Contents

1

A Bayesian approach to galaxy evolution studies

1.1 Discovery space

We, astronomers, mostly work in the 'discovery' space, the region where effects are statistically significant at $\lesssim 3$ sigma's or near boundaries in data or parameter space. Working in the 'discovery space' is a normal astronomical activity: only few, among many, published results are initially found at large confidence. Positive defined quantities (such as mass, fractions, star formation rates, dispersions, etc.) are sometimes found to be negative, or, more generally, quantities are sometimes found at unphysical values (completeness larger than 100 %, $V/V_{max} > 1$ or fractions larger than 1, for example). Working in the 'discovery space' is a normal activity of frontier-line research because almost every significant result usually reach (if any) this status after having appeared first in the 'discovery space', and because a good determination of known effects/trends usually triggers searches for finer, harder to detect, effects, mostly falling in, again, the 'discovery space'.

Many of us are very confident that commonly used statistical tools properly work in the situations in which we use them. Unfortunately, in the 'discovery space', and sometimes outside it, we should not take it for granted, as shown below with a few examples. We cannot avoid working in this grey region, because, in order to move our results in to the statistically significant area, we often need a larger or better sample. In order to obtain this, we first need to convince the community (and the Time Allocation Committee) that an effect is probably there, by working in the 'discover space'. Furthermore, awaiting a larger sample has the unappealing property that someone else will publish our result and he, not us, will be credited for the discovery. Of course, we are assuming that a larger sample exists and it is accessible, which is not always the case: there is just one universe (and sky) already fully observed in the microwave. Or, a group formed by 10 galaxies has no more than 10 galaxies to be used to measure its velocity dispersion. Or, gamma ray bursts (and any transient event) cannot become permanent enough to allow us to collected enough photons to put our aimed measurement, say polarization, in the 'statistically' significant area. Working in the 'discovery space' region is therefore an essential part of the astronomer work.

Standard tools may fail (especially if mis-used) in many ways. In the next two sections we will show some examples of failure in two idealized experiments, and we will show that the Bayesian approach does not suffer from these failures. In the first example we show that the maximum of the likelihood (the best fit) may not be a good estimates of the true value: averaging the likelihood is preferable to maximizing it. The second example shows that sometimes the observed value is biased and highlights the bad things that may occur when the prior is ignored. These two examples are very simple with respect to true problems, and they have been chosen so as to make obvious the fact that the best fit value or observed value may be bad or biased. The third example shows that when the sample size is small, even simple operations on data, such as perform an average or fitting it with a function, is a potentially risky operation and there is nothing that guareantees that what is failing in a simple case works correctly in a difficult one. In Sec ?? and 1.5 consider two realistic examples, showing failures of standard methods more difficult to spot, but of the same nature of easely spotted failures: all of them come from contradicting axioms of probabilities. In the first example we want to measure the width of a distribution in presence of a contaminating population. A mixture modelling on inhomogeneous Poisson processes easely solve this problem. In the second example we want to fit a trend in presence of contaminating population and we will use a mixture of regressions. We finally conclude the chapter showing that the Bayesian approach allows to understand what really is the number returned by tests like Kolomogorov-Smirnov, χ^2, etc., named, in the statistical jargon, $p-$ value.

1.2 Average vs. Maximum Likelihood

Maximum likelihood estimates (called 'best fit' by astronomers) are one of the most used tools in astronomy and it is taken from granted that maximizing the likelihood (or minimizing the $\chi^2 = -2\ln\mathcal{L}$) will always give the correct result.

Mixture distributions naturally arise in astronomy when data come from two populations, in such cases as a) a signal is superposed on a background, b) there are interlopers in the sample, c) there are two distinct (by colour, morphology, dynamical properties, etc.) galaxy populations, d) taking an image of the sky using an instrument with a field of view large enough to accomodate more than one source, or e) observing a galaxy spectrum we note the presence of two stellar populations. Finally, there are many other cases as well. Let's consider the simple case of a mixture (sum) of two Gaussians:

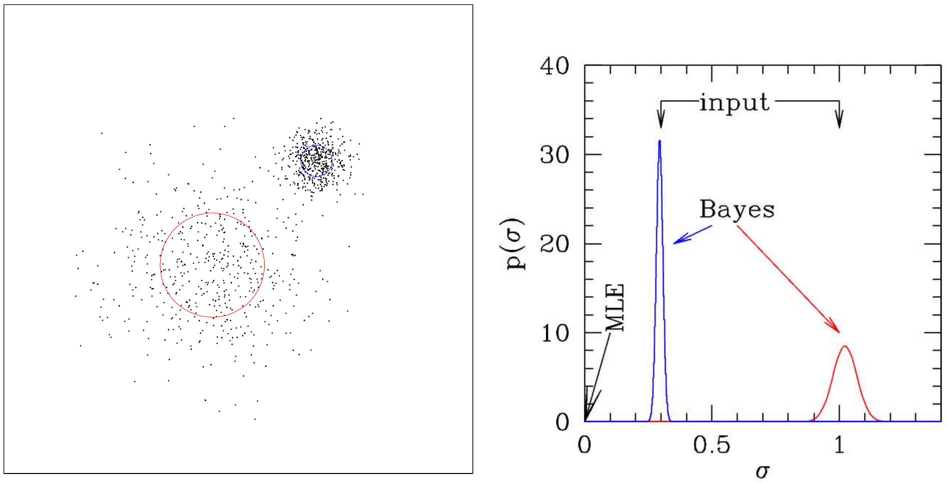

Figure 1.1 Left Panel: An example of mixture distribution in two dimensions. Readers may think that we displayed the spatial distribution of photons coming from two astronomical extended sources as observed by a perfect instrument (no background, perfect angular resolution, etc.), or the distribution of two galaxy populations in a two dimensional parameter space, or whatever else. Right panel: true (input) values, maximum likelihood estimate (MLE) and posterior (Bayes) probability distributions for the example in the left panel.

$$p(y_i|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = \lambda \mathcal{N}(y_i|\mu_1, \sigma_1^2) + (1-\lambda)\mathcal{N}(y_i|\mu_2, \sigma_2^2) \tag{1.1}$$

where (μ_j, σ_j) $j = 1, 2$ are location (or center) and scale (or width) of the two Gaussians $\mathcal{N}(y_i|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}}$, λ and $1-\lambda$ are the proportions of the two components and y_i is the i^{th} datum. Fig. 1.1 shows a two dimensional example.

We want to determine the locations and scales of the two Gaussians with the data at hand. The likelihood of independently and identically distributed data is given by the the product, over the data y_i, of the terms in eq. 1.1:

$$p(y|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = \prod_i p(y_i|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) \tag{1.2}$$

We usually maximize the likelihood in current problems, blindly assuming that the maximum likelihood values are good estimates of the true value. Here though, the parameters that maximize the likelihood above (the 'best' fit) are not near their true value (e.g. those used to draw the points in Figure 1.1), but occur when $\mu_j = y_i$ and $\sigma_j \to 0$. In fact, the likelihood goes $\to \infty$ as $\sigma_j \to 0$, because the i^{th} term of eq 1.1 diverges and the remaining ones take a finite, non zero, value.

The problem is a general characteristic of mixtures of probability distributions of non-fixed variance, not only of Gaussians. The problem does not disappear 'in the long run', i.e. by disposing of a sufficiently large sample (that we don't have, Time Allocation Committes 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 presence of interlopers, or many other quantities in presence of two populations or signals (or an interesting and an un-interesting 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 real applications, as those described in sec 1.4 and 1.5, nothing as bad as an infinity appears, and thus there is no "warning bell" signaling that something is going wrong.

The Bayesian approach is not affected by such problematics: it never instructs us to maximize any unknown parameter, 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.

The right panel of Fig 1.1 shows the posterior distribution for the data shown in the left panel, and adopting a constant prior up to very large values (the precise values are irrelevant for this parameter estimation problem). The posterior is well-behavied and it is centered on the input value. The likelihood, instead, has hundreds of infinities, one per datum, all at $\sigma = 0$.

1.3 Priors and Malmquist/Eddington bias

Number counts are steep. It is well known to astronomers that the true value, μ, of the source intensity differs from the measured counts, n, of the source when n is small: even in presence of symmetric errors an object with n counts is more probably to come from the numerous population of objects having $\mu < n$ than the rare population having $\mu > n$. Therefore, objects with n counts have, likely, $\mu < n$ (e.g. Eddington 1913). A similar effect arises for parallaxes, star counts, velocity dispersions and any noisy determination of a quantity concerning an object drawn from a population that shows an important numerical change over

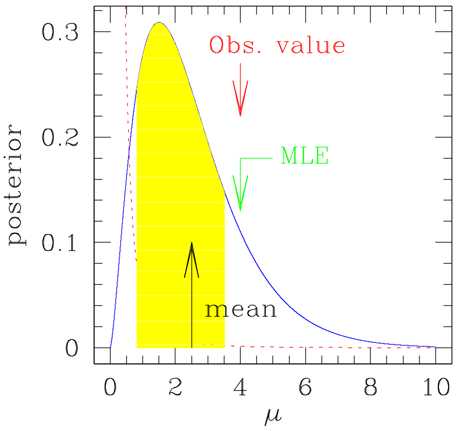

Figure 1.2 Posterior distribution (solid curve) of the source flux, having observed four photons and knowning that source number counts have an euclidean distribution (red dashed distribution). The maximum likelihood estimate (MLE) is also reported.

the range included by the error on the measurand† (on μ, in our example), i.e. most of the times in which a (Malmquist or Eddington) bias is invoked. Restated in the statistical jargon, in parameter estimation problems where the prior, $p(\mu)$, has a large change in the μ range where the likelihood is slowly varying, the prior cannot be neglected.

As a quantitative example, tailored around the X-Boote survey (Kenter et al. 2005), lets consider a Poisson process (i.e. anything Poisson-distributed), $p(n|\mu) = \frac{1}{n!}\mu^n e^{-\mu}$, with rate μ and a power law prior of logarithimic slope α, $p(\mu) = \mu^{-\alpha}$ (the latter is called number counts by astronomers). Having observed four photons (i.e. $n = 4$), the maximum likelihood estimate of the source rate is $\hat{\mu} = 4$, but we, astronomers, known by experience that this value is wrong: with better data we find, most of the times, $\mu < 4$. Assuming an euclidean slope $\alpha = 2.5$ (the observed value of the slope of number counts at the rate of interest), the posterior $p(n|\mu)p(\mu)$ is $\propto \frac{1}{n!}\mu^n e^{-\mu}\mu^{-2.5}$. The posterior mean (or, in astronomical terms, the Eddington corrected value) of the source rate is 2.5 photons, 40 % less than the originally observed value ($n = 4$).

The same holds true, as mentioned, for many noisy quantities, such as the determination of a velocity dispersion with just a few velocities or the estimate of N_{200} (cluster richness, see Andreon 2009 for details on the latter).

What is known in astronomy as Malmquist or Eddington bias is the manifes-

† Measurand it the parameter being quantified. It usually differs from the outcome of the measurement because of the noise.

tation of the important role of the prior when measurements are imprecise, i.e. the fact that a correct inference proceeds along the Bayes theorem. The prior (the correction for Malmquist bias) moves the result away from the observed value, or the maximum likelihood estimate, and brings it near the true value. The example shows that priors (the fact that there are much more fainter system than bright ones) cannot be ignored in inferences, if one does not wish to be wrong most of the time. Priors are specific to the Bayesian approach and, usually, non-Bayesian methods consider them something to be avoided.

It is therefore apparent that prior-free and maximum likelihood methods are in trouble.

1.4 Small samples

Astronomers are often faced to 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 mass increases. In most of the 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, or the chosen parametrization does not matter? Both parametrization have been adopted in recent astronomical papers (and the author, in Andreon 2008 took a third, different, parametrization!). Specific star formation rates (sSFR) and e-folding times, τ, are approximatively reciprocal measures (long e-folding times correspond to small sSFR). To the author knowledge, any of the proposed parametrizations (e.g. L/F vs F/L or sSFR vs τ) has a special status: there is any serious physical though behind the choice of one of them. Unfortunately (for the astronomers), when the sample size is small, results obtained using commonly used formulae (e.g. weighted average, best fit, etc.) does depend on the adopted parameteric form. For example, an average value, computed by a weighted sum, or a fit performed minimizing the χ^2, has a special meaning, because the result depends on which parametrization is being adopted.

As an example, let us consider two, for the sake of clarity, data points, $(f/l)_1 = 3 \pm 0.9$ and $(f/l)_2 = 0.3333 \pm 0.1$. The error weighted average $\langle f/l \rangle$ is 0.37. The reciprocal values ($(l/f)_i = 1/(f/l)_i; 0.3333 \pm 0.1$ and 3 ± 0.9) have error weighted average 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 (obvious, the error on the mean is smaller than the error on any data point). At first sight, by choosing the variable parametrization, the astronomer may select the number he want, 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.

One may argue that in the shown case 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, or looks for trends or fits a function using a number of data points only sliglhy exceeeding the number of parameters. Often, few points are the result of a large observational effort (and obtained through analysing 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 few numbers is almost all we can do. For example, how many estimations of SFR at $z \sim 6$ exist? Should we not combine the very few available in some way to take profit of all them? Small sample problems are often 'hidden' in large samples: even large surveys, such as the SDSS, 2dF, VVDS and CNOC2 surveys including tens or hundreds of thousand of galaxies, estimate galaxy densities using subsamples equivalent to just one to ten galaxies. Finally, how many of us have checked, before performing a fit or an average, if the sample size is large enough to be insensitive to the chosen parametrization?

The described problem originates from the freedom, in the frequentist paradigm, of choosing an estimator of the measurand ($\langle f/l \rangle$ or $1/\langle l/f \rangle$ for example). All estimators (satisfying certain conditions) will converge on the true value of the estimand 'in the long run', but without any assurance, however, that such regime is reached with the sample size in hand. Untill 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.

This example shows that frequentist methods return the value of an estimator of the measurand, neither the measurand itself, nor its probability distribution. While these differences are easy to appreciate in our example, it is not so with real cases, dealing with dispersion, slope or intrinsic scatter, discussed in next sections.

1.5 Measuring a width in presence of a contaminating population

Let's focus now on how to measure the scale (dispersion) of a distribution (say, of velocities v), knowing that the sample is contaminated by the presence of interlopers, but without the knowledge of which object is an interloper. The main idea is not to identify or de-weight interlopers in the scale estimate, but to account for them statistically, precisely as astronomers do with photons when estimating the flux of a source in presence of a background.

We assume that data come from two populations: background galaxies, whose distribution is assumed to be an homogeneous (i.e. the intensity is independent on v) Poisson random process, and cluster galaxies, whose distribution is assumed to be a Poisson process whose intensity is Gaussian-distributed in v, i.e

$$I(v_i|...) = N_{clus}\mathcal{N}(v_i|v_{clus}, \sigma^2_{v_i} + \sigma^2_{clus}) + \frac{N_{bkg}}{\Delta v} \tag{1.3}$$

where Δv is the (velocity) range over which velocities are considered (say, ±5000 km/s from the cluster preliminary velocity center), σ_{v_i} is the velocity error, N_{clus} and N_{bkg} are the number of cluster and background galaxies, and v_{clus} and σ_{clus} are our (perhaps) most interesting quantities: the cluster redshift and velocity dispersion.

Simple algebra shows that the likelihood of independently and identically distributed data, $p(v|I(v))$, is

$$p(v|I(v)) \propto \prod_i I(v_i|...) \ \ e^{-\int_V I(v|...)} \tag{1.4}$$

Combined with prior probability distributions for the parameters, this likelihood function yields, via the Bayes theorem, the posterior distribution for the function parameters θ, given the data. Uniform priors, zero-ed at unphysical values of the parameters are often adequate for the samples usually available. Marchov Chain Monte Carlo with a Metropolis sampler (Metropolis et al. 1953) may be used to sample the posterior. The chain provides a sampling of the posterior that directly gives credible intervals for whatever quantity, either for the parameters θ or any derived quantity: for an interval at the desired credible level it is simply matter of taking the interval that includes the relevant percentage of the samplings.

Most literature estimates of (cluster velocity) dispersions are, instead, based on the family of estimators presented by Beers, Flynn & Gebhardt (1991), often called 'robust'. We now compare the performances of the 'robust' method and the Bayesian approach.

Let us consider a simulated 'cluster' having $\sigma_v = 1000$ km s^{-1} composed of 500 galaxies with Gaussian distributed velocities, superposed over a background of 500 uniformly distributed (in velocity) interlopers, within ±3000 km s^{-1}. Note that within 1000 km s^{-1} from the cluster center there are on average $500 \cdot 0.68 = 340$ members and $500/3 = 83$ background galaxies, i.e. the contamation is here just 20 %. The large sample size has been adopted to leave data to speak by themselves. Applying the methods of Beers et al. (1991) yields $\hat{\sigma_v} = 1400$ km s^{-1} which is an excessively large estimate of σ_v (and hence of mass). The posterior mean is 940±85 km s^{-1} which is closer to the 'true' (input) value. 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 on the sample size, provided the relative fraction of cluster and interlopers is mantained.

We now acknowledge that, in true life experiments, we do not precisely know the model from which data are drawn (i.e. is the velocity distribution perfectly

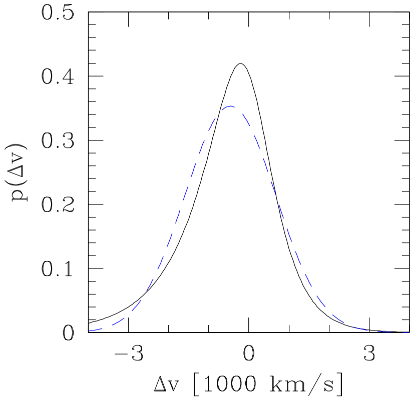

Figure 1.3 Perturbed velocity distribution (solid line), given by eq. 1.5, and a Gaussian with identical first two moments (dashed blue line). The former is used to generate hypothetic data, the latter is assumed to estimate σ_v.

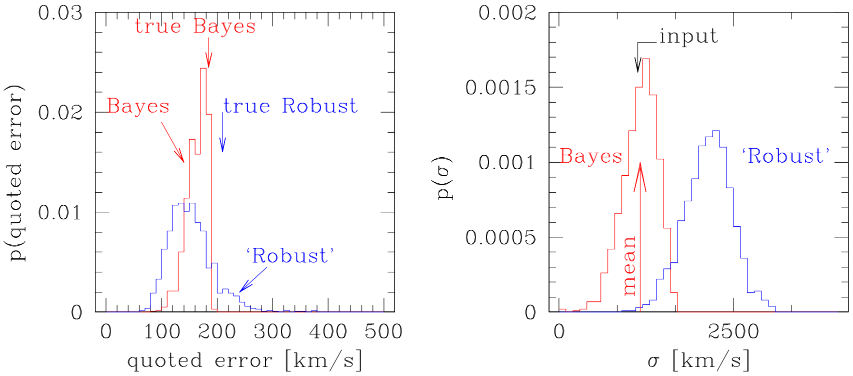

Figure 1.4 Left panel: Comparison between the distributions of the quoted error (histograms) by 'Robust' (biweight estimator of scale, in blue) and by our Bayesian method (in red) for 1000 simulations of 25 galaxies uncontaminated by background. 'Robust' error estimate is noiser (the histogram is wider) 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. Right panel: True (input) value of the velocity dispersion, and histogram of recovered values by the biweight estimator of scale (right histogram) and of by our Bayesian method (left histogram) for 1000 simulations of a sample of 25 galaxies, 50 % contaminated by background. In presence of a background, the 'robust' estimate of the velocity dispersion is biased.

Gaussian? Does it have more power in the wings or is it slightly tilted?). Lets therefore suppose that cluster substructure perturbs the velocity distribution, that we now assume to be described by

$$p(v) \propto e^{v/1000}(1 + e^{2.75v/1000})^{-1} \tag{1.5}$$

depicted in Fig 1.3 (solid line). The function has first and second moments (mean and dispersion) equal to -460 and 1130 km s^{-1}, respectively, excess kurtosis and a non zero skewness. We simulate 1000 (virtual) clusters of 25 members each (and no interlopers) drawn from the distribution above (eq. 1.5), but we compute the velocity dispersion using eq. 1.3, i.e. with a likelihood function appropriate for members drawn from a Gaussian, to make our study more realistic. The average of determined posterior means is 1140 km s^{-1} (vs. the 1130 km s^{-1} input value) with a standard deviation of 185 km s^{-1}. The uncertainty (posterior standard deviation), averaged over simulations, is 163 km s^{-1}, close (as it should be) to the scatter of the posterior means. The uncertainty has a negligible scatter, 18 km s^{-1}, indicating the low noise level of each individual uncertainty determination. The uncertainty of the dispersion error is four time better with a Bayesian estimation than using BCES, displaying a scatter of 70 km s^{-1}, and returning uncertainties as small as 73 km s^{-1} and as large as 865 km s^{-1} for data that are supposed to give a unique, fixed, value of uncertainty (see left panel of 1.4).

As a more difficult situation, we now consider a sample drawn, as before, from a distribution different from the one used for the analysis, but furthermore $\sim 50\%$ contaminated by interlopers and consisting of half as many members: 13 galaxies are drawn from the distribution above (eq. 1.5), superposed to a background of 12 galaxies, uniformly drawn from ±5000 km s^{-1}. Within 1130 km s^{-1} (i.e. $1\sigma_v$) from the center the average contamination is about 20 %. The average of found posterior means is 1160 km s^{-1} (vs. the 1130 km s^{-1} input value). The average uncertainty is 390 km s^{-1}, with a low (80 km s^{-1}) scatter. The biweight estimator returns, on average, a strongly biased estimate, 2135 km s^{-1}, see Fig 1.4.

As mentioned, mixtures often arise in astronomy and our equations 1.3 and 1.4 equally hold for any Poisson signal superposed on a background, such as the distribution of galaxies in colour or the spatial distribution of X-ray photons or galaxies, or whatever. We just need to rename variables with names appropriate to the measurand, and eventually, consider a more complex model, for example if the background distribution is not uniform. In fact, the solution illustrated in this section has been developed to measure the X-ray core radius of a cluster of galaxies barely detected (Andreon et al. 2008) and later used to measure the cluster velocity dispersion.

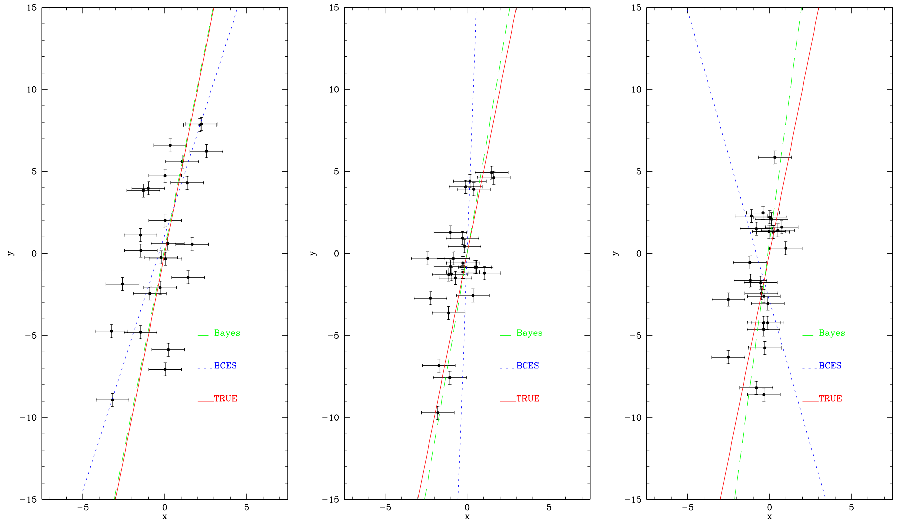

Figure 1.5 Three simulated data sets of 25 objects (points), true trends from which data are generated (red solid line), recovered trend by BCES (dotted blue line) and mean (Bayesian) model (dashed green line). In our 1000 simulations, BCES results worser than those shown in central and right panels occur in about 10 % of the cases (see text for details).

1.6 Fitting a trend in presence of outliers

Let's consider an apparently different problem: we observed some quantities x and y and we want to estimate some parameters describing how these two quantities vary as a function of each other. In astronomy, these regressions are named, say, Tully-Fisher, Faber-Jackson, Colour-Magnitude relations, Fundamental Plane, cluster scaling relation, Ghirlanda relation (for Gamma Ray Burst), etc. Many articles present their own way for the determination of these parameters (direct-, inverse-, orthogonal-, Bivariate Correlated Error and intrinsic Scatter-, Measurement errors and Intrinsic Scatter- fit). The Bayesian approach allows a simple solution, even in the difficult case of a linear fit in presence of heteroscedastic (i.e. of different magnitude) errors on both variables and an intrinsic scatter (i.e. not accounted for by experimental errors), and censored or truncated data. In such case, and for an ignorable data collection process (see below) and for variable having names appropriate for the colour–magnitude relation, slope a, intercept c, intrinsic scatter, σ_{intr} of the colour–magnitude relation, and Gaussian photometric errors, the likelihood is a Gaussian (e.g. D'Agostini 2003; 2005; Gelman et al. 2004):

$$p(m_i, col_i|a, c, \sigma_{intr}, nobkg) \propto \mathcal{N}(col_i|a\, m_i + c, \sigma^2_{intr} + \sigma^2_{col_i} + a^2\sigma^2_{m_i}) \tag{1.6}$$

where σ_{m_i} and σ_{col_i} are the errors on magnitude and colour of the i^{th} galaxy. The solution is quite intuitive: the colour magnitude relation has a width given by the sum in quadrature of the intrinsic scatter, colour errors and magnitude errors propagated on the colour (via the slope a). In spite of the solution simplicity, many pages are spent in journals to decide which approximate procedure (usually far more complicate that the equation above) should be used in which cases, all of which can be shown to be approximations of eq. 1.6. Of course, a change of variable names makes the result useful for whatever scaling relation.

To avoid recourse to maths, let us perform numerical simulations and compare the Bayesian approach and the state-of-the-art non-bayesian astronomical method, BCES (Akritas & Bershady 1996). BCES accounts for intrinsic scatter and for heteroscedastic errors. We considered a sample of 25 objects obeying to a linear trend of slope $a = 5$ with an intrinsic dispersion $\sigma_{intr} = 1$. Data have Gaussian errors, $\sigma_x = 1$ and $\sigma_y = 0.4$. In detail, the true x values have been drawn from a Gaussian having $\sigma_\tau = 1$ centered on $\mu_\tau = 0$. The true y values are given by $y = 5x$ (i.e. $c = 0$ in eq. 1.6). Observed x values and y values are computed by adding to each true x and y some noise (a Gaussian variate with $\sigma_x = 1$ and $\sigma_y = 0.4$ respectively). Because of the intrinsic scatter, y is perturbed by adding a Gaussian variate with $\sigma_{intr} = 1$. Figure 1.5 shows three simulated data sets. Qualitatively, these plots look similar to, or better than, many $L_X - \sigma_v$ seen on astronomical journals. We produced 1000 simulations of 25 data points. For each simulation we compute the slope and slope error as determined by BCES. We also compute the slope posterior mean and standard deviation, assuming uniform priors for all parameters but for the slope a, for which we take an uniform prior on the angle α ($a = \tan\alpha$). Since in our problem σ_x is comparable to the x range and the x distribution is far from being uniform (in the statistical jargon "the data collection process is not ignorable"), the likelihood continues to be described by a Gaussian, as eq. 1.6, but in a 2 dimensional (y, x) space. Performing the algebra associated to the matrix product gives a Gaussian $\mathcal{N}(\mu, \sigma^2)$ with parameters μ and σ^2:

$$\sigma^2 = (a^2\sigma^2_\tau + \sigma^2_{intr} + \sigma^2_y)(\sigma^2_\tau + \sigma^2_x) - a^2\sigma^4_\tau \tag{1.7}$$

$$\begin{aligned}\mu = {} & (\sigma^2_\tau + \sigma^2_x)(y_i - c - a\mu_\tau)^2 - 2a\sigma^2_\tau(y_i - c - a\mu_\tau)(x_i - \mu_\tau) + \\ & + (a^2\sigma^2_\tau + \sigma^2_{intr} + \sigma^2_y)(x_i - \mu_\tau)^2\end{aligned} \tag{1.8}$$

As $\sigma_\tau \to \infty$, the likelihood converges to eq. 1.6 (i.e. eq 1.6 is an approximation

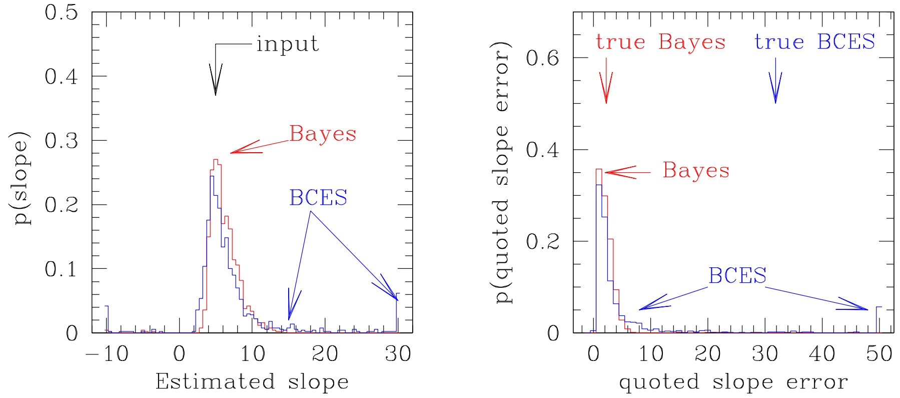

Figure 1.6 Comparison between BCES (blue) and the Bayesian approach (red) for our linear regression problem. We considered 1000 simulations of a sample of 25 objects uncontaminated by interlopers. Left panel: BCES sometimes returns badly wrong slope estimates: the BCES histogram distribution shows outliers, cumulated at −10 and 30. Right panel: True value of the slope error (vertical arrows), as measured by the scatter of returned slope minus input slope, and distribution of quoted errors (histograms). BCES is overly-optimistic about the quality of its error, the very large majority of the error estimates are small when the scatter between input and output slope is large. Furthermore, BCES displays a large scatter in the returned error, for data sample supposed to give identical values of uncertainty.

of the present equation). Of course, the parameters used to produce the data (σ_τ, μ_τ, σ_{intr}, a and c) are assumed to be unknown in both analysis.

The left panel of Fig 1.6 shows that both BCES method and the Bayesian approach return slopes whose distribution is centered on the input value, at least for our setting. However, BCES returns sometimes slopes much different from the input one (look the histogram wings and in particular around -10, 30, where we have cumulated more extreme values). The Bayesian approach does not show such catastrophic failures. The right panel of Fig 1.6 shows the distribution of the quoted errors. The important thing here is not how large (or small) a method claim to be its error, i.e. the location of the plotted histogram, but the veracity of claimed error, i.e. whether the quoted error distribution (i.e. histogram) is located near or far from the true error (vertical arrow). The true error is computed as the scatter between the returned slope and the input slope. On average, BCES optimistically estimates errors by a large factor, mainly because in 10 % of the cases it presents a catastrophical failure. The Bayesian method performs better in this respect: the quoted slope uncertainty is equal to the scatter between the input and output slopes, as it should be. Second, BCES displays a large scatter

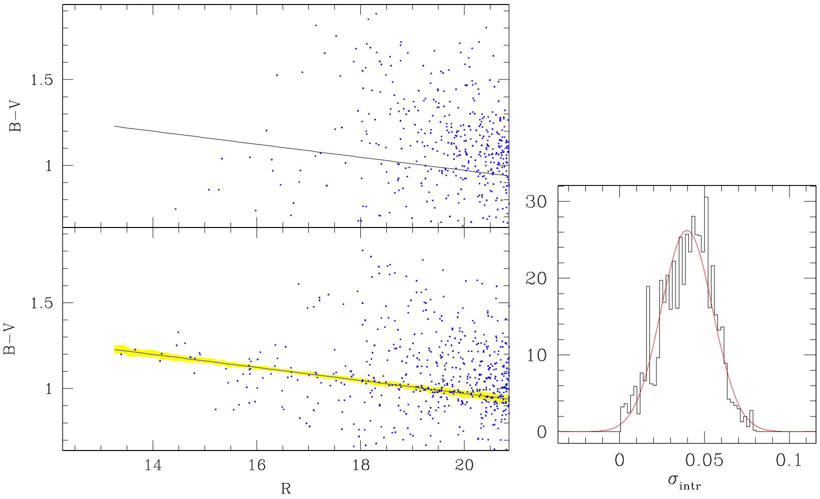

Figure 1.7 Left panels: Colour–magnitude diagrams for background galaxies (upper panel) and cluster+background galaxies (lower panel). These are true data for the cluster Abell 1185, presented in Andreon et al. (2006a). The solid line is the mean colour-magnitude relation of cluster galaxies computed as described in the text. The shaded region marks the 68 % highest posterior interval. Right panel: Posterior probability distribution of the colour-magnitude intrinsic scatter. The jagged nature of the distribution is due to the finite lengh of the used MCMC chain. A Gaussian with first two moments matching the distribution is overplotted to guide the eye.

of the quoted slope error, for data sample supposed to give similar values of uncertainty.

To summarize, although BCES is not systematically in error, in ten percent of our simulations, BCES returns badly wrong slopes with badly underestimated errors. In a real application true values are unknown, and in such a case there is no way to known whether the BCES result is one of the frequently good values or a bad one. The Bayesian method better performs because it is better behavied: there are no such catastrophic failures.

As formulated above, the problem does not account for our everyday experience: real samples are contaminated by interlopers, i.e. objects unrelated to the ones we are interested in. Now our model will be a mixture of two regressions, one carrying the signal (the cluster colour–magnitude relation) and the other describing the background (galaxies, objects on the line of sight), with the usual difficulty that we do not known which galaxy belongs to the cluster and which one is simply projected along the cluster line of sight. A real case (the cluster Abell 1185, from Andreon et al. 2006a) is shown in Fig. 1.7. The distribution of background galaxies in the m,col space is not uniform, and therefore the background is modelled by an inhomogeneous process, $B(m_i, col_i|m, col)$. Therefore, the likelihood of the i^{th} galaxy, $p(m_i, col_i|a, c, \sigma_{intr})$, is given by the mixture of two distributions:

$$\begin{aligned} p(m_i, col_i|a, c, \sigma_{intr}, \alpha, M^*\phi^*) = \Omega_j B(m_i, col_i|m, col)+ \\ + \delta_c \Omega_j \mathcal{N}(col_i|a\, m_i + c, \sigma^2_{intr} + \sigma^2_{col_i} + a^2\sigma^2_{m_i}) S(m_i|\alpha, M^*\phi^*) \end{aligned} \tag{1.9}$$

In this equation we considered the usual case, i.e. we have at least a control field (i.e. data from a sky region uncontaminated by the cluster contribution). In such case, $\delta_c = 1$ for cluster datasets , $\delta_c = 0$ for the other datasets, Ω_j is the studied solid angle. Otherwise, it is just matter of replacing δ_c with a radial profile. S is the usual Schechter (1976) function, with α, M^* and ϕ^* parameters, that describes the luminosity function of galaxies. We have also assumed that the data collection model is ignorable, for mathematical convenience.

As in eq. 1.4, the likelihood of independently and identically distributed data is given by the product, over the data m_i, col_i of the individual likelihood terms. As shown there, the likelihood includes an integral term, given by the integral of the model over the values ranges. The integral should be performed on the appropriate colour and magnitude ranges (those accessible to the data) and it is equal to the expected number of galaxies. This term disfavours models that predict a number of galaxies very different from the observed one. If errors on m (mag) are not negligible, S in eq 1.9 should be replaced by the convolution between the Schechter function and the error function. The inference proceeds as usual, by choosing a prior, computing the posterior, and summarizing the result of the computation above with a few numbers, those of scientific interest.

The problem of determining an instrinsic scatter around a linear trend in presence of outliers or a background population is so difficult that, to our best knowledge, there are no non-Bayesian solutions to be compared with our Bayesian approach. We cannot, therefore, simulate some data and compare the performances of different methods, because of the lack of a contender.

Had we (mis-)used BCES, then the found slope of the colour magnitude-relation shown in fig 1.6 would be completely wrong, and equal to the one of the background population, outnumbering by a factor four the cluster population. This occurs because BCES is not built to be robust against a contaminating population. An ad hoc solution often used in astronomical papers is to remove the slope dependency by subtracting off an expected one (e.g. one observed at $z = 0$, assuming no slope evolution), measuring the scatter using ‘Robust’ methods, as described in previous section, and finally quadratically subtracting colour errors from the measured scatter, following Stanford et al. (1996). This procedure often lead to intrinsic scatter with 68 % error bars extending to negative values, including some examples in Stanford et al. (1996). We have already discussed the shortcoming of using the 'robust' estimate of the scatter. The Bayesian method does not require ad hoc methods, does not make assumptions on the trend slope, and always returns positive intrinsic dispersions, as in the case of Abell 1185 shown in Figure 1.7.

Readers interested in fractions, f_b, bounded in the $[0,1]$ range or hardness ratios, $H-S/(H+S)$, bounded in the physical range for data contaminated by a background may consult Andreon et al. (2006b) and D'Agostini (2004) and remember that the hardness ratio has the same mathematical properties of (is equal to) $1-2f_b$, i.e. one minus twice the (blue) fraction discussed in these two works.

1.7 What is the number returned by tests like χ^2, KS, etc.?

Many articles measure the 'probability of rejecting the null hypothesis' using some statistical tests, for example Kolmogorov-Smirnov's, Kendall's, Spearman's rank correlation, $F-$, Student's $t-$, Wilcoxon rank, χ^2 tests. Many of us have noted oddities with the numbers (called $p-$values) returned by them: by taking two statistical tests we sometimes found widely different 'probabilities', e.g. 0.001 and 0.861. How can this be the case, since the desired result is a single unique value? In our example, which one is the good probability, the one rejecting the null or the other one? The mere existence of a variety of tests, as opposed to a single one, is an indication that no test always gives the desiderated number. Actually, p-values are not the probability of the hypothesis, that is the desired probability. They are the probability of observing more discrepant values of the chosen statistic for hypothetical data drawn from null hypothesis, that is the probability of rejecting the null hypothesis when it is true. There is nothing strange that two different statistics (measures, say height and width) of data drawn under the null hypothesis takes different values.

The difference between the $p-$values and probability of the hypothesis can be better understood with an astronomical example: the detection of faint sources. In such case, the null hypothesis to reject is "no source is there". Let I_0 be the flux measured at the target position. A usual way to compute the detection confidence is by measuring how frequently one observes larger values, $> I_0$, under the null, i.e. in areas free from sources: $p(> I_0|background)$. The $p-$value is, precisely, the measured frequence. For many famous tests, like those mentioned at the start of the section, the probability distribution of the test statistic is analytically known and there is no need of further data (the background) in order to compute the distribution of the test statistics.

Let us suppose to have found a $p-$value of 0.003, i.e. that measurements free of sources gives $p(> I_0|background) = 0.003$ (= 0.3 %, readers may of course choose any other value). Should this means that the target is real at one minus the $p-$value confidence, $p(source|I_0) = 1 - 0.003 = 0.997$, i.e. is real at 99.7 % confidence ? Certaintly not. Qualitatively, if sources fill a small portion of the sky there is a lot of sky left to the background. Then, statistical fluctuations of the background, even rare ones, may overwhelm the number of true sources. In such a case, only a very tiny fraction of detections are true, not 99.7 % as the p-value leads us to believe. More quantitatively: let x be the portion of sky occupied by sources (when observed in the same observational set up that gives a $p-$value of 0.003), and N the number of independent beams in the sky. Note that x is the probability a priori that there is a source in a beam. Then, xN beams are occuped by sources and $(1-x)N$ are not. Assuming a 100 % detection efficiency, xN are true sources detected, and $0.003(1-x)N$ will be instead false positive detections. Thus, there will be $xN + 0.003(1-x)N$ detections, but among them only xN are true sources. The probability to be a true source, $p(source|I_0)$, is given by the fraction of true detected sources over total number of detections: $x/(x + 0.003(1-x))$. If $x = 0.07$ %, a value appropriate for typical Chandra exposures, then sources believed to be detected at 99.7 % confidence (or better, with a $p-$value of 0.003) have 19 % probability of being real. Adopting instead a 5 % $p-$value, we end up with a catalog composed by entries that are junk 99 times out 100, instead of being true sources 95 % of the times, as the $p-$value leave to think. Only in fortunate cases (appropriate values of x) one may have similar numbers for the $p-$value and the probability of rejecting the null hypothesis. Therefore, these two probabilities are conceptually different and take different values.

As shown in the example, the desired probability does depend on the a priori probability of the (null) hypothesis (x in the example). However, virtually all non Bayesian astronomical papers compute $p-$values but call them “the probability of rejecting the null hypothesis”. For example, in testing the reality of a trend, the Sperman’ rank correlation test is often used, and the one minus the $p-$value is quoted as probability to reject the null (“no trend”) hypothesis. Such a practice ignores the essential role played by the a priori probability of the competing hypothesis, which, in principle, may convert a “95 %” confident result into an inconclusive result, as in our example. The Bayesian approach is based on probabilities for the hypothesis, it cannot ignore them, and in our example the Bayes theorem takes the expression we have used to evaluate the desired probability.

1.8 Summary

The Bayesian approach solves some difficulties encountered with other procedures. It works in the regime of typical researcher activity: when looked-for effects are marginally significant, or near boundaries, such as when the small intrinsic dispersion of the colour-magnitude relation is to be determined, or when there is no agreement among astronomers how to set up the right procedure, as for the regression problem in absence of a background. It also offers a solution when none is there, as in the case of the fit of a trend in presence of a contaminating population. It works when otherwise obtained results are unsatisfactory, as for velocity dispersions or for cases when other procedures return unphysical values. The Bayesian approach already includes corrections for biases, as for the Eddington bias. The ultimate reason for its good performances is highlighted in two idealized cases at the start of this chapter, a) it obeys to the sum axiom of probability and thus make averages (marginalize) over unknown quantities, instead of maximize the value of some ad hoc estimators, and b) it performs inferences following the Bayes' theorem instead of considering priors as something to be avoided. Finally, the Bayesian approach clarifies what other methods are actually computing, for example what is the meaning of the number returned by Kolmogorov-Smirnov, χ^2, Wilcoxon rank tests.

Let's conclude this chapter by remembering that the scientific method suggests to always prefer a procedure known to work over one whose reliability is uncertain.

Journal or Book (with editors in brackets), volume, pages

Andreon, S. (2008), The history of mass assembly of faint red galaxies in 28 galaxy clusters since $z = 1.3$, MNRAS 386, 1045

Andreon, S. (2009), Calibration of N_{200} (richness) masses of galaxy clusters, MNRAS, submitted.

Akritas, M. G., & Bershady, M. A. 1996, Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter, ApJ, 470, 706

Andreon, S.; Cuillandre, J.-C.; Puddu, E.; Mellier, Y. (2006a), New evidence for a linear colour-magnitude relation and a single Schechter function for red galaxies in a nearby cluster of galaxies down to $M^* + 8$, MNRAS 372, 60

Andreon, S.; De Propris, R.; Puddu, E.; Giordano, L.; Quintana, H. (2008), Scaling relations of the colour-detected cluster RzCS 052 at $z = 1.016$ and some other high-redshift clusters, MNRAS, 383, 102

Andreon, S., Quintana, H., Tajer, M., Galaz, G., & Surdej, J. (2006b), The Butcher-Oemler effect at z 0.35: a change in perspective, MNRAS, 365, 915

Beers, T.; Flynn, K.; Gebhardt, K. (1990) Measures of location and scale for velocities in clusters of galaxies - A robust approach. The Astronomical Journal, 100, 32

D'Agostini, G. (2003) Bayesian reasoning in data analysis. A critical introduction (World Scientific Publishing)

D'Agostini, G. (2004) Inferring the success parameter p of a binomial model from small samples affected by background (arXiv:physics/0412069)

D'Agostini, G. (2005) Fits, and especially linear fits, with errors on both axes, extra variance of the data points and other complications, (arXiv:physics/0511182)

Eddington, A. S. (1913) On a formula for correcting statistics for the effects of a known error of observation, MNRAS 73, 359

Gelman A., Carlin J., Stern H., Rubin D. (2004) Bayesian Data Analysis, (Chapman & Hall/CRC)

Kenter et al. (2005) XBootes: An X-Ray Survey of the NDWFS Bootes Field. II. The X-Ray Source Catalog The Astrophysical Journal Supplement, 161, 9

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller A., Teller, E., 1953, J. Chem. Phys, 21, 1087

Schechter, P. (1976), An analytic expression for the luminosity function for galaxies, ApJ, 203, 297

Stanford, S. A., Eisenhardt, P. R., & Dickinson, M. 1998, The Evolution of Early-Type Galaxies in Distant Clusters, ApJ, 492, 461