Figure 10.5 replaces obsy with y, i.e., shows the true y values. In research activities, y is not observable, and therefore this plot will never appear on a scientific paper. When obsx= 3, the histogram of the appropriate y values is not centered on 3, i.e., on the value predicted by y=obsx. Again, the average y value of the data having obsx= 3 is 2.1, not 3. By scanning obsx on a grid of values, we build the right panel that shows the difference between the true value y given its corresponding obsx, and the value predicted by y= 1$*$obsx, i.e. the residuals <y|obsx> $-1*$obsx. As the figure shows, the line with slope 1 overestimates y at large values, and underestimates y at small values.

The left panel of Fig. 10.5 shows the source of the "problem" for this example: points at large obsx values are more likely x<obsx than x$\approx$obsx, as fairly obvi-

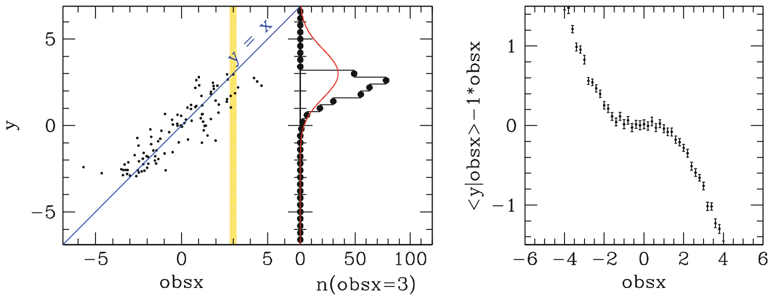

Fig. 10.5 *Left Panel:* As previous figure, but now the unobservable y is plotted in abscissa. *Right panel:* plot of the mean of y at a given obsx, after removing a slope 1 dependence

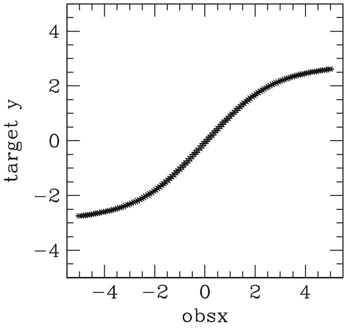

Fig. 10.6 Target value, per small bins in obsx, that regression models should predict as a function of obsx. The line bends at the extremes because y never exceeds the [−3,3] range by definition

ous for every obsx> 3 datum (all x< 3 by definition). Since y=x, then y<obsx. The same effect occurs (mutatis mutandis) at the other end of the obsx range. With a little thinking, all the range obsx> $3-3$ ($3=3\sigma$, for $\sigma=1$) and obsx< $-3+3$, i.e., all obsx values, is/are affected, although with a different degree of importance (larger at the extremes, lower in the middle). Astronomers in the audience may recognize a sort of Malmquist bias. Readers of this book may notice that this has been a reoccurring theme.

Figure 10.6 shows target y values per small bins across obsx that all regression models should predict as a function of obsx. The line bends at the extremes because y never exceeds the [−3,3] range by definition, while obsx values go outside this interval because of errors.

10.4.2.2 Testing How Fitting Models Perform for a Regression Without Intrinsic Scatter

Let us compare the performances of various fitting algorithms with this simple example.

We first fit the sample of 100 elements shown in Fig. 10.4. We present these 100 obsx and obsy and their errors (all equal to one) to all considered fitting algorithms. Non-Bayesian fitting algorithms return the slope $\widehat{\mathtt{beta}}$ and intercept $\widehat{\mathtt{alpha}}$ (and errors, usually) of the best fit regression, and we use them to compute the predicted value $\tilde{\mathrm{y}} = \widehat{\mathtt{beta}} * \mathtt{obsx} + \widehat{\mathtt{alpha}}$ of a testing sample of 10^4 obsx values drawn from the same model. The Bayesian model (listed below) directly returns $\tilde{\mathrm{y}}$ for the 10^4 data points having only obsx (and error).

Let us see which fitted model best predicts y for the testing sample.

- Bayes. For the Bayesian model, we assumed a uniform prior on the angle b, which is mathematically equivalent to a Student-t distribution on the angular coefficient beta, and a uniform prior on the intercept alpha (formally, a Gaussian of large σ).

```
model {
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
for (i in 1:length(obsx)) {
 obsx[i] ~ dnorm(x[i],1)
  x[i] ~ dunif(-3,3)
  # modeling y
  obsy[i] ~ dnorm(y[i],1)
  # modeling y-x
  y[i] <- alpha+beta*(x[i]-0.)
  }
}.
```

The fit to the data gives:

$$\mathtt{y} = (1.02 \pm 0.07)\,\mathtt{x} - 0.07 \pm 0.12 \tag{10.3}$$

i.e., the Bayesian fitting recovers within the errors the input regression with slope one and intercept zero.

As mentioned, to predict $\tilde{\mathrm{y}}$ for the testing data of size 10^4 with obsx, we use the posterior predictive distribution. In practice, we list these 10^4 obsx values in the data file and we put NA (for Not-Available) at the place of the non-available obsy. The top-left panel of Fig. 10.7 shows the residuals (true y minus predicted values) for the Bayesian model. The model performs very well, yields small residuals, and does not exhibit any major trend with obsx.

- Naive, $\tilde{y} = obsx$

Suppose, for example, that theory tell us that $y = x$, and that someone will attempt to predict $\tilde{y}$ as $\tilde{y} = obsx$. This is a bad idea, recall the previous section. As already shown in the right panel of Fig. 10.5, this way of predicting y performs poorly, re-shown in Fig. 10.7 on the same scale of other fitting algorithms.

- OLS, Least-square

The least-square fit ignores the difference between observed and true values and finds $\hat{a}$ and $\hat{b}$ that minimize the sum of the square of the residuals between obsy and $\widehat{\mathtt{beta}}*\mathtt{obsx} + \widehat{\mathtt{alpha}}$. On the sample of 100 obsx and obsy gives:

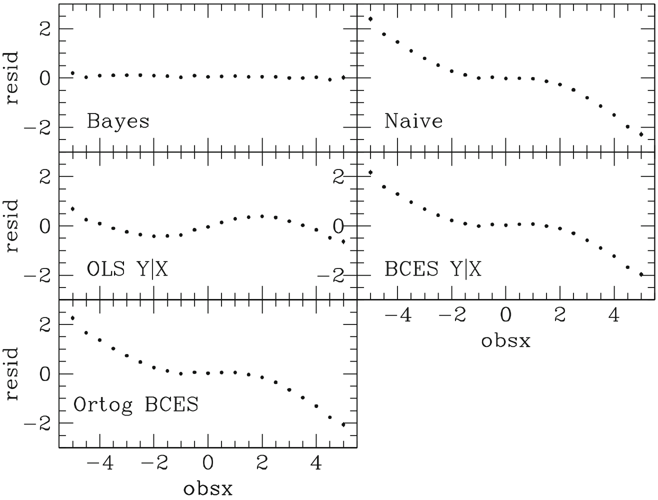

Fig. 10.7 Performances of various candidate models in predicting the true value y. On the abscissa we plot residuals (true y minus predicted values, for small ranges in obsx). The model plotted in the bottom panel is not recommended for predictions, but unfortunately is widely used

$$\texttt{obsy} = 0.74 * \texttt{obsx} - 0.09 \tag{10.4}$$

The slope is well below 1, i.e., the recovered slope is tilted away from the true slope. The center-left panel of Fig. 10.7 shows that this model performs worse than the Bayesian one, but still better than freezing the slope to the true value (one, i.e., case b). Using least-squares, the price to pay to have better predictions is a bias in the slope estimate.

Some researchers, feeling that this fitted line is not steep enough to capture the data's trend, sometimes combine this slope with the one derived from swapping obsx with obsy, obtaining a fitting line with a slope near one, as in our case b). The performance of this further fitted line is therefore near to the, already discussed, case b).

- BCES $Y|X$

  This model is one of the most used in astronomy. Fitting this model on the 100 obsx and obsy pairs, the $Y|X$ fit is:

  $$\texttt{y} = (0.95 \pm 0.08) * \texttt{x} - 0.05 \pm 0.15 \tag{10.5}$$

  BCES estimated parameters recover the input parameters (slope one and intercept zero) within the errors.

  However, the bad news is that for *not* having a biased slope, BCES performance in predicting y is very bad, as for the naive fit, as shown in Fig. 10.7.

- BCES orthogonal and bisector.

Table 10.1 Summary table of comparisons for the regression without intrinsic scatter

Method	Parameters	Predicted values
Bayes	Good	Best
Naive	Assumed known	Very bad
OLS	Biased	Intermediate
BCES Y—X	Good	Very bad
Ortog BCES	Good	Very bad

Akritas and Bershady (1996) offer several fitting models, among which are the orthogonal and bisector BCES fits. The former minimizes residuals orthogonally to the regressed line, the latter returns the bisector of the $Y|X$ fit and the fit obtained swapping `obsx` with `obsy` (BCES $X|Y$).

The BCES orthogonal and bisector fits, for the sample of 100 data `obsx` with `obsy`, are identical:

$$\mathtt{y} = (0.96 \pm 0.08) * \mathtt{x} - 0.05 \pm 0.15 \tag{10.6}$$

The slope and intercept of the input regression is recovered within the errors, and, as a consequence, these fits performances in predicting $\tilde{\mathtt{y}}$ are very bad, as shown in the bottom panel of Fig. 10.7.

The performances of the various methods are summarized in Table 10.1. In summary, for the considered example, non-Bayesian fits either return a good value for the slope and a very bad $\tilde{\mathtt{y}}$ (e.g., naive or BCES fits), or a better, but still poor, $\tilde{\mathtt{y}}$ for a biased value of the slope `beta` (e.g., OLS fit). The Bayesian model returns the right slope with very good $\tilde{\mathtt{y}}$. We leave you, the researcher, with the choice of deciding which fitting model you want to adopt.

The reason why the Bayesian model performs better than other considered methods should be, by now, clear. As shown in Fig. 10.6, the relation between `y` and `obsx` is non-linear, even if the relation between `y` and `x` is linear. This immediately tells us that a method that adopts a linear scaling between `y` and `obsx`, or use it to predict `y` from `obsx`, will perform worse than one that is able to adapt itself to fit this curved relation. The Bayesian way does not strictly follow this path, and this is advantageous, because real samples are often small, and therefore the shape of the curved line is imprecisely determined when one attempts to fit `y` vs `obsx`. The Bayesian fitting method assumes a linear model between true values (`x` and `y`), not between observed values (`obsx` and `obsy`). It allows a non-linear relation between `y` vs `obsx`, but does not require the determination of supplementary coefficients, because the relation is fitted in the true data space. The Bayesian model therefore works better than every competing model attempting to fit a straight line (or to predict $\tilde{\mathtt{y}}$ values using a straight line) on observed values.

If we were talking about predicting (observationally expensive) masses ($\tilde{\mathtt{y}}$) from (observationally parsimonious) mass proxies (`obsx`, such as richness or X-ray luminosity), non-Bayesian fitting models are either wrong in predicting the trend between mass and proxy (i.e., returns a wrong slope of the `y` vs `x` relation) or perform poorly in predicting masses from observed values of mass proxies. The Bayesian fit returns the right (input) slope and the best predicted masses.

10.4.3 One More Comparison, with Different Data Structures

In this section we test whether the results found in the previous section are unique to the adopted data structure (data were drawn from a truncated uniform distribution, often named “top-hat” in astronomy and physics) or are more general. We also consider other fitting algorithms, and, in the case of the Bayesian approach, a few priors that do not match the data structure in location, width, asymmetry, and importance of tails. The purpose of these priors is to simulate the situation in which the prior is imprecisely known.

We generate 100 values of x from a Student-t distribution with 10 degrees of freedom centered on 0.1 and with scale 1.1, x[i]∼dt(0.1,pow(1.1,-2),10), these x are then perturbed by homoscedastic Gaussian noise with $\sigma^2 = 1$ to give 100 obsx. As in previous example, y=x with no intrinsic scatter, and these y are also perturbed by homoscedastic Gaussian noise with $\sigma^2 = 1$ giving 100 obsy. In practice, the JAGS code used to generate the sample is:

```
model {
 x ~ dt(0.1,pow(1.1,-2),10)
 y <- 1.0 * x + 0.0
 obsx ~ dnorm(x,1)
 obsy ~ dnorm(y,1)
}.
```

The resulting 100 (obsx,obsy) pairs are used as the data for fitting the line $\mathtt{y}= a*\mathtt{x}+b$ using various fitting models.

In addition to the Bayesian model fully described below, and the least square and BCES considered in the previous section, we also consider the χ^2 and the MLE fits. For the MLE, we adopt a Bayesian flavor, because as already remarked there is no single maximum likelihood estimate (the number of parameters, 102, is larger than the number of data points, 100). We marginalize over the 100 obsx[i] and adopt the marginal likelihood (D’Agostini 2005)

$$L(a,b) = \frac{(2\pi)^{-N/2}}{\prod_{i=1}^{N}\sqrt{\sigma_y^2 + a^2\sigma_x^2}} \exp\left[-\frac{\sum_{i=1}^{N}(y_i^{obs} - ax_i^{obs} - b)^2}{2(\sigma_y^2 + a^2\sigma_x^2)}\right]. \tag{10.7}$$

Note the χ^2 in the exponential term and, at the same time, the a dependence in the leading non-exponential term which makes χ^2 and maximum likelihood different. Therefore, when there are errors on the predictor quantity, χ^2 is not the MLE estimate, as already remarked.

Concerning the Bayesian modeling, we consider several options for the prior.

1. In some astronomical problems, the x data structure is fairly well known (e.g., from previous experiments or from theory). If this were the case here, the prior would be

   ```
   x[i] ~ dt(0.1,pow(1.1,-2),10)
   ```

   since this is actually the true population model.
2. Of course it may well be the case that we do not have such a high level of prior knowledge and so we also consider an alternative prior

   ```
   x[i] ~ dnorm(0,1)
   ```

   which has a difference in location and scale as well as a lighter tail behavior, all of which are providing examples of a possible mismatch between the true data structure and what is known about it. This prior is still relatively informative; it might perhaps be arrived at in a not strictly Bayesian sense by looking at a histogram of the 100 obsx[i] values.
3. Finally, we also consider a less informative prior, with parameters to be determined at the same time as the other regression parameters; we have adopted a Gaussian prior with parameters treated as additional unknowns

   ```
   x[i] ~ dnorm(mu,prec)
   ```

   with weak hyperpriors on the parameters

   ```
   mu ~ dnorm(0,1.0E-4)
   prec ~ dunif(0.0,10.0) .
   ```

Therefore, the model reads:

```
model {
 alpha ~ dnorm(0.0,1.0E-4)
 beta ~ dt(0,1,1)
 for (i in 1:length(obsx))
 {
  obsx[i] ~ dnorm(x[i],prec.obsx[i])
  obsy[i] ~ dnorm(y[i],prec.obsy[i])
  y[i] <- alpha+beta*x[i]
# t prior for the x population OR
  x[i] ~ dt(0.1,pow(1.1,-2),10)
# N(0,1) prior for the x population OR
  x[i] ~ dnorm(0,1)
# Normal prior for the x population with hyperparameters
  x[i] ~ dnorm(mu,prec)
 }
  mu ~ dnorm(0,1.0E-4)
  prec ~ dunif(0.0,10.0)
} .
```

Now, we are ready to test how well different methods perform in recovering the regression parameters and in predicting y. For the latter we used a sample of 10000 obsx generated from the true model, as in previous section. Figure 10.8 summarizes our tests:

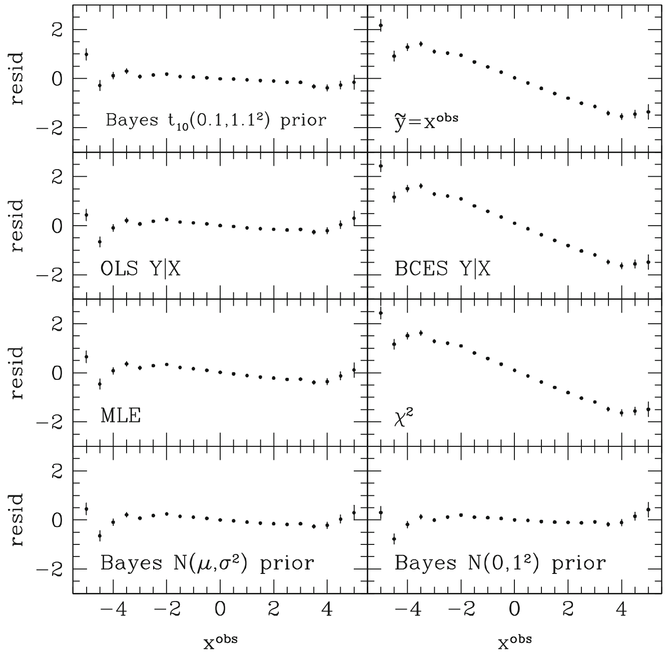

Fig. 10.8 Performance of the various approaches in predicting the true value y given obsx. Each point shows the average residual per small bin of obsx. The error bars indicate the standard error of the mean residuals. Reproduced from Andreon and Hurn (2013) with permission

Top row, left. Bayes regression with a dt(0.1,pow(1.1,-2),10) prior performs very well with small residuals across the entire range of obsx as is to be expected given that the correct model is being fitted. The point estimates of the parameters themselves are 1.09 and 0.00 for the slope and intercept, respectively.

Top row, right. Suppose that one naively attempts to predict ỹ by ỹ =obsx, in effect making use of the knowledge that y=x, but ignoring the noise structure. The residual in this case is the perturbation x−obsx, i.e., a dnorm(0,1) value in this example, however, as a result, it is negatively correlated with obsx so that the largest residuals occur at the extremes of obsx

Second row, left. OLS generates better predictions than the naive estimate. It achieves this better performance by biasing the fitted slope, returning an estimated slope of 0.65 in place of the input value of 1.

Second row, right. Predicted ỹ values computed using the BCES method are similar to those using the naive ỹ =obsx method. By improving the estimates of the slope and intercept, the prediction performance is degraded.

Third row, left. MLE Bayesian-flavored approach returns reasonable predictions and a biased slope, as OLS.

Third row, right. χ^2 performs less well than MLE and differs from it because MLE has a leading term missing in the χ^2 expression.

Bottom row, left and right. Bayesian errors-in-variables regression with either a `dnorm(mu,prec)` prior for the `x[i]` or a `dnorm(0,1)` prior performs almost as well as the Bayesian model using the true distribution as a prior, showing that an accurate knowledge of the prior is not necessary to achieve good performance in this example.

To summarize, for this example the methods based on sound statistical models applied appropriately (MLE Bayesian-flavored, Bayesian regression) outperform the predictions of BCES and χ^2. OLS also does well here paradoxically because of its sheer simplicity (but returns a biased slope, as indeed the MLE does). This example confirms results from the previous section and shows in addition that one may achieve good prediction performances even when taking a prior for the data structure that does not match the data structure in location, width and importance of tails.

We leave to the reader the exercise of showing that similar results are also obtained for data drawn from asymmetric distributions and for regressions with heteroscedastic errors and an intrinsic scatter.

10.5 Summary and Experience of a Former Non-Bayesian Astronomer

We have illustrated that Bayesian methods may easily outperform the common non-Bayesian methods used in astronomy and physics (χ^2, maximum likelihood, OLS, etc.).

The first author was, for the first 20 years of his career, almost unaware of Bayesian methods. He used the common methods described (and criticized) in this chapter, and with some difficulties improved them (and only by a bit) to make them usable for research applications (e.g., he first published the correct likelihood expression for fitting the luminosity function in the presence of a background). Nevertheless, he prefers the Bayesian approach for various reasons:

1. It returns what the researcher is looking for: a probability distribution for the quantity being measured, allowing him to say "the Hubble constant (or the mass of the Higgs boson or whatever else) is in this range with such a probability," instead of forcing the researcher to say something alternative to that.
2. It does not require the researcher to check whether regularity conditions, necessary for the use of other methods, hold, and neither to delay the publication of a result for the moment when the sample is large enough to be in the asymptotic regime.
3. It is easy to model the complex relation between the many variables needed in research applications, and that these are often intrinsically uncertain (i.e., should be included in the model and if so in what form) and, often, stochastic.

4. The Bayesian method is effective in extracting information from the data in an optimal way (with little or no loss of information) even with a small sample size.
5. When data size increase, the learning (i.e., the way uncertainty decreases) is always optimal. In this book we never mention the word “optimal.”
6. It allows us to merge information of different types (theory, measurements, expert opinions).
7. As seen several times in this book, the Bayesian method is easy to configure and is flexible. For example, one may easily add components to a model (e.g., an intrinsic scatter, a noisy predictor) or just modify the model (change a relation from linear to quadratic, or the scatter from Gaussian to Cauchy).

As briefly mentioned, many non-Bayesian methods exist. Some of them are very good. Some of them are mathematically equivalent to the Bayesian method, and therefore it is hard to believe that their relative performances may differ. However, they are rarely used in astronomy and physics. We further emphasize that some non-Bayesian methods are often misused and sometimes their performance is criticized on ill-posed bases.

To summarize, the key advantage of the Bayesian way is just a pragmatical one: an average Bayesian astronomer can easily and quickly solve by himself a statistical problem he encounters in his research activity, finding a solution competitive with the one of a leading non-Bayesian statistician, without the need of losing a lot of his own time involved in researching the non-Bayesian literature (especially because it is written in a language that is hard-to-understand for non-statisticians). Not only is the solution competitive, but the human time required to obtain the solution (i.e., to solve the problem) is much shorter in the Bayesian paradigm. CPU computation times may be larger, however, but a computer works 24 h per day 365 days/y and potentially across many computers.

Once the basics of probability are learned (the few pages of Chap. 2), an average Bayesian astronomer/physicist can address a very complex analysis without needing to learn anything new. As mentioned in the preface, the example in Sect. 8.2 was written by a student *alone and unsupervised* after 12 h of studying the content of this book. Strictly speaking, our italic statement is correct for parameter estimation problems only. Problems involving choosing, or comparing, competing models requires learning additional theory and concepts.

References

S. Andreon. A Bayesian approach to galaxy evolution studies. *Bayesian Methods in Cosmology, edited by Michael P. Hobson, Andrew H. Jaffe, Andrew R. Liddle, Pia Mukeherjee and David Parkinson. Published: Cambridge University Press, New York, Cambridge, UK; 2010, p. 265*, page 265, 2010b.

S. Andreon. The history of mass assembly of faint red galaxies in 28 galaxy clusters since $z = 1.3$. *Monthly Notices of the Royal Astronomical Society*, 386: 1045–1052, 2008.

T. C. Beers, K. Gebhardt, W. Forman, J. P. Huchra, and C. Jones. A dynamical analysis of twelve clusters of galaxies. *The Astronomical Journal*, 102: 1581–1609, 1991.

S. Andreon, R. de Propris, E. Puddu, L. Giordano, and H. Quintana. Scaling relations of the colour-detected cluster RzCS 052 at $z = 1.016$ and some other high-redshift clusters. *Monthly Notices of the Royal Astronomical Society*, 383: 102–112, 2008.

S. Andreon and M.A. Hurn. Measurement errors and scaling relations in astrophysics: a review. *Statistical Analysis and Data Mining*, 6:15–33, 2013.

M.G. Akritas and M.A. Bershady. Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*, 470:706, 1996.

Appendix A
Probability Distributions

A.1 Discrete Distributions

A.1.1 Bernoulli

For $\mathtt{f} \in (0,1)$ and $\mathtt{obsx} \in \{0,1\}$, $\mathtt{obsx}$ follows the Bernoulli distribution, denoted as `obsx ~ dbern(f)` if its distribution is

$$p(\mathtt{obsx}) = \mathtt{f}^{\mathtt{obsx}}(1-\mathtt{f})^{1-\mathtt{obsx}}.$$

For the Bernoulli distribution, $\mathrm{E}(\mathtt{obsx}) = \mathtt{f}$ and $\mathrm{Var}(\mathtt{obsx}) = \mathtt{f} * (1-\mathtt{f})$, where $\mathrm{E}(\cdot)$ is the mean, and $\mathrm{Var}(\cdot)$ is the variance (square of the standard deviation). Notice that the Bernoulli distribution is a binomial distribution with $\mathtt{n} = 1$.

A.1.2 Binomial

For $\mathtt{f} \in (0,1)$ and $\mathtt{n} = 0,1,2,\ldots$, $\mathtt{obsx}$ follows the binomial distribution, denoted as `obsx ~ dbin(f,n)` if its distribution is

$$p(\mathtt{obsx}) = \binom{\mathtt{n}}{\mathtt{obsx}} \mathtt{f}^{\mathtt{obsx}}(1-\mathtt{f})^{\mathtt{n}-\mathtt{obsx}}$$

for $\mathtt{obsx} = 0,1,\ldots,\mathtt{n}$ and where

$$\binom{\mathtt{n}}{\mathtt{obsx}} = \frac{\mathtt{n}!}{\mathtt{obsx}!(\mathtt{n}-\mathtt{obsx})!}.$$

For the binomial distribution, $\mathrm{E}(\mathtt{obsx}) = \mathtt{n} * \mathtt{f}$ and $\mathrm{Var}(\mathtt{obsx}) = \mathtt{n} * \mathtt{f} * (1-\mathtt{f})$.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5

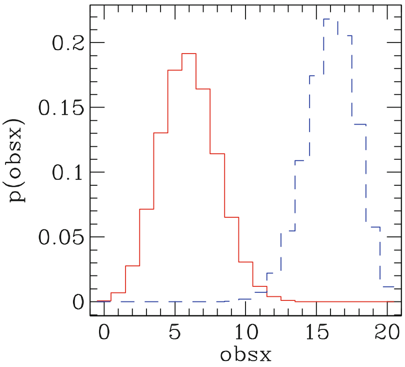

Fig. A.1 Binomial distributions both with $\texttt{n} = 20$ and with $\texttt{f} = 0.3$ (*solid line*) and $\texttt{f} = 0.8$ (*dashed line*)

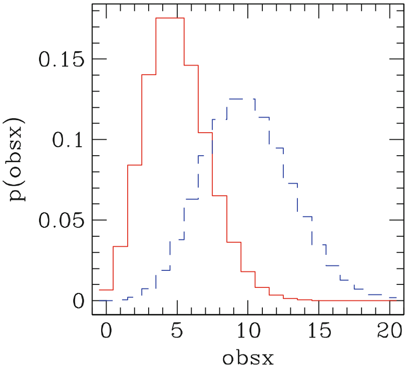

Fig. A.2 Poisson distributions with $\texttt{s} = 5$ (*solid line*) and $\texttt{s} = 10$ (*dashed line*)

Figure A.1 displays two binomial distributions both with $\texttt{n} = 20$ and with $\texttt{f} = 0.3$ (solid line) and $\texttt{f} = 0.8$ (dashed line).

A.1.3 Poisson

For $\texttt{s} > 0$, obsx follows the Poisson distribution, denoted as `obsx ~ dpois(s)` if its distribution is

$$p(\texttt{obsx}) = \frac{\texttt{s}^{\texttt{obsx}} \exp(-\texttt{s})}{\texttt{obsx}!},$$

for $\texttt{obsx} = 0, 1, \ldots$. For the Poisson distribution, $\mathrm{E}(\texttt{obsx}) = \texttt{s}$ and $\mathrm{Var}(\texttt{obsx}) = \texttt{s}$. Figure A.2 displays two Poisson distributions with $\texttt{s} = 5$ (solid line) and $\texttt{s} = 10$ (dashed line).

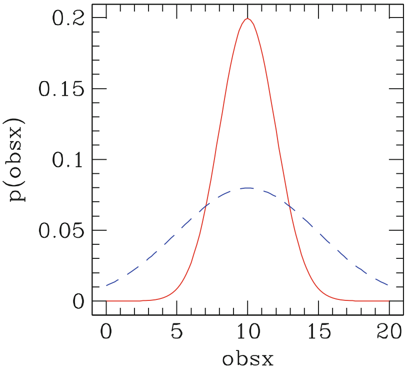

Fig. A.3 Gaussian distributions both with $\mathtt{m} = 10$ and with $\mathtt{prec} = 1/4$ (*solid line*) and $\mathtt{prec} = 1/25$ (*dashed line*)

A.2 Continuous Distributions

A.2.1 Gaussian or Normal

For $\mathtt{m}, \mathtt{obsx} \in \mathbb{R}$ and $\mathtt{prec} > 0$, obsx follows the Gaussian distribution, denoted as `obsx ~ dnorm(m,prec)` if its distribution is

$$p(\mathtt{obsx}) = \frac{\sqrt{\mathtt{prec}}}{\sqrt{2\pi}} \exp\left(-\mathtt{prec} * (\mathtt{obsx} - \mathtt{m})^2/2\right).$$

For the Gaussian distribution, $\mathrm{E}(\mathtt{obsx}) = \mathtt{m}$ and $\mathrm{Var}(\mathtt{obsx}) = 1/\mathtt{prec} = \mathtt{s}^2$. An alternative parameterization of this model is in terms of the variance $\mathtt{s}^2$,

$$p(\mathtt{obsx}) = \frac{1}{\mathtt{s}\sqrt{2\pi}} \exp\left(-\frac{(\mathtt{obsx} - \mathtt{m})^2}{2\mathtt{s}^2}\right).$$

Figure A.3 displays two Gaussian distributions both with $\mathtt{m} = 10$ and with $\mathtt{prec} = 1/4$ (solid line) and $\mathtt{prec} = 1/25$ (dashed line).

A.2.2 Beta

For $\mathtt{obsn}, \mathtt{n} > 0$, and $\mathtt{obsx} \in (0,1)$, obsx follows the beta distribution, denoted as `obsx ~ dbeta(obsn,n)` if its distribution is

$$p(\mathtt{obsx}) = \frac{1}{B(\mathtt{obsn}, \mathtt{n})} \mathtt{obsx}^{\mathtt{obsn}-1} (1 - \mathtt{obsx})^{\mathtt{n}-1}$$

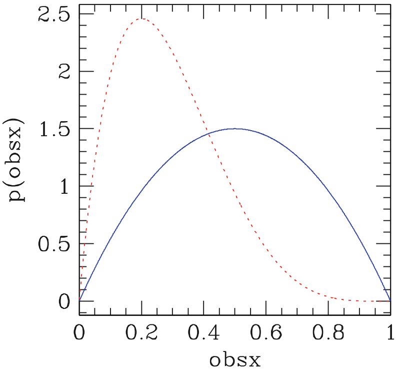

Fig. A.4 Beta distributions with $\texttt{obsn} = 2, \texttt{n} = 2$ (*solid line*) and $\texttt{obsn} = 2, \texttt{n} = 5$ (*dashed line*)

where $B(\cdot,\cdot)$ is the beta function and is defined as

$$B(\texttt{obsn},\texttt{n}) = \int_0^1 x^{\texttt{obsn}-1}(1-x)^{\texttt{n}-1}dx.$$

For the beta distribution, $\mathrm{E}(\texttt{obsx}) = \frac{\texttt{obsn}}{\texttt{obsn}+\texttt{n}}$ and

$\mathrm{Var}(\texttt{obsx}) = \frac{\texttt{obsn}*\texttt{n}}{(\texttt{obsn}+\texttt{n})^2*(\texttt{obsn}+\texttt{n}+1)}$.

Figure A.4 displays two beta distributions with $\texttt{obsn} = 2, \texttt{n} = 2$ (solid line) and $\texttt{obsn} = 2, \texttt{n} = 5$ (dashed line).

A.2.3 Exponential

For $\texttt{s}, \texttt{obsx} > 0$, obsx follows the exponential distribution, denoted as obsx ~ dexp(s) if its distribution is

$$p(\texttt{obsx}) = \texttt{s}\exp(-\texttt{s}*\texttt{obsx}).$$

For the exponential distribution, $\mathrm{E}(\texttt{obsx}) = \frac{1}{\texttt{s}}$ and $\mathrm{Var}(\texttt{s}) = \frac{1}{\texttt{s}^2}$.

Figure A.5 displays two exponential distributions with $\texttt{s} = 2$ (solid line) and $\texttt{s} = 0.5$ (dashed line).

A.2.4 Gamma and Schechter

For $\texttt{nu}, \texttt{prec}, \texttt{obsx} > 0$, obsx follows the gamma distribution, denoted as obsx ~ dgamma(nu,prec) if its distribution is

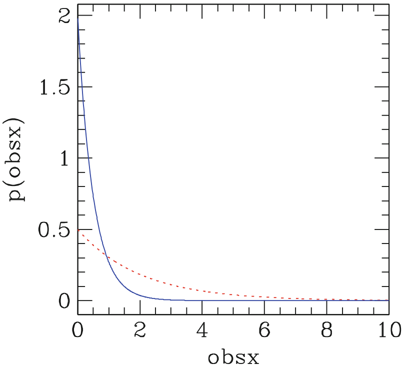

Fig. A.5 Exponential distributions with $\texttt{s} = 2$ (*solid line*) and $\texttt{s} = 0.5$ (*dashed line*)

$$p(\texttt{obsx}) = \frac{1}{\Gamma(\texttt{nu})\texttt{prec}^{\texttt{nu}}}\texttt{obsx}^{\texttt{nu}-1}\exp\left(-\frac{\texttt{obsx}}{\texttt{prec}}\right)$$

where $\Gamma(\cdot)$ is the gamma function and is defined as

$$\Gamma(y) = \int_0^\infty t^{y-1}\exp(-t)dt.$$

For the gamma distribution, E($\texttt{obsx}$)=$\texttt{nu}*\texttt{prec}$ and Var($\texttt{obsx}$)=$(\texttt{nu}*\texttt{prec})^2$. A common parameterization of this distribution, which is also implemented in JAGS, is

$$p(\texttt{obsx}) = \frac{\texttt{c}^{\texttt{nu}}}{\Gamma(\texttt{nu})}\texttt{obsx}^{\texttt{nu}-1}\exp\left(-\texttt{c}*\texttt{obsx}\right),$$

where $\texttt{c} = 1/\texttt{prec}$.

The chi-square distribution is a special case of the gamma distribution and is obtained by setting $\texttt{nu} = \texttt{v}/2$ and $\texttt{prec} = 2$. Similarly, the exponential distribution is a special case of the gamma distribution and is obtained by setting $\texttt{nu} = 1, \texttt{prec} = \texttt{s}^{-1}$.

Related to the Gamma distribution is the Schechter function. If $\texttt{obsx}$ has a gamma distribution with parameters $(\texttt{nu}, \texttt{prec})$, then the distribution for the random variable $\texttt{obsy} = \log_{10}(\texttt{obsx})$ is the Schechter distribution given as

$$p(\texttt{obsy}) = \frac{\ln(10)}{\Gamma(\texttt{nu})}10^{\texttt{nu}(\texttt{obsy}-\texttt{b})}\exp\left(-10^{\texttt{obsy}-\texttt{b}}\right)$$

where $\texttt{b} = \log_{10}(\texttt{prec})$.

The left panel of Fig. A.6 displays two gamma distributions with $\texttt{nu} = 2, \texttt{prec} = 2$ (solid line), and $\texttt{nu} = 9, \texttt{prec} = 9$ (dashed line). The right panel displays the corresponding Schechter function for these gamma distributions.

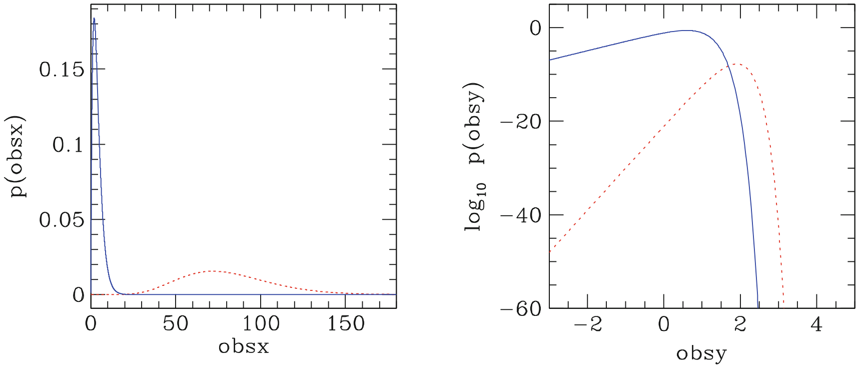

Fig. A.6 Gamma distributions (*left panel*) and Schechter functions (*right panel*) with $\texttt{nu} = 2, \texttt{prec} = 2$ (*solid line*) and $\texttt{nu} = 9, \texttt{prec} = 9$ (*dashed line*)

A.2.5 Lognormal

For $\texttt{m} \in \mathbb{R}$ and $\texttt{obsx}, \texttt{prec} > 0$, `obsx` follows the lognormal distribution, denoted as `obsx ~ dlnorm(m,prec)` if its distribution is

$$p(\texttt{obsx}) = \frac{\sqrt{\texttt{prec}}}{\texttt{obsx}\sqrt{2\pi}} \exp(-\texttt{prec} * (\ln(\texttt{obsx}) - \texttt{m})^2/2).$$

If `obsx` is a lognormal random variable with parameters `m` and `prec`, then $\log(x)$ is a normal distribution with mean equal to `m` and precision `prec`.

For the lognormal distribution, $\mathrm{E}(\texttt{obsx}) = e^{\texttt{m}-1/(2\texttt{prec})}$ and $\mathrm{Var}(\texttt{obsx}) = (e^{1/\texttt{prec}} - 1) * e^{2\texttt{m}+1/\texttt{prec}}$. An alternative parameterization of this model is

$$p(\texttt{obsx}) = \frac{1}{\texttt{obsx} * \texttt{s}\sqrt{2\pi}} \exp\left(-\frac{(\log(\texttt{obsx}) - \texttt{m})^2}{2\texttt{s}^2}\right).$$

Figure A.7 displays two lognormal distributions both with $\texttt{m} = 0$ and with $\texttt{prec} = 16$ (dashed line) and $\texttt{prec} = 1.778$ (solid line).

A.2.6 Pareto or Power Law

For, $\texttt{a}, \texttt{b} > 0$ and $\texttt{obsx} > 0$, `obsx` follows the Pareto distribution, denoted as `obsx ~ dpar(a,b)` if its distribution is

$$p(\texttt{obsx}) = \begin{cases} \texttt{a} * \texttt{b}^{\texttt{a}} \texttt{obsx}^{-(\texttt{a}+1)} & \text{if } \texttt{obsx} > \texttt{b}, \\ 0 & \text{otherwise} \end{cases}.$$

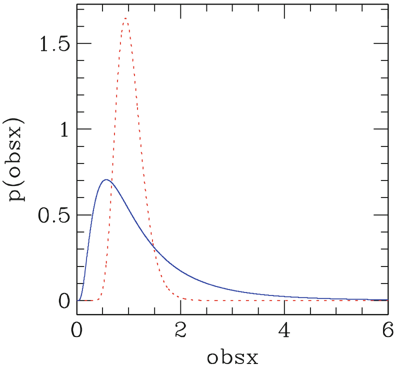

Fig. A.7 Lognormal distributions both with $\mathtt{m} = 0$ and with $\mathtt{prec} = 16$ (*dashed line*) and $\mathtt{prec} = 1.778$ (*solid line*)

For the Pareto distribution, $\mathrm{E}(\mathtt{obsx}){=}\frac{\mathtt{a*b}}{\mathtt{a-1}}$ for $\mathtt{a}{>}1$ and $\mathrm{Var}(\mathtt{obsx}){=}\frac{\mathtt{a*b}^2}{(\mathtt{a-1})^2(\mathtt{a-2})}$ for $\mathtt{a} > 2$. Note that a Pareto distribution is also known as a power-law distribution.

A.2.7 Central Student-t

For $\mathtt{nu} > 0$ and $\mathtt{obsx} \in \mathbb{R}$, obsx follows the central Student-t distribution, denoted as `obsx ~ dt(0,1,nu)` if its distribution is

$$p(\mathtt{obsx}) = \frac{\Gamma((\mathtt{nu}+1)/2)}{\Gamma(\mathtt{nu}/2)} \left(\frac{1}{\mathtt{nu}*\pi}\right)^{1/2} \left(1+\frac{\mathtt{obsx}^2}{\mathtt{nu}}\right)^{-\frac{\mathtt{nu}+1}{2}}.$$

For the central Student-t distribution, $\mathrm{E}(\mathtt{obsx}) = 0$ for $\mathtt{nu} > 1$and $\mathrm{Var}(\mathtt{obsx}) = \frac{\mathtt{nu}}{\mathtt{nu}-2}$ for $\mathtt{nu} > 2$. Note that the central Student-t distribution is similar in shape to a Gaussian distribution, only it has heavier tails. Also, as nu increases in value, the central Student-t distribution converges to a standard Gaussian distribution (mean 0 and variance 1). Figure A.8 displays two central Student-t distributions with $\mathtt{nu} = 50$ (solid line) and $\mathtt{nu} = 3$ (dashed line).

A.2.8 Uniform

For $\mathtt{a},\mathtt{b} \in \mathbb{R}$ where $\mathtt{a} < \mathtt{b}$ and for $\mathtt{obsx} \in [\mathtt{a},\mathtt{b}]$, obsx follows the uniform distribution, denoted as `obsx ~ dunif(a,b)` if its distribution is

$$p(\mathtt{obsx}) = \begin{cases} \frac{1}{\mathtt{b-a}} & \text{if } \mathtt{a} < \mathtt{obsx} < \mathtt{b}, \\ 0 & \text{otherwise} \end{cases}.$$

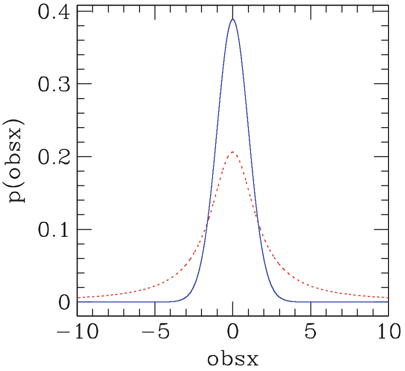

Fig. A.8 Central Student-t distributions with $\mathtt{nu} = 50$ (*solid line*) and $\mathtt{nu} = 3$ (*dashed line*)

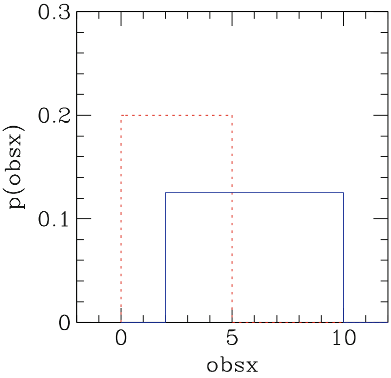

Fig. A.9 Uniform distributions with $\mathtt{a} = 2, \mathtt{b} = 10$ (*solid line*) and $\mathtt{a} = 0, \mathtt{b} = 5$ (*dashed line*)

For the uniform distribution, $\mathrm{E}(\mathtt{obsx}) = (\mathtt{a} + \mathtt{b})/2$ and $\mathrm{Var}(\mathtt{obsx}) = (\mathtt{b} - \mathtt{a})^2/12$. Figure A.9 displays two uniform distributions with $\mathtt{a} = 2, \mathtt{b} = 10$ (solid line) and $\mathtt{a} = 0, \mathtt{b} = 5$ (dashed line).

A.2.9 Weibull

For $\mathtt{a}, \mathtt{b}, \mathtt{obsx} > 0$, `obsx` follows the Weibull distribution, denoted as `obsx ~ dweib(a,b)` if its distribution is

$$p(\mathtt{obsx}) = \frac{\mathtt{a}}{\mathtt{b}} \left(\frac{\mathtt{obsx}}{\mathtt{b}} \right)^{\mathtt{a}-1} \exp((-\mathtt{obsx}/\mathtt{b})^{\mathtt{a}}).$$

For the Weibull distribution, $\mathrm{E}(\mathtt{obsx}) = \mathtt{b}\Gamma(1 + 1/\mathtt{a})$ and $\mathrm{Var}(\mathtt{obsx}) = \mathtt{b}^2\{\Gamma(1+2/\mathtt{a}) - [\Gamma(1+1/\mathtt{a})]^2\}$.

Appendix B
The Third Axiom of Probability, Conditional Probability, Independence and Conditional Independence

B.1 The Third Axiom of Probability

Readers more theoretically familiar with probability theory may feel that we are a little loose in our statement of the third axiom of probability in Sect. 2.3. Our view for this book is to present Bayesian statistics in a more intuitive fashion and so it is enough for the reader to be familiar with the axioms we have set forth. The reader interested in a more formal development of probability theory (through measure theory) should consult the books that we recommended at the beginning of Chapter 2. In these he/she will find a more formal definition, such as:

$$p(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} p(E_i).$$

for an infinite collection of disjoint sets $E_1, E_2, \ldots$ from some measure space. Our marginalization axiom stems from this more basic axiom.

B.2 Conditional Probability

For two events y and x, the conditional probability of observing the event y given that we have observed the event x is defined as

$$p(y|x) = \frac{p(y \cap x)}{p(x)}.$$

Intuitively, this definition says the conditional probability is the ratio of the probability of both events occurring $(y \cap x)$ and the marginal probability of event x.

S. Andreon, B. Weaver, *Bayesian Methods for the Physical Sciences*,
Springer Series in Astrostatistics 4, DOI 10.1007/978-3-319-15287-5

B.3 Independence and Conditional Independence

Suppose we are interested in two events A and B and say that event A is that I decide to buy coffee that is decaffeinated and event B is a supernova event occurring in the Andromeda galaxy. One would find it hard to argue that my choice of coffee would have any influence on the occurring of the supernova. In probability theory we call the events A and B independent when the occurrence of event A has no influence on the occurrence of event B, and vice versa (for example, the supernova explosion would not have any influence right now on my choice of coffee). Mathematically, two random variables x and y are independent if their joint distribution, $p(x,y)$ is the product of their marginal distributions, $p(x)$ and $p(y)$. Symbolically,

$$p(x,y) = p(x)p(y).$$

Independence is a very useful property because it allows us a simple way to write the joint distribution of a collection of random variables (and hence their likelihood). If $x_1, \ldots, x_n$ are independent random variables, then their joint distribution is just the product of their marginal distributions:

$$p(x_1,\ldots,x_n) = p(x_1)\cdots p(x_n).$$

Related to independence is conditional independence. Suppose now that two events, each one not influencing the other one, such as the explosion of a supernova in our Galaxy and in Andromeda. Both depend on, and are useful to measure, the rate of supernovae occurrences in galaxies. Mathematically, two random variables (measurements in the example), x and y, are conditionally independent, given a third random variable z (the supernovae rate in the example), if their joint conditional distribution, $p(x,y|z)$ can be written as the product of their marginal conditional distributions, $p(x|z)$ and $p(y|z)$,

$$p(x,y|z) = p(x|z)p(y|z).$$

Conditional independence is what is usually stated by “independent” in many articles in physics and astronomy: measurements are independent in the sense that they do not influence each other, yet they depend on a parameter z (often the one we are interested in studying and that triggered our interest in the data x and y). Mathematically, if $x_1,\ldots,x_n$ (representing our data) are independent random variables given the parameter θ, i.e., apart from the dependence on θ, then observing x_i does not affect the measurements x_j. The likelihood is then the product of the individual likelihood terms:

$$p(x_1,\ldots,x_n|\theta) = p(x_1|\theta)\cdots p(x_n|\theta) = \Pi_i p(x_i|\theta).$$

A note of warning, however, independence does not imply conditional independence and conditional independence does not imply independence.