Chapter 2

A Cloudy Night QSO

2.1 Introduction

One of the greatest problems in studying the Lyman α forest region of QSO spectra is dealing with the uncertainties inherent in the data analysis procedures. Often high-resolution ($\lesssim 0.5$ Å FWHM) QSO spectra are of low to moderate signal-to-noise (S/N) ratios ($\sim$ 5–20) because of observational constraints. This can cause significant problems in many of the steps of the data reduction and analysis processes.

Two crucial steps in the process are the fitting of a continuum level and the fitting of Voigt profiles to the absorption features. Many computer programs exist to help perform these tasks, and they fall into two basic categories:

Interactive programs: These aid the user by allowing her to set various fit parameters and examine the results visually. If required, the parameters are modified until an acceptable fit is achieved. This approach relies on user experience and commonsense discretion in interpreting the often complex features of a spectrum. The first programs for fitting Voigt profiles were of this type (Davenhall and Pettini, 1984; Davenhall and Pettini, 1989).

Automated programs: These take most of the decisions away from the user and rely on statistical formalism in an attempt to produce fits which are objective and, in some sense, "best fits".

The central disagreement between advocates of these two approaches to parameter fitting is on the value of human input to the process of data analysis.

Interactive fitting allows a human to "keep an eye on" the data analysis at all times. Some unusual patterns or peculiarities in the data can easily be recognised by an experienced person and dealt with appropriately, whereas a computer program might continue blindly and produce anomalous results. A proponent of interactive analysis might consider human input to be a valuable, if not essential, addition to analysing complex data which could otherwise be misinterpreted.

On the other hand, the aim of a completely objective approach has a strong philosophical appeal, in that the results become clearly reproducible. The underlying mathematics can be optimised so that, formally, the results produced have the

highest probability of being correct, given the information available. Advocates of automated fitting procedures might consider human input to be "pollution", resulting in irreproducible and biased results.

Clearly there are advantages in both approaches. Interactive fitting allows the injection of thought, a process so far beyond the capabilities of computer programs. Automated fitting provides a rigorous mathematical basis, which no human can emulate merely by seeing a graphic on a computer terminal. Ideally, *both* these abilities would be utilised fully, but unfortunately no software has yet been designed to allow this.

Some of the specific problems affecting the fitting of continua and absorption profiles are presented in Sections 2.1.1 and 2.1.2.

2.1.1 Continuum Fitting

To understand the problems associated with fitting a continuum to a real spectrum, it is helpful to consider a series of idealisations, beginning with an absorption-free spectrum with a continuum of constant intensity at all wavelengths. If Poissonian noise is added, then the mean and median intensities over all wavelength bins are equivalent estimators of the continuum level. Additionally, knowledge of the photon counting statistics allows uncertainty estimates of these estimators to be made.

If the continuum level varies with wavelength (and the spectrum is still free of absorption), a function which gives continuum estimators versus wavelength must be found. The most straightforward approach here is to divide the spectrum into wavelength intervals, calculate a continuum estimator on each interval, and fit a function to the resulting points. This broad class of procedures includes all algorithms for smoothing or function-fitting noisy data.

At this stage several decisions must be made. The size of the intervals is important. If they are too large, the variation of the continuum level over the interval will bias the estimators. If they are too small, the number of wavelength bins included will not be enough to provide a statistically reliable estimator. The estimator itself can be of several different types. The mean, median, and several other defensible choices are available, most producing different values. The choice of how to fit a function to the chosen points is also crucial. Many choices are available, each with advantages and disadvantages.

The complexity of the problem rises again when absorption features are present. Many, or most, of the wavelength intervals which can be chosen will be affected by absorption, which lowers the apparent continuum level and biases the estimators. Strong features are easily recognised by eye and can be excluded from an interactive fitting process. A computer program designed to fit a continuum relies on statistical inference to recognise and eliminate such regions.

Weak lines are much more problematic. A single weak line may not be recognised as such, either by eye or algorithm, and instead be put down to statistical fluctuations. Since a program relies on the statistical properties of the data, this inevitably biases the result it produces. The effect on an interactive process is less

clear. It may be that a human, while not consciously detecting the absorption, can be trained to account for it by a comparison with other parts of the spectrum. Multiple weak lines appearing in close proximity are even worse. The partial blending can result in a region of spectrum where no light is unabsorbed and which convincingly mimics a smooth continuum, but at a lower level than the true continuum. Neither human nor computer can identify such regions without appealing to data beyond the extremes of the area in question and applying assumptions about the overall shape of the continuum.

The combined effect of strong and weak lines is that the points chosen as continuum estimators will be sparse and unevenly spaced, because regions with strong lines provide no estimates, and biased, because undetected lines will lower the apparent continuum. This will be true no matter how the points were determined. Whether a computer program or a human can do a better job is unclear.

It is important to note that a computer-encoded algorithm for fitting a continuum still relies heavily on human input. Choices must be made for the interval size over which estimators are calculated, the method of calculating the estimators, the inference method and rejection threshold for eliminating absorption features, and the method of fitting a function to the chosen estimator points. The only real advantage a program has over a human is that it can perform the chosen operations consistently, quickly, and with mathematical precision. Possible disadvantages of the automated approach include the fact that the chosen algorithms may not be sophisticated enough to deal with the complexities of the input.

Several researchers rely on computer-encoded algorithms for fitting continua. A typical process is the one described by Young *et al.* (1979). Briefly, the process involves the following steps:

1. Splitting the spectrum into lengths of consecutive pixels. Young *et al.* chose sections of spectrum 100 pixels long.
2. Calculating the observed standard deviation σ_{obs} of the photon counts in the pixels of each length of spectrum.
3. Calculating the theoretical standard deviation σ_{theory} for the same length of spectrum using photon statistics. This value is given by Equation 1.10 applied to the pixels being considered.
4. Comparing σ_{obs} to σ_{theory}. Because the spectrum contains absorption features which increase the variance of the photon counts, generally $\sigma_{\text{obs}} > \sigma_{\text{theory}}$. If this is the case, the pixel with the lowest photon count is removed.
5. Iterating steps 2–4 until $\sigma_{\text{obs}} \leq \sigma_{\text{theory}}$. The mean count of the remaining pixels is taken as a continuum estimator at the mean wavelength of those pixels. If the number of pixels remaining is below some threshold, no estimator is assigned. Young *et al.* set this threshold at 30 pixels.
6. Fitting splines to the set of continuum estimators produced in this way.

It is known that the process of continuum fitting described by Young *et al.* is inaccurate. Parnell and Carswell (1988) performed a series of simulations in

which they mimicked real QSO absorption line data with a S/N ratio of 15. For a number of absorption features consistent with observations in the redshift range $3.0 < z_{\rm abs} < 3.8$, they determined that the automatically fitted continuum levels were 5–10% below the true continuum.

It would be instructive to compare the results of interactive and automated continuum fitting on realistic simulated data. This would allow the assessment of the relative strengths and weaknesses of each process, and perhaps allow the construction of an algorithm incorporating the best features of each method.

2.1.2 Profile Fitting

There are dilemmas of a similar nature to those encountered in continuum fitting when attempting to fit Voigt profiles to noisy data.

At high S/N ratios, single absorption lines in the linear part of the curve-of-growth can usually be fitted accurately by an automated procedure (Carswell *et al.*, 1991). At low S/N ratios ($\sim$ 10 or less), however, the presence of a substantial noise contribution distorts the Fourier transform of the spectrum and often leads to spurious results (Rauch *et al.*, 1993). This is because the noise is uncorrelated (or very poorly correlated) between adjacent pixels, so absorption features usually appear more "spiky" than they really are. This leads to underestimates of the b value of the line.

Noise also makes a determination of equivalent widths difficult. The random nature of the noise implies a mean addition of zero to $W_{\rm obs}$ over the extent of a line, but it also affects the wavelength interval which the line appears to cover. A positive noise spike in the wing of a line often extends above the continuum level, so the extent of the line may be underestimated, resulting in a bias in the $W_{\rm obs}$ measurements. This is particularly problematic in the case of broad, shallow lines, where it is extremely difficult to determine how many pixels must be summed to best estimate $W_{\rm obs}$ because the line wings and continuum are virtually indistinguishable.

Further complications arise when lines are blended. A complex absorption feature must be deconvolved into its components. The shape of the feature is the only information available, from which must be determined the number of components and their Voigt parameters. Unfortunately, if the wrong number of components is chosen, the parameters λ, b, and N for the fitted components will all be biased, since these components must be made to fit the shape of the feature produced by a different number. Mistakes can have serious implications for further studies and conclusions based upon the data.

One option for an automated approach is to minimise the χ^2 value of the fitting residuals, using the minimum number of components necessary to achieve a fit which passes some statistical test at a selected significance level. Mathematically, it can be argued that this is the best procedure. However, it is possible, even highly probable, for the number of components chosen to be incorrect, because of statistical fluctuations. In some cases in which an automated program has fitted an incorrect number of components it may be obvious to an experienced human examining the

spectrum what the correct number of components is. The Voigt profile shapes are visually distinctive, so a trained eye has an advantage over a program, which must rely on the statistical behaviour of the numbers to make a determination.

As in the case of continuum fitting, it must be noted that automated line-fitting routines cannot be completely objective. The χ^2 example mentioned above is only one possibility for a test statistic, and a significance level for testing must also be chosen. Even with careful choices, mistakes obvious to a human can be made.

2.1.3 The Cloudy Night QSO Simulation

With the competing advantages and disadvantages of the automated and interactive approaches to analysing absorption spectra, it is important to know which method offers better performance and more reliable results.

To investigate the reliability of interactively fitting absorption line data, and to compare the results with those obtained by other procedures, interactive data reduction and analysis procedures were applied to a detailed simulation of a QSO spectrum. A separate group of researchers, Saskia Besier, John Webb (University of New South Wales), and Bob Carswell (Institute of Astronomy, Cambridge), applied automated routines to the same simulation. The goal was to compare the results both with each other and with the "true" answers, which were unknown to either group until their analysis was completed.

This approach is very instructive because it allows the comparison of "experimental" results with the true answers—a luxury seldom encountered in scientific work. Once the accuracy of the data analysis has been determined by this process, it is possible to apply that knowledge either to correct for systematic data reduction errors or, if such corrections are not feasible, to exercise an appropriate degree of caution in interpreting the results.

The simulated QSO spectrum used to test the data analysis procedures (dubbed the "Cloudy Night QSO", or CNQ) was supplied by Professor Edward Jenkins (Princeton University). He aimed to produce a synthetic spectrum as close as possible in character to one typically observed for a real QSO with current instruments. However, the parameters of the absorption lines present were generated randomly under a certain set of rules and were unknown until after the analyses.

The rules (drafted in consultation with Richard Hunstead) for Jenkins to follow were designed to produce a spectrum closely mimicking spectra obtained with an echelle spectrograph. In particular, spectra taken with UCLES and the IPCS detector on the AAT were used as a model, and the rules were chosen with the aim of reproducing all the instrumental effects of these devices. The goal of the simulation was to produce what looked like reduced echelle spectral orders, just prior to the continuum fitting step.

The rules used in generating the simulation were as follows:

1. The spectrum was to be presented in data files with pairs of numbers corresponding to wavelength in Angstroms and intensity in photon counts.

2. The "instrumental" velocity resolution was to be constant at $7\,\mathrm{km\,s^{-1}}$(gaussian FWHM) across the entire spectrum.
3. The bin size was to be determined by a constant $\Delta\lambda/\lambda = (2.5\,\mathrm{km\,s^{-1}})/c$ across the entire spectrum.
4. The spectrum was to be un-normalised, with a varying continuum level.
5. The continuum count levels were to vary within the range 60 to 600 counts, with a spectral shape consistent with real QSOs (*i.e.* including a reasonable Lyman α emission peak).
6. The spectrum was to be given in discrete sections, similar to echelle spectrograph orders, with a blaze-type fall-off in count level toward the edges of each section. Each section should cover approximately $5000\,\mathrm{km\,s^{-1}}$in velocity space, with small inter-order gaps.
7. Lyman α absorption lines from unknown (except to Jenkins) redshift, column density and velocity dispersion distributions were to be superimposed on the continuum. These Voigt profile lines were to be convolved with the chosen gaussian "instrumental" resolution of $7\,\mathrm{km\,s^{-1}}$FWHM. The number density of Lyman α lines was to be consistent with that observed in real QSOs, and the lines were to be distributed randomly in redshift.
8. Jenkins was to add some heavy element absorption systems, with a number density and an ionisation and velocity structure consistent with systems observed at similar redshifts to that of the CNQ.
9. Each simulated absorption line was to be logged by Jenkins, with its identification, wavelength, column density and velocity dispersion, for later comparison with the fitted results.
10. Poissonian noise based on the object count level plus an average sky count of 5 per bin was to be added. The noise was to be uncorrelated from one bin to the next.
11. The simulated spectrum was to cover the entire range from observed Lyman α emission to observed Lyman β emission, except for the small inter-order gaps.
12. Any details not specifically covered in the rules were to be chosen or extrapolated by Jenkins in a reasonable manner.

Jenkins produced the CNQ spectrum according to these rules and then passed on the result for analysis. He also provided a lower resolution spectrum, which could be considered to have been a pre-existing "observation" published in the literature. This low-resolution spectrum overlapped with the high-resolution spectrum and covered wavelengths to the red of the Lyman α emission peak, with an average S/N ratio of 40, and a gaussian resolution of $100\,\mathrm{km\,s^{-1}}$ FWHM. In Jenkins' words:

> The purpose of this spectrum is to give you a chance to identify the stronger metal line systems on the basis of lines that show up at low resolution for wavelengths to the red of your coverage. I feel that this is a fair piece of information to furnish for this test.

Table 2.1 Details of Cloudy Night spectral data. The mean count includes all pixels, both in the continuum and in absorption lines.

Order	$\lambda_{\rm min}$/Å	$\lambda_{\rm max}$/Å	Mean count
66	3392.8	3440.0	58.1
65	3442.2	3492.9	66.1
64	3496.0	3547.5	81.8
63	3551.4	3603.8	80.4
62	3608.7	3661.9	95.6
61	3667.9	3722.0	119.3
60	3729.0	3784.0	104.9
59	3792.2	3848.1	88.3
58	3857.6	3914.5	181.0
57	3925.3	3983.1	380.1
56	3995.4	4054.3	630.9
Low Res.	3399.1	5196.1	311.5

Jenkins also revealed, prior to my analysis, that he had included a variation in Lyman α line number density with redshift of the usual form:

$$\frac{d\mathcal{N}}{dz} \propto (1+z)^{\gamma}, \tag{2.1}$$

with a reasonable value of γ, which was not revealed. This effect was not explicitly included in the list of rules because it would be a small effect over the observed CNQ wavelength range, but Jenkins chose to include it. Jenkins stated that there was no attempt to simulate the "proximity effect", in which the number density of Lyman α lines is observed to fall off at redshifts close to the emission redshift of the QSO (see Section 1.5.4).

The CNQ had an emission redshift $z_{\rm em} = 2.30$. Details of the high resolution spectral orders and the low resolution spectrum provided by Jenkins are given in Table 2.1. The spectral orders themselves are shown unnormalised in Figure A.1, and after continuum fitting and normalisation in the lower panels of Figure A.3, both in Appendix A.

2.2 Continuum Determination and Line Fitting

The CNQ spectrum was analysed using the same procedures as used to analyse that of the real objects Q1101−264 and Q2348−147 (see Section 4.2.3).

Since the spectrum was supplied unnormalised, the first step was to fit a continuum level to each of the echelle orders. This was done in the spectral analysis program DIPSO (Howarth *et al.*, 1993). I judged the continuum fits by eye, selecting points in apparently line-free regions which the DIPSO routine CFIT then

interpolated with cubic splines[1].

The continua for all the orders were fitted independently. In a real spectrum this is done because there is no reason to assume the echelle blaze function is identical in each order. It was possible that Jenkins had used identical blaze functions in each order—whether this was the case or not was unknown at this stage—but no cross-checking of continuum fits between orders to exploit this possibility was done.

Max Pettini examined the continuum fits and advised on minor changes, some of which were incorporated. The data were then normalised by dividing by the adopted continua. The adopted continua, shown plotted on the unnormalised spectral orders, are shown as solid lines in Figure A.1 in Appendix A.

Likely absorption lines were picked from the echelle spectral orders and the low resolution spectrum by visual inspection. Wavelengths and equivalent widths of these lines were measured using the EW command in DIPSO. Lines with equivalent widths above 6σ, with σ given by the EW command based on the unnormalised continuum level near each line, were compiled into a list of detected lines. Apparent lines close to the 6σ limit were measured in an effort to ensure no lines above the cut-off were missed.

Although every effort was made to locate absorption lines, it is nevertheless possible that some significant lines escaped detection. This is most likely to occur with broad, shallow lines, whose equivalent width is spread over many pixels, or lines heavily blended in the wings of very strong lines. In both these cases, illustrated in Figure 2.1, the presence of noise makes it difficult even to notice that a line may be present.

Heavy element systems were identified by searching for absorption line wavelength matches to metal lines at common redshifts. The procedure is described in detail in Section 5.2. Briefly, some systems were first identified by a visual examination for obvious metal lines such as C IV $\lambda\lambda$ 1548, 1550 doublets, but a computer program was then used to search for possible systems. Once a redshift system was identified, lines from common metal transitions were searched for at that redshift. Since metal lines often have $b < 10\,\mathrm{km\,s^{-1}}$, all unidentified narrow lines were cross-checked with each other to find any possible matches for metals at a common redshift. Some possible matches were rejected based on a lack of absorption in other transitions of a tentatively identified species, while others proved to be further systems. A few systems could not be conclusively established or rejected. Such lines were identified as metals, but annotated as having uncertain identifications—these are indicated as such in Table B.3. The narrow lines still unidentified were then tentatively assigned the redshift of each identified system in turn and checked against a metal line list to see if they could be a metal in a known system. A list of the metal lines used in these searches, taken from Morton *et al.* (1988), is given in Table B.1 in Appendix B.

The identified metal systems are listed in Table B.2 in Appendix B. Detailed

[1]The possibility of fitting continua with an automatic computer routine, and possible pitfalls in such a process, are discussed in Section 2.1.1.

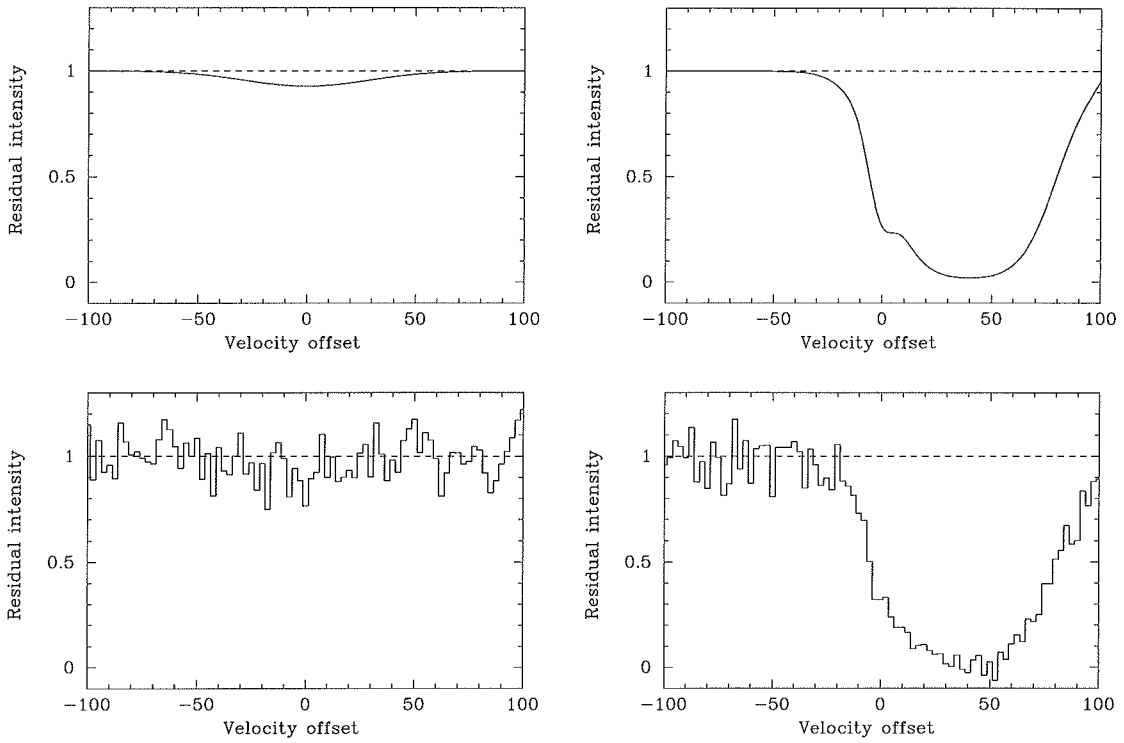

Figure 2.1 Selection effects in line detection. Each upper panel shows an absorption line (centred at 0 velocity offset) which would be difficult to detect in a noisy spectrum. *Left:* A single, weak line, with $b = 40\,\mathrm{km\,s^{-1}}$, $\log N = 12.6$, and $W_o = 21\,\mathrm{m\AA}$. Note where this line would fall in Figure 2.2. *Right:* A stronger, narrower line, blended in the wing of a strong line. The centre line has $b = 6$, $\log N = 12.8$, and $W_o = 27\,\mathrm{m\AA}$. The rest equivalent width for the centre lines in both panels is greater than that for several of the Lyman α lines actually detected in the CNQ spectrum. The lower panels show the same lines as they might appear in a real spectrum, with S/N ratio ~ 10 and $2.5\,\mathrm{km\,s^{-1}}$ pixels (the same pixel size as the CNQ spectrum).

descriptions of the identified metal line systems are given in Section 2.2.1. Lines not identified as due to metals were assumed to be due to Lyman α for the remainder of the analysis.

All of the absorption lines identified in the echelle spectrum (whether identified as metal lines or not) were fitted with Voigt profiles. The profile fitting was performed interactively with the program Xvoigt (see Appendix C). The general procedure was to fit each discrete absorption feature with as few Voigt components as possible in order to achieve a visually acceptable fit. This involved examining the residues after dividing out the fits to ensure they were consistent with Poissonian noise in amplitude and character. The noise level was taken into account when deciding whether or not extra components should be used to obtain a better fit. Generally, where the noise level was high fewer components were used than if the feature had been less noisy, since spurious structure may be introduced if the noise level is high

(Rauch *et al.*, 1993).

Where features were obviously composed of more than one line, overlapping only slightly, independent fitting would have produced spurious results because of the inclusion of extra equivalent width from the neighbouring line. In such cases, the lines were fitted simultaneously to ensure the line parameters were valid.

All of the lines were fitted *as if they were Lyman α lines*. That is, the values of the velocity dispersion parameter b and the column density of absorbing material N were calculated under the assumption that the absorbing material was neutral hydrogen and the absorbing transition was Lyman α, *whether or not* the line had been identified as a metal. Furthermore, lines identified as being due to the same heavy element, ionisation, and redshift were fitted *independently*, not simultaneously with the constraint of having identical b and N values as would be expected. The reason for doing this was to obtain a larger data set of independently determined profile fits, for comparison with the true values of the line parameters (revealed later by Ed Jenkins).

Where necessary, the following transformations were used to convert between metal line parameters and line parameters determined under the assumption that the line was due to Lyman α:

$$b_{\rm metal} = b_{{\rm Ly}\alpha}, \tag{2.2}$$

$$\log N_{\rm metal} = \log N_{{\rm Ly}\alpha} + \log f_{{\rm Ly}\ \alpha} \lambda_{{\rm Ly}\alpha} - \log f_{\rm metal} \lambda_{\rm metal}. \tag{2.3}$$

Here $N_{\rm metal}$ is the true column density of metal ions, while $N_{{\rm Ly}\alpha}$ is the column density of H I which would be required to produce a line as strong as the given metal line.

Parameters were not recorded for saturated lines, because such parameters were uncertain by large amounts. Velocity dispersions and column densities of lines in the saturated region of the curve-of-growth are not uniquely determined by the profile or equivalent width, so determining these parameters to any reasonable accuracy is not possible.

The results of the profile fitting are shown in Table B.3 in Appendix B.

2.2.1 Description of Metal Line Systems

This Section describes each of the metal line systems identified in the CNQ spectrum.

$z_{\rm abs} = 0.000$

This system is identified by the calcium H and K lines (Ca II $\lambda\lambda$3934, 3969), which are the only lines present. There are two velocity components within $\sim 10\,{\rm km\,s^{-1}}$ of $z_{\rm abs} = 0$, plus a possible additional weak component $\sim 30\,{\rm km\,s^{-1}}$ blueward. The weak component is seen in the λ3969 line, but the matching component in the λ3934 line is blended with a strong Lyman α line. This system corresponds to calcium absorption at zero redshift within the Galactic halo.

$z_{\rm abs} = 0.250$

This is only a possible system, tentatively identified by the presence of what may be a Mg II λ2803 line. The corresponding λ2796 line would fall in an inter-order gap, and no other strong lines are expected in the spectrum. This tentative identification is based on the strength and narrowness of the identified line, which look characteristic of a metal line. It has been marked as a metal line to avoid possible contamination of the Lyman α line sample.

$z_{\rm abs} = 0.252$

This is only a possible system, tentatively identified by the presence of what may be a Mg II λ2803 line. The corresponding λ2796 line would fall in the centre of a saturated Lyman α line, and no other strong lines are expected in the spectrum. This tentative identification is based on the strength and narrowness of the identified line, which look characteristic of a metal line. It has been marked as a metal line to avoid possible contamination of the Lyman α line sample.

$z_{\rm abs} = 0.364$

This is a very complex system, containing at least eight components identifiable in the Mg II $\lambda\lambda$2796, 2803 transitions. Intermingled with the λ2796 lines are some Si III λ1206 lines of the $z_{\rm abs} = 2.160$ system. Overlapping several of the λ2803 lines are strong Al III λ1862 lines from the $z_{\rm abs} = 1.052$ system, and a broad Lyman α line. Several additional likely components are present in the λ2796 line, with their λ2803 counterparts unidentifiably blended with these additional lines. There is also a possible Fe II λ2600 line, matching extremely well in redshift with one of the strongest Mg II lines (visible in both transitions). Other possible Fe II lines are either blended with a strong Lyman α feature or in an inter-order gap. Since a good match for more than one Fe II line cannot be found, this line is marked as only a possible identification.

$z_{\rm abs} = 1.052$

This system contains three readily identifiable components, all present in C I λ1656, Al II λ1670, Ni II $\lambda\lambda$1709, 1741, 1751, Si II λ1808, and Al III $\lambda\lambda$1854, 1862. Some of the stronger lines are saturated with the two bluemost components partially blended. A tentative identification is placed on three lines which match extremely closely the expected positions of Mg I λ1827, but which are quite strong when this transition is only expected to be seen weakly, if at all. The two strongest components are also seen in Fe II $\lambda\lambda$2344, 2374, 2384, and one component in Fe II λ2260 in the low resolution spectrum.

$z_{\rm abs} = 1.199$

This system is based on a single C IV $\lambda\lambda$1548, 1550 doublet. The redshift match is excellent, and the doublet ratio is consistent with expectation, but no other lines are seen with this redshift. There is a narrow feature which precisely matches the expected location for Al II λ1670, but it is not significant at the chosen 6σ detection level for lines.

$z_{\rm abs} = 1.302$

This system contains four strong C IV $\lambda\lambda$1548, 1550 components, with the central two being heavily blended and perhaps containing additional components. There are also two Si II λ1526 lines matching the strongest C IV components. One might also expect to see Al II λ1670 in this system, but the appropriate wavelengths are covered by a strong Lyman α line.

$z_{\rm abs} = 1.502$

This is only a possible system, tentatively identified by the presence of what may be a C IV λ1548 line. The corresponding λ1550 line would fall near the centre of a strong Lyman α line, and no other strong lines are expected in the spectrum. This tentative identification is based on the strength and narrowness of the identified line, which look characteristic of a metal line. It has been marked as a metal line to avoid possible contamination of the Lyman α line sample.

$z_{\rm abs} = 1.523$

This is only a possible system, tentatively identified by the presence of what may be a C IV λ1548 line. The corresponding λ1550 line would fall near the centre of a blend of at least three other lines. There is one component of the blend close to the expected λ1550 wavelength, so this could be taken as supporting evidence for this system identification.

$z_{\rm abs} = 1.594$

This system contains a C IV $\lambda\lambda$1548, 1550 doublet and a corresponding Si IV λ1393 line. The wavelength matches and C IV doublet ratio are in excellent agreement.

$z_{\rm abs} = 1.604$

This system contains a C IV $\lambda\lambda$1548, 1550 doublet and Si II $\lambda\lambda$1304, 1526 lines. The Si II lines are offset by $\sim 10\,{\rm km\,s^{-1}}$ blueward of the C IV redshift, and the λ1526 line is blended with a stronger feature.

$z_{\rm abs} = 1.616$

This system has a strong component seen in O I λ1302, C II λ1334, Si IV λ1393, and C IV λ1548. There is an additional component seen in C IV, and there is a corresponding feature in O I, but the O I feature does not exceed the 6σ detection limit. Although the additional component might be expected to be visible in Si IV, no feature is discernible, but this may be because of S/N effects. The corresponding C IV λ1550 line is beyond the red end of the echelle spectrum, but corresponds to a feature of appropriate strength in the low resolution spectrum. The expected position of Si II λ1304 is within a strong Lyman α feature, Si II λ1526 falls in an inter-order gap, and there is no discernible Al II λ1670 feature.

$z_{\rm abs} = 1.791$

This is a system with a complex structure of at least six components visible in the Si II λ1260 and Si IV λ1393 lines. An additional component in the Si II line is lost in a saturated Lyman α line at the Si IV wavelength, and there is an obvious weak Si IV component which is not evident in Si II, possibly because of noise. The strongest component is also visible in N V $\lambda\lambda$1238, 1242, O I λ1302, Si II λ1304. A saturated blend of Lyman α components appears at the bluemost end of the echelle spectrum, with some components lost outside the wavelength coverage.

$z_{\rm abs} = 1.864$

This system consists of a strongly saturated Lyman α line and three Si III λ1206 lines. The Lyman α is presumably a blend of separate lines, but no structure is discernible. A C IV $\lambda\lambda$1548, 1550 doublet is present in the low resolution spectrum. No Si II $\lambda\lambda$1260, 1526 or Al II λ1670 are visible.

$z_{\rm abs} = 1.924$

This system comprises a saturated Lyman α line and corresponding Si II λ1260 and Si III λ1206 lines. There is also a C IV $\lambda\lambda$1548, 1550 doublet in the low resolution spectrum.

$z_{\rm abs} = 2.006$

This is another system identified by a saturated Lyman α line and a C IV $\lambda\lambda$1548, 1550 doublet in the low resolution spectrum. There are also Si III λ1206 and C II λ1334 lines present. The expected position of Si II λ1260 falls in an inter-order gap.

$z_{\rm abs} = 2.160$

This system incorporates two components at $z_{\rm abs} = 2.1593$ and $z_{\rm abs} = 2.1612$. Each component is seen in N I $\lambda\lambda$1134, 1200, Fe II λ1144, Si II λ1193, and Si III λ1206. The bluemost component is also seen in Si II λ1260 (the redmost component is beyond

the red end of the echelle spectrum) and the redmost is seen in Si II λ1190 (where the bluemost component is blended with a strong Lyman α line). Both components have strongly saturated (or even slightly damped) Lyman α lines, which are blended and resemble a single line. The line D I λ1215, corresponding to the Lyman α transition in deuterium, is seen in the wing of the Lyman α lines for the bluemost component. Although there is a feature close to the expected wavelength of C IV $\lambda\lambda$1548, 1550 in the low resolutions spectrum, that feature appears to be entirely due to Fe II λ2382 in the $z_{\rm abs} = 1.052$ system. If C IV is present in this system it is extremely weak.

$z_{\rm abs} = 2.164$

This is another system with very strong Lyman α lines and multiple components. Three components are seen in Si II $\lambda\lambda$1190, 1193, N I λ1199, and Si III λ1206. Two components are seen in Fe II λ1144 and N I λ1200. Many of these lines are blended with other features—either metal lines in other systems or broad lines which are most likely Lyman α. The deuterium line D I λ1215 is seen for the bluemost component. Like the $z_{\rm abs} = 2.160$ system, there appears to be no evidence for the existence of the C IV $\lambda\lambda$1548, 1550 doublet in the low resolution spectrum.

$z_{\rm abs} = 2.202$

This system contains a heavily saturated Lyman α line as well as Si III λ1206 and Si II λ1260. The Si II line is weak, and there is no sign of the weaker Si II $\lambda\lambda$1190, 1193 doublet. There is a C IV $\lambda\lambda$1548, 1550 doublet in the low resolution spectrum.

$z_{\rm abs} = 2.229$

This system has only a saturated Lyman α line and a Si III λ1206 line in the echelle spectrum. The expected position of the Si II λ1260 line is beyond the red end of the echelle spectrum, and the $\lambda\lambda$1190, 1193 doublet lines are in an inter-order gap and blended with a Lyman α feature respectively.. There is a C IV $\lambda\lambda$1548, 1550 doublet in the low resolution spectrum.

2.3 Distributions of Lyman α Lines

The Lyman α lines from Table B.3 were formed into two samples:

Sample 1 All the Lyman α lines (205 lines).

Sample 2 The subset of Lyman α lines with well-determined parameters (i.e. those not marked as uncertain in Table B.3) (146 lines).

These samples were used to examine various statistical properties of the measured absorption line parameters. The remainder of this Section deals only with these samples of Lyman α lines.

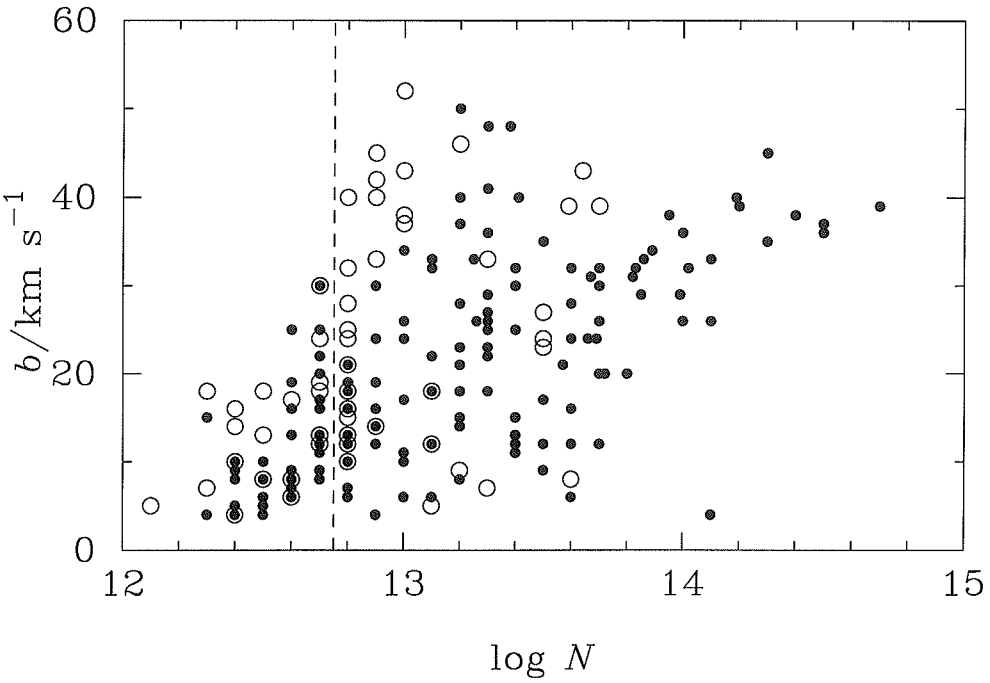

Figure 2.2 Relationship between fitted values of b and $\log N$ for lines identified as Lyman α in the CNQ spectrum. The filled circles represent lines with well-determined parameters, while the open circles represent those lines marked as uncertain in Table B.3. There is an obvious "quantisation" of the points into discrete vertical and (not so obvious) horizontal lines. This is purely an artefact of the fitting procedure, in which $\log N$ was usually rounded to the nearest 0.1 and b to the nearest integer. The uncertainties in $\log N$ are mostly 0.1 and in b always > 1, so the quantisation is certainly not real and can be ignored. The conspicuous outlier point at $\log N = 14.1$ and $b = 4$ is discussed in Section 3.6. The dashed line shows the adopted completeness limit of $\log N \geq 12.75$, discussed in Section 2.3.2.

2.3.1 Correlation of b and $\log N$

The fits to the lines identified as Lyman α are shown plotted on the b–$\log N$ plane in Figure 2.2.

To determine whether the distributions of b and $\log N$ were independent or correlated, linear regression analyses were carried out for the two functional forms used in Pettini *et al.* (1990):

$$b = A + B \log N \tag{2.4}$$

and

$$N = A' b^{B'}. \tag{2.5}$$

The fitted values for A, B, A', and B' in each sample are shown in Table 2.2, along with the calculated correlation coefficient r. The resulting correlation coefficients have very small ($\ll 10^{-5}$) probabilities of occurring in uncorrelated data of the same number of points, so the null hypothesis that the b and $\log N$ values are not

Table 2.2 Results of regression analyses between the measured b and $\log N$ values for the Lyman α lines in the CNQ spectrum. Note that the uncertainties in the fit parameters are correlated.

$b = A + B \log N$				
Sample	No. lines	A	B	r
1	205	-207 ± 7	17.2 ± 0.6	0.524
2	146	-216 ± 9	17.8 ± 0.7	0.621

$N = A' b^{B'}$				
Sample	No. lines	A'	B'	r
1	205	$(5 \pm 2) \times 10^8$	3.34 ± 0.15	0.509
2	146	$(8 \pm 5) \times 10^8$	3.23 ± 0.17	0.592

correlated can be rejected with high confidence. The highest correlation coefficient occurs for the well-determined subsample (Sample 2), fitted to Equation 2.4.

2.3.2 The Lyman α Column Density Distribution

A maximum likelihood (ML) analysis was performed on the column density values of the Lyman α samples to derive the parameters A_0 and β in the canonical number density distribution (Carswell *et al.*, 1984)[2]:

$$d\mathcal{N} = A_0 N^{-\beta} dN. \tag{2.6}$$

The ML method has maximal efficiency and the advantage that it does not require any binning or other loss of information from the data set. The results, for various subsamples defined by $\log N$ cut-offs, are shown in Table 2.3 and a histogram of the distribution with two selected fits to Equation 2.6 is shown in Figure 2.3. Note that the ML calculations assume that lines of different strengths have equal probabilities of being detected. If this assumption is true, the fit parameters should be constant for different subsamples. If it is not (*i.e.* there is some incompleteness of the data), the parameters will change between subsamples.

From these results, it appears the Lyman α line list is complete down to a value of $\log N \sim 12.75$, and that the value of $\beta \sim 1.7$. This judgment is based on the assumption that the N distribution follows a simple power law (as given by Equation 2.6), and the observation that the fitted slope drops sharply when values of $\log N < 12.75$ are considered. Also, as shown in Figure 2.3, the $\beta = 1.71$ fit clearly describes the data with $\log N > 12.75$ well, whereas the power law fits for the ranges $\log N > 12.00$ (also shown) and $\log N > 12.50$ (not shown on the figure, for clarity) are poor fits to the observed distribution in their respective $\log N$ ranges.

[2]The column density distribution for real data was introduced in Section 1.5.1 and is discussed in detail in Section 6.2.

Table 2.3 Parameters derived from ML fits of the CNQ Lyman α lines to Equation 2.6. The parameter β' is the estimated value of β after the line blanketing correction described in Section 2.3.2.

Sample	log N range	No. lines	A_0	β	β'
1	12.00–15.00	205	4.9×10^4	1.24 ± 0.04	1.26 ± 0.04
1	12.25–15.00	204	5.9×10^6	1.39 ± 0.04	1.41 ± 0.04
1	12.50–15.00	182	5.6×10^8	1.54 ± 0.06	1.57 ± 0.05
1	12.75–15.00	149	1.1×10^{11}	1.71 ± 0.07	1.73 ± 0.07
1	13.00–15.00	94	2.8×10^{10}	1.66 ± 0.10	1.70 ± 0.09
2	12.00–15.00	146	5.5×10^4	1.18 ± 0.04	1.18 ± 0.04

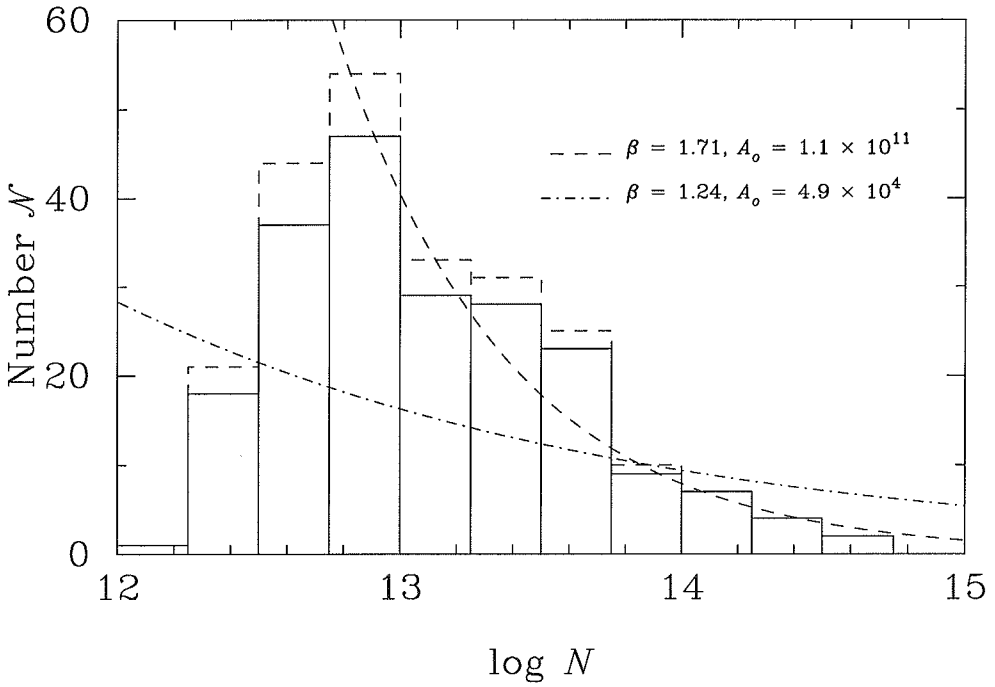

Figure 2.3 Number density histogram of the log N values for the CNQ Lyman α lines. The solid histogram shows the actual number of lines observed in each bin; the dashed histogram shows the number after correction for the blanketing effect, given by Equation 2.8. The curved lines show two of the ML parameter fits to the observed distribution (solid histogram) from Table 2.3.

Incompleteness is caused by three effects:

1. The difficulties of detecting weak lines against a noisy continuum.
2. The blending of weak lines. This may cause two or more weak lines to be identified and fitted as a single line of somewhat larger column density, thus lowering the number of weak lines identified as such. The resulting fit will be a somewhat stronger line, which may also skew the distribution slightly, causing more lines to appear just above the completeness threshold.
3. The blanketing effect of strong lines. Any weak lines in the wings or saturated cores of strong lines may simply not be detected. This effectively reduces the wavelength region over which it is possible to detect weak lines.

It is important to note that incompleteness is implied only if the adopted power law form for the distribution (or one similar in shape, with no turnover at low column densities) is correct. However, because of the difficulties in detecting weak lines discussed above, some level of incompleteness is to be expected, becoming progressively more important with decreasing $\log N$. Given this, and the acceptable fit to a simple power law at high $\log N$, there is no justification for invoking a more complex form for the distribution function.

There are many possibilities for choosing more complex forms of the distribution function (such as a broken power law, as discussed in Section 6.2), but the data do not allow for discriminating between them. It is also important to realise that in a real QSO observation any functional form will merely be an empirical description of the data, unless there is a theoretical model which predicts a particular form. Since there are no widely accepted theories which predict a shape for the column density distribution, attempting to fit a complex form is not worthwhile unless the data are good enough to rule out a simple power law.

The Effect of Line Blanketing on β

An attempt was made to estimate how severely the apparent value of β may be influenced by the blanketing effect mentioned above. Since strong lines (both metal and Lyman α) blanket a wavelength interval approximately equal to their $W_{\rm obs}$, the wavelength interval $\Delta\lambda(\log N_0)$ over which lines with $\log N_{{\rm Ly}\alpha} > \log N_0$ can be detected is given by

$$\Delta\lambda(\log N_0) \simeq \Delta\lambda(0) - \sum_{\log N_{{\rm Ly}\alpha} > \log N_0} W_{\rm obs}, \tag{2.7}$$

where $\Delta\lambda(0)$ is the total wavelength range covered by the spectrum and the sum is over both metal and Lyman α lines. The values of $N_{{\rm Ly}\alpha}$ for metal lines are defined as in Equation 2.3, and are simply equal to N for Lyman α lines. For each of the bins in Figure 2.3, the number of lines which would have been detected if there was no blanketing, $\mathcal{N}_{\rm blanket}$, was estimated by

$$\mathcal{N}_{\rm blanket} = \frac{\Delta\lambda(0)}{\Delta\lambda(\log N_0)}\,\mathcal{N}. \tag{2.8}$$

These estimates (rounded to the nearest integers) are shown as the dashed histogram in Figure 2.3.

An attempt was made to estimate quantitatively the effect of line blanketing on the calculated value of β. The ML estimator could not be applied directly to the bin counts as modified by Equation 2.8 because it requires unbinned data. As an approximation, however, a number of extra, fictitious lines were distributed in the appropriate bins according to Equation 2.8, and evenly spaced in $\log N$ within each bin. When β was recalculated with this modified data set, the result in each of the $\log N$ ranges shown in Table 2.3 was a higher value of β, but only by about half the quoted uncertainty estimates. The results are shown in Figure 2.3 as β'. The differences between β and β', although not formally significant, imply that line blanketing does lower the apparent value of β somewhat, as might be expected from an examination of Figure 2.3, and that it is possible to obtain a quantitative estimate of the effect.

Assuming a power law gives a valid description of the N distribution, the line blanketing correction has not significantly changed the incompleteness for $\log N < 12.75$, so it can be concluded that line blanketing is a minor cause of such incompleteness. The majority of lines at these column densities is then not detected because of either line blending or signal-to-noise effects.

2.4 Summary

Several of the problems which arise when analysing absorption lines in noisy spectral data have been described. In order to study the biases introduced by the effects of limited signal-to-noise ratios, Professor Ed Jenkins constructed a detailed simulation of a QSO spectrum—the "Cloudy Night QSO" spectrum—which was then analysed using procedures identical to those used on real data.

After continuum fitting, the parameters of the absorption lines were measured by Voigt profile fitting and heavy element absorption systems were identified. The observed correlation between Lyman α line velocity dispersion and neutral hydrogen column density was quantified. The distribution of column densities was analysed and fitted to the distribution function given in Equation 1.12. In calculating the power law index β, a method was developed to correct for line blanketing, in which strong lines can obscure weaker ones.

Chapter 3

A Cloudy Night Comparison

3.1 Introduction

When the analysis of the Cloudy Night QSO (CNQ) spectrum was complete (as far as analysis of a real spectrum), Ed Jenkins revealed the details which went into producing it. A full list of all the absorption lines present as well as their important parameters and identifications was supplied. Jenkins also provided the true continuum levels of the unnormalised data and noise-free, normalised spectral orders, showing all of the absorption lines. These figures were of great help in identifying absorption features and deciding which identifications corresponded to which line in the spectrum I studied. This noise-free spectrum is reproduced in the upper panels of Figure A.3 in Appendix A.

3.2 Analysis of the Continuum Fits

The true continuum shape of the CNQ data was generated by Jenkins in a manner designed to mimic carefully the blaze profile recorded in real echelle spectra at that time[1].

The starting point was an initial continuum based on a digital version of the spectrum of Q0143−015 from Sargent *et al.* (1989). A 50 point median filter was applied at every tenth point to eliminate the narrow absorption features. The resulting flux blueward of Lyman α emission was adjusted upwards to compensate for the difference in average continuum depression (D_A as defined by Oke and Korycansky, 1982) caused by Lyman α absorption between the Q0143−015 $z_{\rm em} = 3.138$ and the CNQ $z_{\rm em} = 2.300$. This coarse continuum was interpolated by spline fits to give continuum values at each of the synthetic spectrum wavelength points.

The echelle blaze efficiency was simulated by segmenting the continuum into separate orders, then multiplying by a sinc^2 function centred slightly to the right of the centre of each order. Finally, the loss of detector efficiency and increased effect

[1]Recent data from the 10 m Keck Telescope have higher count levels and thus S/N ratios than the AAT echelle data studied in this thesis.

Table 3.1 The mean true and fitted continuum levels for each order of the Cloudy Night QSO spectrum, the ratios of fitted to true levels, and typical true S/N ratios.

	Mean continuum			Typical
Order	True	Fitted	Ratio	S/N
66	75.9	72.1	0.950	8.2
65	81.6	74.1	0.909	8.5
64	101.7	93.9	0.923	9.6
63	107.8	100.7	0.934	9.9
62	115.1	106.7	0.928	10.3
61	146.4	136.3	0.931	11.7
60	147.7	139.2	0.943	11.7
59	170.2	153.8	0.904	12.7
58	230.6	220.5	0.956	14.9
57	462.6	444.4	0.961	21.3
56	662.1	658.1	0.994	25.5

of atmospheric absorption to the blue was mimicked by multiplying each order by a linearly interpolated value ranging from 1 at order 56 to 0.25 at order 66.

This synthetic continuum was then used to multiply the noise-free, normalised spectrum, and gaussian noise, adjusted appropriately to simulate the increased noise caused by sky subtraction, was added to produce the spectral orders which I analysed.

The unnormalised data, shown with both the true continuum levels and my adopted continuum fits, are shown in Figure A.1 in Appendix A.

3.2.1 Comparison of True and Fitted Continua

As Figure A.1 shows, there are considerable differences between the adopted continuum levels and the true levels. The adopted fits generally improve with increasing S/N, but are always too low except for two regions in order 56 where the fitted continuum rises slightly above the true level.

The mean fitted and true continuum levels for each order, as well as the ratios of mean fitted level to true level, are shown in Table 3.1. The final column of the Table gives the typical S/N ratio of each order, calculated using Equation 1.10 with

$$n_{\rm obj} = n_{\rm sky} = 1, \tag{3.1}$$

$$\mathcal{N}_{\rm obj} = \text{True mean continuum}, \tag{3.2}$$

$$\mathcal{N}_{\rm sky} = 5, \tag{3.3}$$

as appropriate for the CNQ data. Generally, the fitted continua have been set $\sim$ 5–10% too low, similar to the findings of Parnell and Carswell (1988) using an automated continuum fitting procedure.

The trend of improved fits with increased count level (and hence S/N) can be seen from the ratios in Table 3.1. The two obvious exceptions to this are order 59, which has strongly saturated lines at the red end which provide no unabsorbed reference for fitting a continuum, and order 66, which has a better fit than might be expected from the trend because it has some large regions of relatively unabsorbed continuum compared with other orders.

In general, regions dense with absorption lines have the most poorly defined continuum levels. Relatively large regions which appear to be unabsorbed provide the best fits to the true continuum level, but even these have almost invariably been set too low.

The presence of numerous, low column density absorption lines has clearly affected the perceived "unabsorbed" continuum level. There are very few regions of more than a few ($\sim$ 10) pixels where the mean flux is greater than or equal to the true continuum level. This is shown in Figure A.2 in Appendix A, in which an eleven-point mean box-filter has been applied to the spectrum. In a spectrum with no absorption, half the points should lie above the true continuum level.

In reality, very few points of the box-filtered spectrum lie on or above the true continuum level. A few small "spikes" in each order reach or exceed the continuum but mostly the continuum seems to define a loose upper bound on the intensities of the box-filtered spectrum.

This observation suggests a new, objective, and perhaps superior method of fitting continua to noisy absorption line spectra such as that of the CNQ. One method of automatically fitting continua (that of Young *et al.*, 1979) has been described in Section 2.1.1, but that method is known to be unreliable. The new method would involve mean-filtering the spectral data, then taking a suitably defined set of locally maximal points and fitting a low order polynomial to them. This may produce a more accurate continuum than the method described in Section 2.2. Such a method would, however, still have difficulties coping with heavily absorbed regions such as, for example, at the red ends of orders 64, 59, and 58 of the CNQ.

Attempting to use an automated algorithm to fit continua to the CNQ spectrum would be an instructive exercise, since the true continuum levels are known, but this was not done in this study. It is hoped that a comparison can be made with an automated continuum-fitting procedure used by Saskia Besier, John Webb, and Bob Carswell in their analysis of the CNQ spectrum.

3.3 Analysis of the Line Fitting Procedure

A comparison was done to find the true parameters of each of the lines previously identified in the spectrum. The line list shown in Table B.8 was compared to the line list supplied by Jenkins and each true line which contributed a significant amount of equivalent width to a line identified by me was noted. The results are shown in Table B.4 in Appendix B. The 362 lines identified by me fall into three distinct categories:

Table 3.2 Statistics of line parameter fitting errors normalised by uncertainty estimates in the Cloudy Night QSO. The last five columns are respectively the sample mean (μ), standard deviation (σ), mean absolute deviation ($D_{\rm mean}$), median absolute deviation from the median divided by 0.65 ($D_{\rm med}$), and fraction of values with modulus > 1 ($f_{>1}$).

Parameter	Min.	Max.	μ	σ	$D_{\rm mean}$	$D_{\rm med}$	$f_{>1}$
$(b_{\rm fit} - b_{\rm true})/\Delta b$	−13.85	3.44	−0.465	2.02	1.22	1.04	0.337
$(\log N_{\rm fit} - \log N_{\rm true})/\Delta \log N$	− 6.77	6.77	0.162	1.37	0.92	0.92	0.307

1. 31 lines were saturated and so no parameter measurements were made by me.
2. 129 lines were found to be composed of two or more blended lines with significant equivalent width from Jenkins' list. In these cases, comparison of the fitted parameters with the true parameters is complicated by the fact that a blend has been fitted as a single line. These lines are not considered in this Section, but are discussed again in Section 3.7.
3. 202 lines were found to correspond to a single significant line from Jenkins' list. In these cases a direct comparison between the measured line parameters and the true ones could be made.

In order to study the accuracy of the Voigt profile fitting procedure with the largest possible sample of lines, the identifications of the species producing the lines were ignored and a sample of all the non-saturated lines which corresponded to a single line from Jenkins' list (202 lines) was compiled. Column densities and velocity dispersions for this sample were set assuming the lines were Lyman α using equations 2.2 and 2.3 where necessary. This sample is discussed in the remainder of Section 3.3.

3.3.1 Examining the Uncertainty Estimates

In order to examine the appropriateness of my uncertainty estimates for b and $\log N$, the differences between the fitted values and the true values were normalised by dividing by the uncertainties. For a "reasonable" distribution of random measurement errors, the distribution of these normalised values should resemble a normal distribution of zero mean and unit variance[2]. Several statistics based on these normalised samples are shown in Table 3.2.

The sample means are a simple method of investigating systematic bias in the errors. A more sophisticated analysis is carried out in Section 3.3.2, so it will not be pursued here.

The standard deviation is an indicator of the spread in normalised error values. It is significantly greater than 1 for both parameter samples, indicating a greater than normal variance in the distribution of errors. However, the sample standard

[2]It should be noted that errors in b and $\log N$ are not independent variables. For a line of given equivalent width, the errors in b and $\log N$ are anticorrelated.

deviation is not a very robust estimator of the distribution standard deviation, being subject to heavy influence from outlying data values.

Two more robust estimators of the error distribution's standard deviation were calculated, the mean absolute deviation,

$$D_{\text{mean}}(x_1, \ldots, x_N) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|, \tag{3.4}$$

where $\overline{x}$ is the mean of the values $x_1, \ldots, x_N$, and the median absolute deviation from the median, divided by 0.65,

$$D_{\text{med}}(x_1, \ldots, x_N) = \frac{1}{0.65}(\text{the median of the values } |x_i - \tilde{x}|), \tag{3.5}$$

where $\tilde{x}$ is the median of the values $x_1, \ldots, x_N$. The factor of $1/0.65$ is required to convert the median absolute deviation from the median into an estimator for the standard deviation (Rice, 1988).

The values of these robust statistics, shown in Table 3.2, are similar for each parameter, and are significantly smaller than the σ value, indicating that outlying points have strongly influenced the value of σ. Also, each of the robust estimators is close to 1, which is the expected value for a normal distribution.

Another simple method of examining the dispersion of the fitting errors is to calculate the fraction of normalised errors with modulus > 1. This number, designated $f_{>1}$, is 0.317 for a normal distribution. Like the robust statistics described, it is also relatively insensitive to outliers. Again, the values shown in Table 3.2 are close to that expected from the normal distribution.

The conclusion from these results is that the uncertainty estimates of the fitting errors in both b and $\log N$ are good approximations to 1σ uncertainties, but the actual distribution of fitting errors is not well approximated by the normal curve. This can be seen by plotting histograms of the normalised fitting errors, overlaid with gaussian curves of identical area and standard deviation, and mean $\mu = 0$, as shown in Figure 3.1 and Figure 3.2. As seen in the figures, there are gross outlying points and the centre of the distribution in Figure 3.2 appears to be more peaked than a gaussian. These features imply leptokurtic distributions, and indeed both the $(b_{\text{fit}} - b_{\text{true}})/\Delta b$ and $(\log N_{\text{fit}} - \log N_{\text{true}})/\Delta \log N$ distributions have large, positive kurtoses: 14.5 ± 0.7 and 5.2 ± 0.7 respectively.

The outlying points are discussed in detail in Section 3.4.

3.3.2 Testing for Bias

Wilcoxon signed-rank tests (Hollander and Wolfe, 1973) were carried out to determine if the measured values of line wavelengths, b, and $\log N$ were biased with respect to the actual values supplied by Jenkins. A 95% significance level was chosen *a priori* for the rejection of the null hypothesis that there is no systematic measurement bias. The results are summarised in Table 3.3. For large samples, as is the

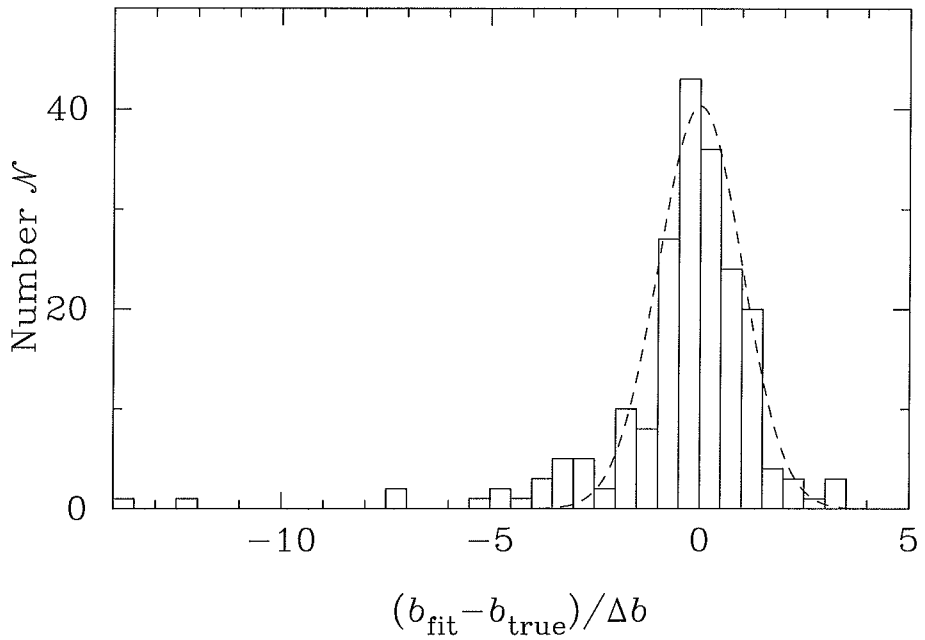

Figure 3.1 The distribution of the b fitting errors, normalised by the fit uncertainty estimates. The dashed curve is a gaussian with area and standard deviation equal to the distribution of all the data (including outliers), and mean $\mu = 0$.

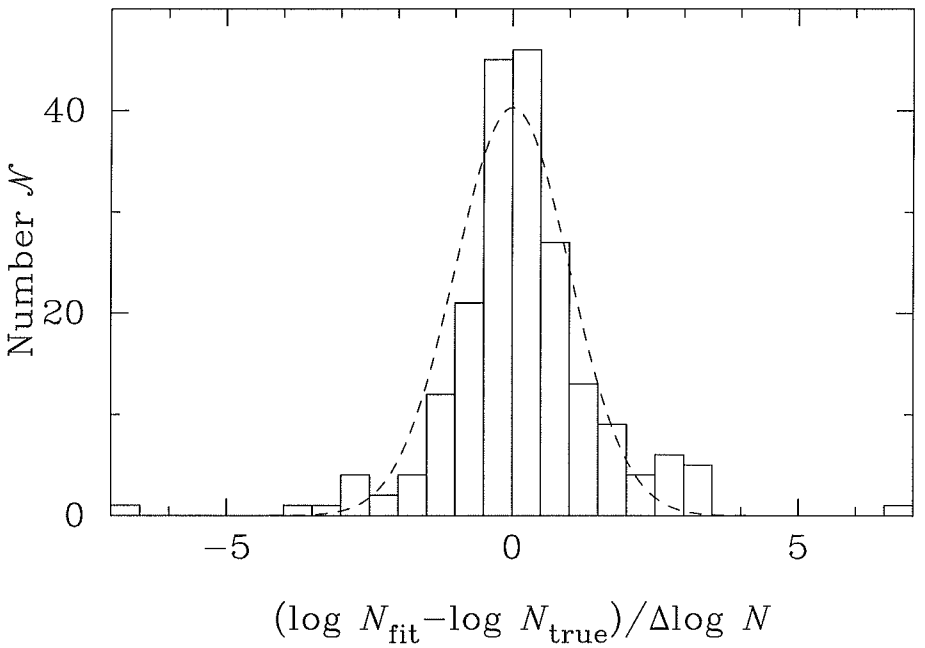

Figure 3.2 The distribution of the $\log N$ fitting errors, normalised by the fit uncertainty estimates. The dashed curve is a gaussian with area and standard deviation equal to the distribution of all the data (including outliers), and mean $\mu = 0$.

Table 3.3 Results of Wilcoxon signed-rank tests on differences between measured and actual values of various parameters for non-saturated CNQ lines.

Parameter	Number of non-zero differences	W_+	$Z(W_+)$	p
$(\lambda_{\rm fit} - \lambda_{\rm true})$	193	9199	−0.208	0.835
$(b_{\rm fit} - b_{\rm true})$	200	8057	−2.432	0.015
$(b_{\rm fit} - b_{\rm true})/\Delta b$	200	8252	−2.194	0.028
$(\log N_{\rm fit} - \log N_{\rm true})$	202	11565	1.579	0.114
$(\log N_{\rm fit} - \log N_{\rm true})/\Delta \log N$	202	11815	1.880	0.060

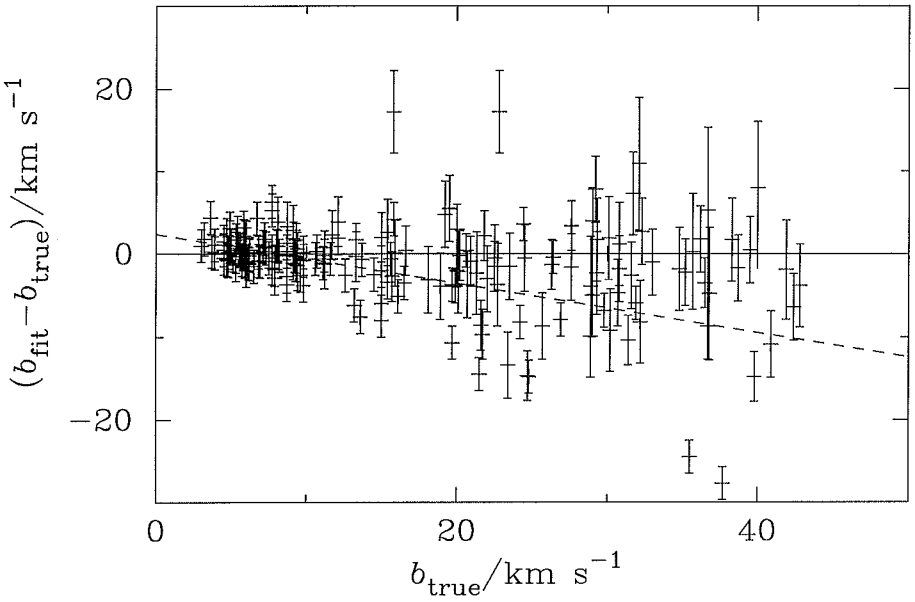

Figure 3.3 Plot of differences between fitted and true b values versus the true b values for the CNQ.

case here, the Wilcoxon test statistic W_+ is normally distributed with known mean and variance. The value $Z(W_+)$ is the statistic normalised to zero mean and unit variance; p is the two-tailed probability of W_+ falling as far from the expected mean under the null hypothesis as measured.

The signed-rank tests show that there is no reason to reject the null hypothesis in the cases of either wavelength (λ) or $\log N$ measurements, but the b measurements are shown to be biased at the 98.5% significance level. There are differences between the results for the raw measurements and those normalised by the uncertainty estimates, but not large enough to change the outcome of the hypothesis tests.

The differences between the measured and true values of b and $\log N$ are plotted against the true values in Figure 3.3 and Figure 3.4 respectively. It is clear from Figure 3.3 and the sign of $Z(W_+)$ in the signed-rank test that the measured b values are biased towards smaller values. A linear least-squares regression of $b_{\rm fit} - b_{\rm true}$

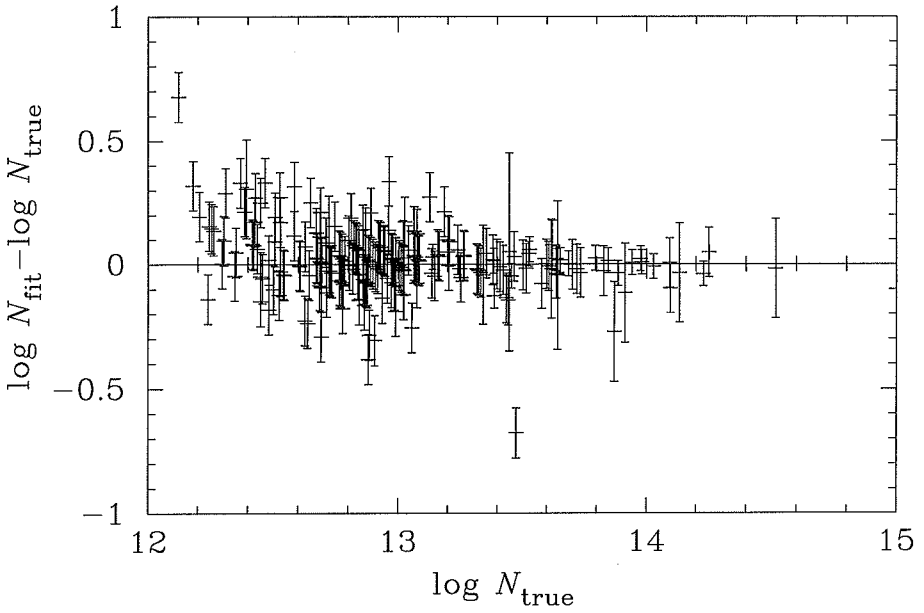

Figure 3.4 Plot of differences between fitted and true $\log N$ values versus the true $\log N$ values for the CNQ.

against $b_{\rm true}$ gives a slope of -0.30 ± 0.02, indicating that the bias is greater for lines which are intrinsically broader; *i.e.* lines with a larger b value are more likely to have underestimated b measurements than lines with smaller b values.

Attempted Calibration of b Measurements

With a systematic bias in the measurement of b values revealed, an effort was made to produce a calibration curve to correct the measured values. The distribution of $b_{\rm fit} - b_{\rm true}$ against $b_{\rm fit}$ was examined. A plot of this distribution is shown in Figure 3.5.

In this case, least squares regression to the line $b_{\rm fit} - b_{\rm true} = A + B b_{\rm fit}$ produced the values $A = -1.1 \pm 0.3$ and $B = 0.00 \pm 0.02$. The non-parametric correlation statistics Spearman's ρ and Kendall's τ had values consistent with no significant correlation. These results mean it is not possible to calibrate the slope of the $b_{\rm fit} - b_{\rm true}$ versus $b_{\rm true}$ curve by using the measured values $b_{\rm fit}$. However, the value of A shows that some correction may still be made by adding a constant value of $1\,{\rm km\,s^{-1}}$ to all the measured b values. The value of doing this is questionable, though, since the scatter observed in the b differences is many times greater than the size of this possible correction. The size of this correction may also be dependent on the particular spectrum examined, so it is doubtful that any meaningful correction can be applied to QSO spectra in general.

Trying one more avenue for a possible significant correction, the distribution of the b differences against the fitted $\log N$ values was examined. This is shown plotted in Figure 3.6. In this case the line $b_{\rm fit} - b_{\rm true} = A + B \log N_{\rm fit}$ was fitted with

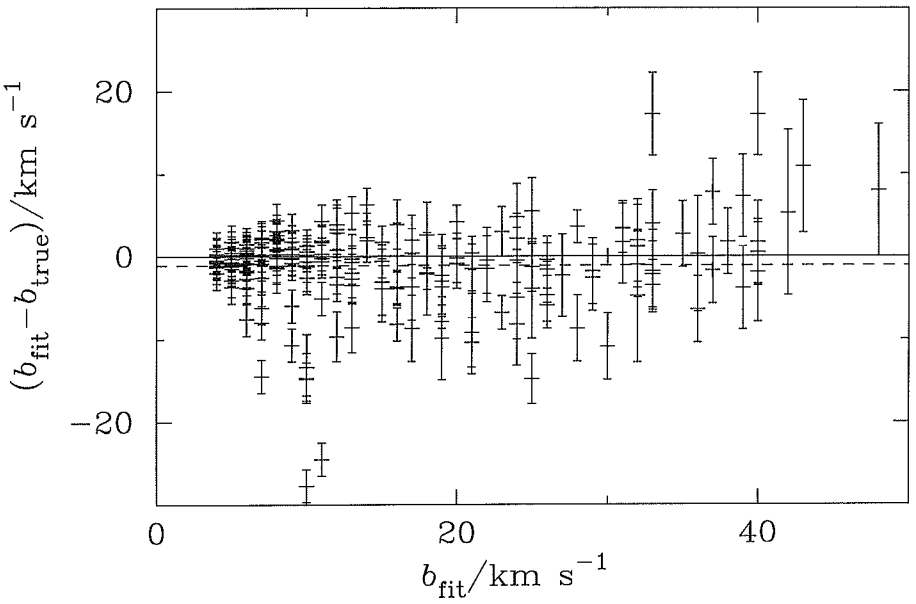

Figure 3.5 Plot of differences between fitted and true b values versus the fitted b values for the CNQ.

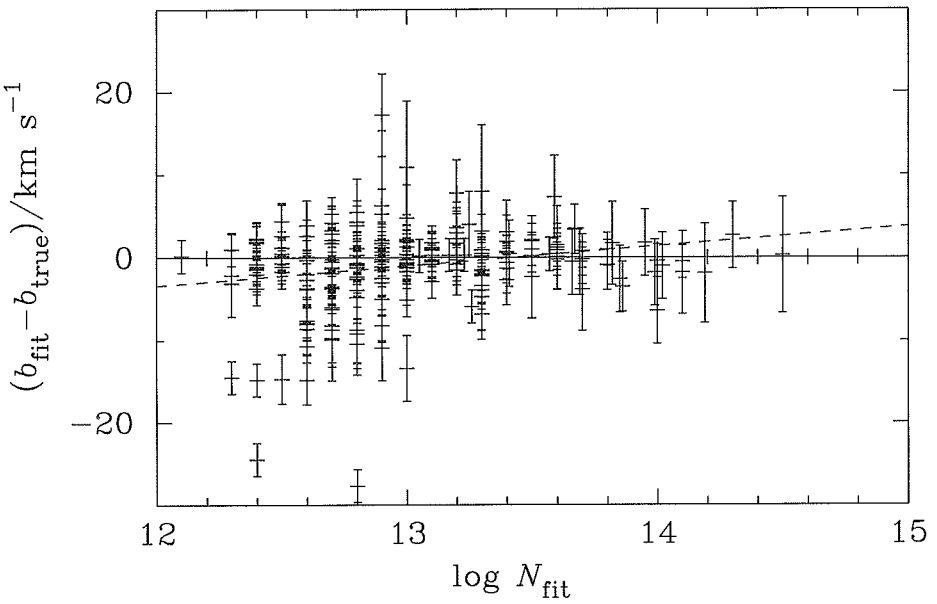

Figure 3.6 Plot of differences between fitted and true b values versus the fitted $\log N$ values for the CNQ.

parameters $A = -32 \pm 5$ and $B = 2.4 \pm 0.4$. The Spearman and Kendall parameters implied a significant correlation. However, the uncertainty in the value of A, as well as the large scatter evident in the plot, imply that any correction based on this fit would be scarcely any better than using the uncorrected values, if at all. Again, applying such a correction to data from real QSOs would be even more dangerous.

3.3.3 Migration of Lines in the b–$\log N$ Plane

Another instructive way of looking at the fitting errors is to see the effect they produce in the b–$\log N$ plane. Plots of the 202 lines in the sample of non-saturated lines with a single corresponding true line are shown in Figure 3.7. Allowing for the obvious structure produced in the upper panel by rounding the b values to whole numbers and many of the $\log N$ values to one decimal place, the two plots are very similar. The only clear difference is the existence of a handful of points in the top left region of the lower panel, which is conspicuously empty in the upper panel.

A method of more easily visualising the differences between these sets of data is to plot the fitted and true line parameters in the same b–$\log N$ plane and join points corresponding to the same absorption feature with a line. Such a diagram is shown in Figure 3.8.

This diagram reveals several interesting features. It contains three slightly overlapping regions with distinct differences in the plotted vectors.

- Towards the bottom, where $b < 10$, there are many points with large deviations in the $\log N$ direction. Points in this region have been fitted with $\log N$ both greater than and less than the true value, but most have $b_{\rm fit} \sim b_{\rm true}$.
- To the right, where $\log N > 13.4$, most of the points are well determined and have small errors in both b and $\log N$.
- To the left, where $\log N < 13.4$, there is a distinct tendency for the points to have migrated considerable distances along a line of positive slope (mostly between about 45° and vertical in Figure 3.8). These large b errors are mostly in the sense of $b_{\rm fit} < b_{\rm true}$, but there are several in the reverse sense.

These results show that broad and weak lines tend to have the greatest fitting errors in b, a fact reflected also in Figure 3.3 and Figure 3.6, whereas narrow lines have the greatest $\log N$ errors.

Broad, weak lines are the most susceptible to having their overall shapes altered by noise, which then gives rise to different apparent b values. Since such lines are on the linear part of the curve-of-growth, where the curves for high b values are very close together, these apparent changes in b do not significantly affect the apparent $\log N$ value.

Narrow lines cover a small number of pixels in the spectrum, so are more likely than broader lines to be subject to changes in apparent equivalent width due to noise. This means the apparent $\log N$ value can be significantly different from $\log N_{\rm true}$, but the apparent breadth of the line, and hence b value, will not change much.

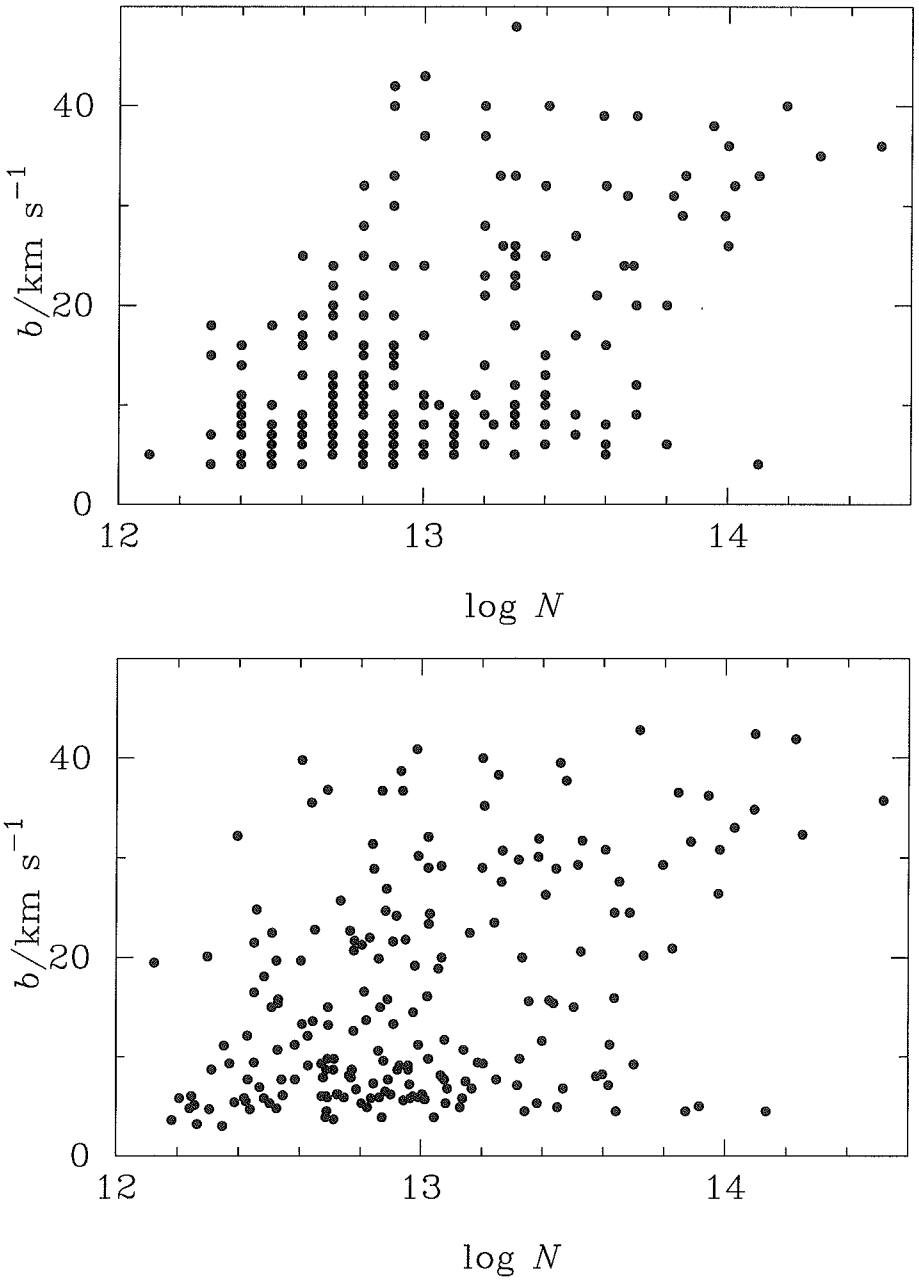

Figure 3.7 Comparison between the fitted line parameters and the corresponding true parameters in the b–log N plane for fitted lines with one corresponding true line in the CNQ. The upper panel shows the fitted values. The lower panel shows the true values for the same lines. The obvious structure in the upper panel is due to the rounding of measured b values to whole numbers and log N values to one decimal place.

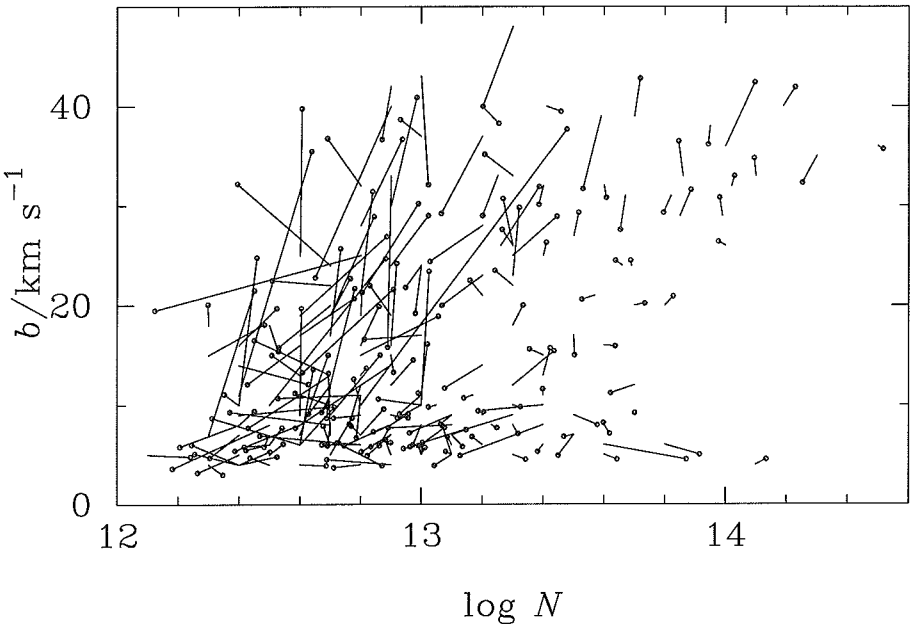

Figure 3.8 The migration of line parameters in the b–log N plane due to the fitting procedure for the CNQ. The small circles show the true parameters of the absorption lines, the attached lines extend to the position of the fitted parameters.

An important point to note is that spectra of the resolution of the Cloudy Night spectrum (which includes the real QSO echelle spectra studied in this thesis) highly resolve the absorption profiles. With no noise, a profile sampled at the same resolution would appear quite smooth, with many pixels defining its shape. Noise adds high frequency components to the Fourier transform, and can give the line a "spiky" appearance and change its apparent b value. In some instances the apparent b value may be almost entirely accounted for by the instrumental resolution, making the deconvolved line seem exceptionally narrow.

The effect of these migrations in the b–log N plane on any apparent correlation between the two parameters is investigated in Section 3.7.

Another instructive method of visualising the behaviour of line parameters under the fitting process is to plot the fitting errors in b against those in log N. Such a plot is shown in Figure 3.9.

The appearance of this Figure is striking in that the first and third quadrants are heavily populated while the second and fourth are largely empty. This demonstrates the strong tendency for lines to be fitted with b and log N either both greater than their true values or both smaller than their true values. This implies that lines which appear to be narrower than they truly are also appear to be weaker and, conversely, lines which appear broadened also appear to be stronger than they are in reality.

This behaviour is a natural consequence of the effects of random noise on the shapes of the line profiles. Lines which have their apparent equivalent widths lowered by noise will tend to be both narrower and less deep than in reality, while lines with

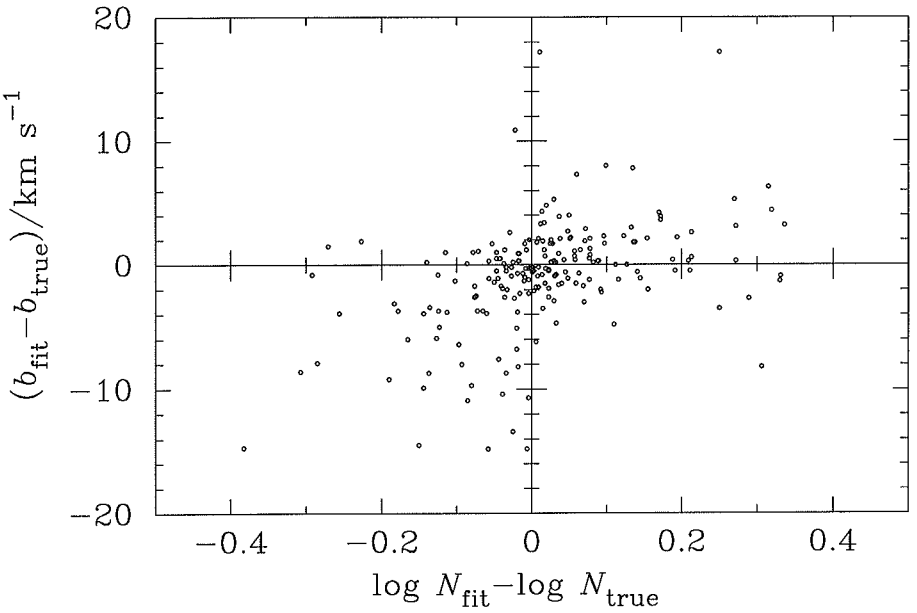

Figure 3.9 The fitting errors in b plotted against the fitting errors in $\log N$ for the Cloudy Night QSO. The axes have been chosen to best illustrate the positions of the many points near the origin, resulting in three points falling outside the region shown. These points are at coordinates (0.67, 5.5), (−0.24, −24.5), and (−0.68, −27.7).

increased apparent equivalent width will tend to be both broader and deeper. A line which, because of noise, appears broader and shallower, or narrower and deeper, than it really is will probably be fitted with an inaccurate b value. If such a line is unsaturated and the equivalent width has not changed substantially, it will be assigned the correct $\log N$ value. In none of these cases is a substantial journey into the second or fourth quadrant of Figure 3.9 produced.

Similar effects occur when the continuum level is incorrect. If the continuum is too low, the broadest part of a line no longer appears to be absorption, so the line appears both shallower and narrower. If the continuum is too high (which is rarely the case), lines will appear deeper and broader.

Rauch *et al.* (1993) have arrived at some similar conclusions on the effects of noise on profile fitting, pointing out that noise can artificially lower the apparent b values of lines. The findings presented here confirm this, but show that many other effects can occur as well.

3.3.4 Conclusions on the Fitting Procedure

The comparison of the results from my line-fitting of the CNQ data to the reality revealed by Ed Jenkins demonstrated the following points:

- The wavelength measurements of absorption lines were in very good agreement with reality.

Table 3.4 Comparison of $\tilde{\chi}^2$ values for selected lines as fitted and with the true line parameters. $\mathcal{N}_{\rm fit}$ and $\mathcal{N}_{\rm true}$ are the number of components fitted and actually present in the complex containing the line, respectively.

Line	$\lambda_{\rm fit}$	$\mathcal{N}_{\rm fit}$	$\tilde{\chi}^2{}_{\rm fit}$	$\mathcal{N}_{\rm true}$	$\tilde{\chi}^2{}_{\rm true}$
15	3410.13	1	1.59	1	2.70
42	3473.66	1	0.49	1	0.90
46	3490.58	1	1.34	1	1.46
151	3695.19	3	1.42	4	0.80
237	3817.37	2	0.86	2	0.98

- The column density measurements did not show any systematic bias or correlation of measurement error with the value of $\log N$, at the 95% confidence limit.
- The velocity dispersion measurements showed both a bias toward measuring systematically lower values of b, and a significant correlation of the measurement error with the actual value of b. All attempts to construct a suitable calibration to correct for this error gave unsatisfactory results.

Although these results do not allow the construction of a method to eliminate all (or any) bias from the fitting procedure, they do give an important insight into what sort of biases exist. With this knowledge, it will be possible to apply an appropriate degree of caution to any conclusions drawn from the fitting of lines in real QSO spectra.

3.4 Detailed Examination of Some Line Fits

In order to appreciate the difficulties associated with fitting theoretical profiles to spectral lines, several individual lines were selected for a detailed comparison between the fitted line parameters and the true ones. Most interesting are the cases in which the fitting process failed to produce anything near a reasonable result.

Listed below are some of these cases, labelled by the line numbers used in Table B.3 and Table B.4. The profile fits are shown in Figure 3.10, and some comparisons between the $\tilde{\chi}^2$ values for my fits and the true line parameters are shown in Table 3.4.

Line 15: This is a weak, noisy line from the bluemost spectral order. It was fitted with a grossly high value of $\log N_{\rm fit} = 12.8 \pm 0.1$ when the true value was 12.12, which is below the detection limit. The true continuum level is almost identical to the fitted continuum near this line, so random noise is responsible for making this line appear much stronger than it is. This line is the obvious outlier point in the top left corner of Figure 3.4.

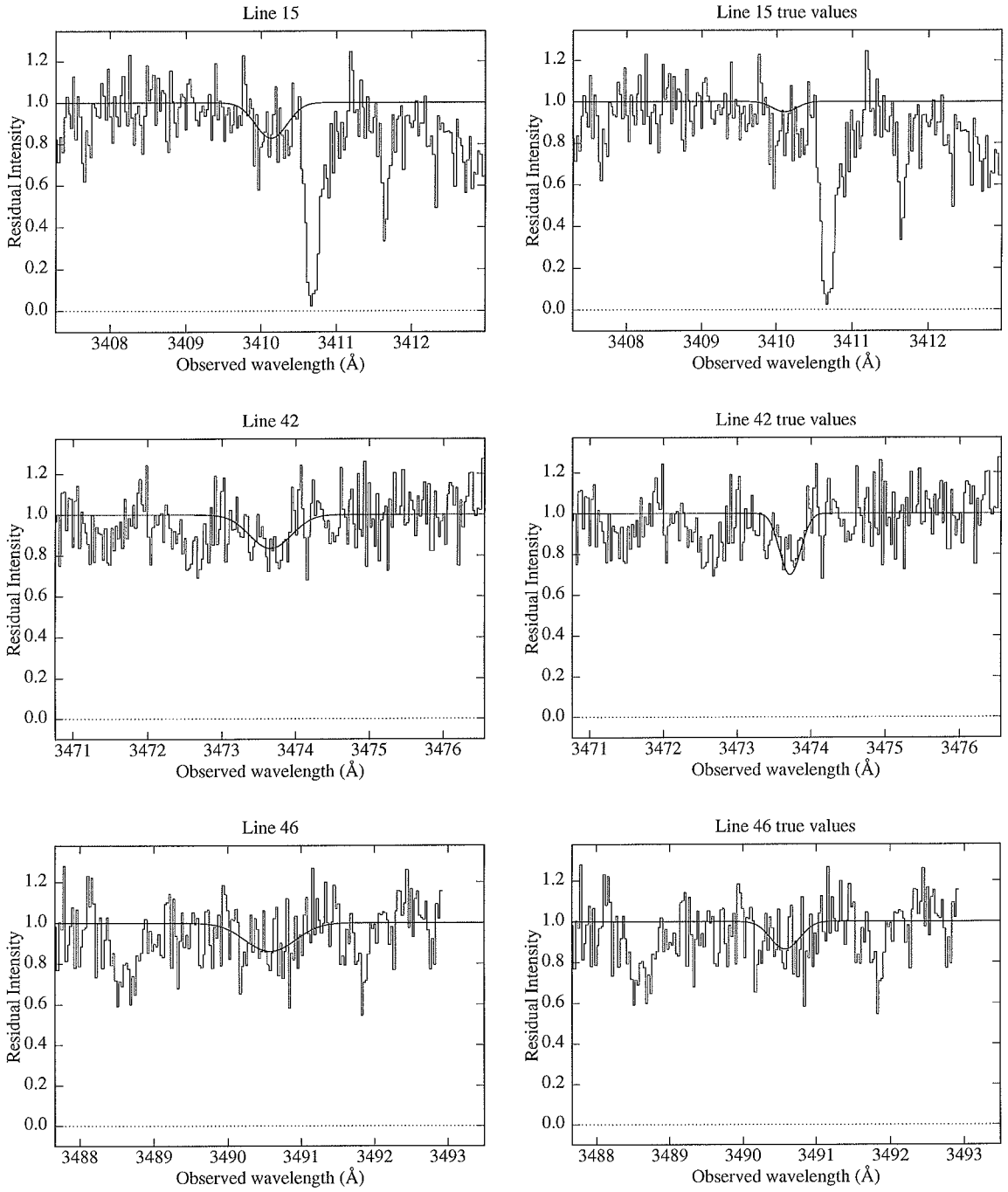

Figure 3.10 Profile fits to some selected lines from the CNQ spectrum. On the left are the fits adopted, on the right are profiles generated by the true line parameters. Note that the true line parameters are shown with the fitted continuum, which is not necessarily the same as the true continuum.

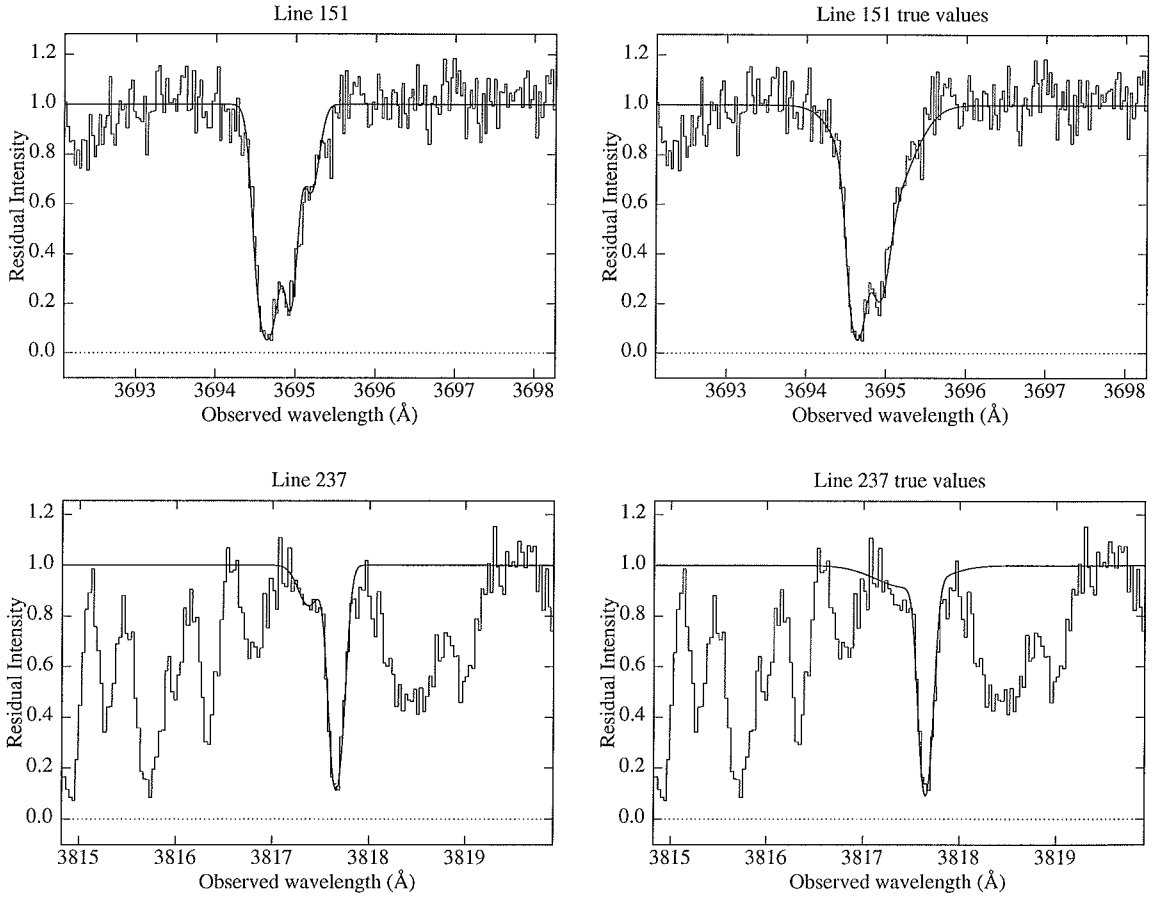

Figure 3.10 *Continued.*

Line 42: This weak line was fitted with $b = 33 \pm 5\,\mathrm{km\,s^{-1}}$ while $b_{\mathrm{true}} = 15.8\,\mathrm{km\,s^{-1}}$. As seen in Figure 3.10, the noise has distorted the apparent line shape. The difference in $\tilde{\chi}^2$ between the adopted fit and the true line profile, shown in Table 3.4, shows that the noise has significantly altered the true line shape. Another factor is that the fitted continuum is $\sim 10\%$ too low, making the line appear shallower than it is in reality. This line accounts for the left-most of the two outlier points in the top half of Figure 3.3.

Line 46: Another weak line, this is in a very noisy part of the spectrum near an order edge. The values $b_{\mathrm{fit}} = 40 \pm 5\,\mathrm{km\,s^{-1}}$ and $\log N_{\mathrm{fit}} = 12.9 \pm 0.1$ are both significantly high. The continuum here was set $\sim 5\%$ too low, so it cannot account for the apparent increase in equivalent width. The increase has been caused purely by the noise. This point is the right-most of the two outliers in the top half of Figure 3.3.

Line 151: This line forms part of a four line complex which was fitted with only three components. Lines 149 and 150 each contain significant equivalent width from at least two of the components, but the feature associated with Line 151 is attributable only to the very broad ($b_{\mathrm{true}} = 37.7\,\mathrm{km\,s^{-1}}$) component and so

appears in the analyses of Section 3.3. The narrow ($b_{\rm fit} = 10 \pm 2\,{\rm km\,s^{-1}}$) line used to fit this feature is therefore much weaker than the true line, and this produces the bottom-most outlier point in both Figure 3.3 and Figure 3.4. The fitted continuum here is $\sim$ 7% too low, which gives the fitted narrow component a smaller equivalent width than the true feature.

Line 237: This is another weak line, close to a strong, narrow line. The noise has made this line appear deeper and narrower than it is, despite the continuum being fitted $\sim$ 7% too low. The adopted fit with $b_{\rm fit} = 11 \pm 2\,{\rm km\,s^{-1}}$ gives a smaller $\tilde{\chi}^2$ value than the true value $b_{\rm true} = 35.5\,{\rm km\,s^{-1}}$. This line produces the second lowest outlier point in Figure 3.3.

With the exception of the complex containing line 151, in which fewer components were fitted than were actually present, the value of $\tilde{\chi}^2{}_{\rm fit} < \tilde{\chi}^2{}_{\rm true}$ for all of these lines. This indicates that the adopted fits are reasonable, given the appearance of the noisy data, and that no obvious errors have been made in the fitting process.

3.5 Analysis of the Lyman α Column Density Distribution

Jenkins revealed that the Lyman α lines in the CNQ spectrum were distributed randomly according to the density distribution function

$$\frac{\partial^2 \mathcal{N}}{\partial N({\rm H\,I})\partial z} = (A_1 N({\rm H\,I})^{-1.3} + A_2 N({\rm H\,I})^{-2.0})(1+z)^{2.33}, \qquad (3.6)$$

with $A_1 = 3.156 \times 10^4$ and $A_2 = 8.436 \times 10^{13}$. This function was truncated outside the range $1 \times 10^{12} < N({\rm H\,I}) < 4 \times 10^{14}\,{\rm cm^{-2}}$, the lower limit judged by Jenkins to be well below the detection threshold and the upper limit chosen to reflect the lack of observed lines above that limit in published spectra of Q1101−264 (Carswell *et al.*, 1991) and Q2206−199N (PHSM).

A plot of Equation 3.6 superposed on the measured Lyman α $\log N$ distribution histogram is shown in Figure 3.11. The two curved lines in the figure show the maximum likelihood fit to Equation 2.6 for $\log N > 12.75$ and the true distribution function used by Jenkins (Equation 3.6), normalised to the number of lines observed with $\log N > 12.75$. The two curves are close over most of the relevant $\log N$ range, diverging slightly only at high $\log N$ values, where there are few lines. Given that only a single power law was fitted, for reasons discussed in Section 2.3.2, the recovery of the shape of the distribution is remarkably good.

The good recovery of the distribution shape is encouraging in that it demonstrates that the problem of incompleteness can be overcome to determine the shape of the underlying distribution. It also shows, however, that it is difficult to distinguish between various functional forms for the distribution. The limited number of lines in a single spectrum, combined with fitting errors, precludes anything more

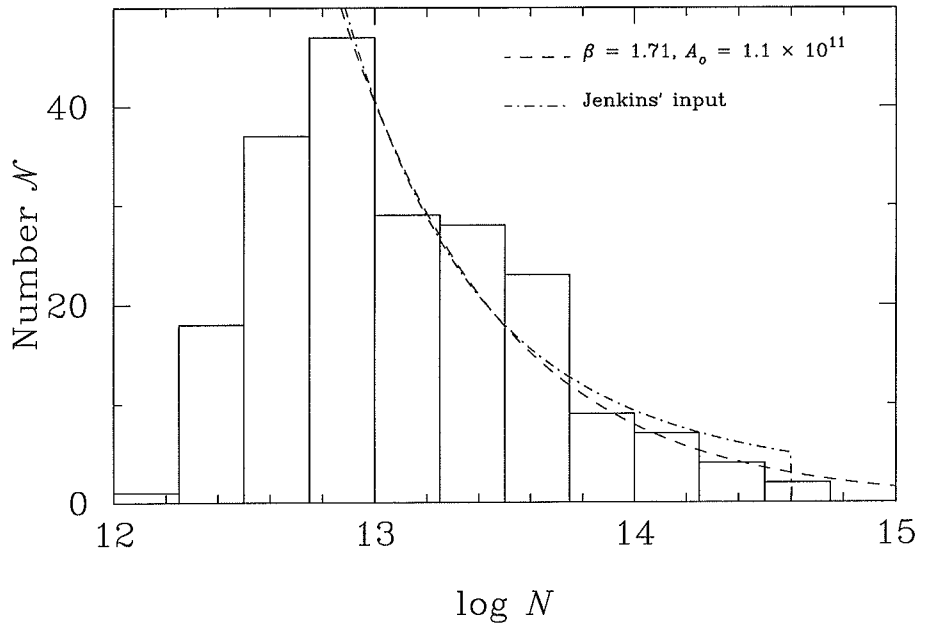

Figure 3.11 Number density histogram of the measured $\log N$ values for the CNQ Lyman α lines, with the dot-dashed line showing the line density distribution function used by Jenkins to generate the spectrum.

complex than a simple low-order fit. If the Lyman α forest arises in several populations of objects, each with their own distributions, this could not be determined from a single QSO observation. The combined data from many objects, including good data at high and low extremes of column density, would be needed to establish a distribution more complex than a single power law. Collecting such data is difficult because there are very few Lyman α forest lines with $\log N \gtrsim 14$ and lines with $\log N \lesssim 12.5$ are subject to severe incompleteness effects.

3.6 Analysis of the Metal Identifications

Of the 362 absorption lines identified and measured in the CNQ spectrum, all 362 could be associated with real lines present in the simulation (*i.e.* no lines were generated spuriously by noise). Many of the identified lines were blends of two or more significant components, sometimes produced by different species. Full details of the line identifications are shown in Table B.4.

The lines were assessed as either correctly or incorrectly identified, and were put into the following categories, summarised in Table 3.5:

- Lines correctly identified, with every significant component in the simulation identical to the adopted identification: 242 lines, plus 1 extra line which was correctly identified as a blend of two different metal lines.
- Lines correctly identified, but blended with additional lines which were not

Table 3.5 Summary of the correctness of the metal line identifications in the CNQ. The percentages are rounded to the nearest 0.1%.

Description	No.	%
1 species correctly identified	242	66.8
2 species correctly identified	1	0.3
Total correctly identified	243	67.1
Lyman α correct, metal unidentified	22	6.1
metal correct, Lyman α unidentified	23	6.4
metal correct, metal unidentified	7	1.9
metal correct, C I* unidentified	1	0.3
Lyman α correct, H_2 unidentified	1	0.3
Total correct, with blend	54	14.9
Total correct	**297**	**82.0**
Lyman α identified as metal	8	2.2
H_2 identified as Lyman α	6	1.7
H_2 identified as metal	4	1.1
C I* identified as Lyman α	4	1.1
metal (known system) identified as metal	2	0.6
metal (unknown system) identified as metal	4	1.1
metal (known system) identified as Lyman α	25	6.9
metal (unknown system) identified as Lyman α	12	3.3
Total incorrect	**65**	**18.0**
Total, all lines	**362**	**100.0**

recognised: 22 lines identified as Lyman α also contained metal lines (necessarily at different redshifts), 23 lines identified as metals also contained absorption due to Lyman α, 7 lines identified as metals also contained other unidentified metal lines, 1 metal line also contained a C I* fine structure line, 1 Lyman α line also contained a molecular hydrogen line. This accounted for a total of 54 lines.

- Lyman α lines incorrectly identified as metals: 8 lines.
- Molecular hydrogen lines in the $z_{\rm abs} = 2.16$ damped Lyman α complex. H_2 had not been searched for because molecular lines were not expected to be present: 6 of these lines were incorrectly identified as Lyman α and 4 were incorrectly identified as metals. Of particular note, the outlying point at $\log N = 14.1$ and $b = 4$ in Figure 2.2 was an unidentified H_2 line.
- C I* and C I** fine structure lines in the $z_{\rm abs} = 1.05$ complex (identified by several other lines), which had not been searched for because of an oversight in the search lists: 4 of these lines were incorrectly identified as Lyman α.
- Metal lines identified as the wrong metal species at the wrong redshift: 2 lines from otherwise identified redshift systems and 4 lines from unidentified

systems were incorrectly identified in this way.

- Metal lines with no identification (thus assumed to be Lyman α): 25 lines from otherwise identified redshift systems and 12 lines from unidentified systems were incorrectly identified in this way.

It is worth noting that the CNQ spectrum contained an unusually high number of heavy element absorption systems (and in fact absorption lines in general). Twenty distinct absorption systems were identified, and 157 of the 362 measured absorption lines ($\sim$ 43 %) were identified with metals. In contrast, the real QSOs studied in this thesis, Q1101−264 and Q2348−147, which span a similar redshift range to the simulated redshift of the CNQ, had 8 and 3 heavy element absorption systems, comprising 97/160 ($\sim$ 61%) and 59/384 ($\sim$ 15%) of the measured lines respectively.

The mean number of identified metal lines per absorption system is lower in the CNQ data (at 7.8) than in either the Q1101−264 or Q2348−147 data (12.1 and 19.7 respectively). The high mean values for the real QSO spectra reflect the multiplicity of velocity components in many of the heavy element systems—which is another factor tending to make the systems more easily identifiable. With fewer lines—and velocity components—per absorption system able to be identified, the CNQ spectrum presented a particularly difficult case for the discrimination of metal and Lyman α lines.

3.6.1 Metal Line Contamination

One of the most important considerations for studies of the Lyman α forest is the amount of "contamination" of the Lyman α sample by unidentified metal lines. A total of 205 lines were identified as Lyman α and 47 of these were misidentifications. However, ten of these misidentifications were H_2, C I* or C I** lines, which were not included in the heavy element search list. Had these lines been in the search list, it is likely they would have been identified, since all were in redshift systems identified by several other lines. In a search including all likely lines, the rate of contamination therefore appears to be $\sim$ 20% of the identified Lyman α lines. This is a considerable fraction, certainly higher than the "negligible" metal contamination which is often claimed in published work.

The remaining 37 misidentified metal lines are examined in more detail in the Sections below.

Unidentified Lines in Known Systems

The 25 unidentified metal lines which were members of known heavy element absorption systems consisted of:

- One Ca II λ3934 line in the $z_{\rm abs} = 0.000$ complex was unidentified. The corresponding Ca II λ3969 line, at $z_{\rm abs} = 0.00003$, was identified, but it was weak and the corresponding line at the expected position of Ca II λ3934 was much

stronger. Since the λ3969 line has a higher oscillator strength, it was concluded that the line in the position of λ3934 must have consisted mostly of Lyman α absorption.

- Eight lines in the $z_{\rm abs} = 1.052$ complex. These included C I λ1656, Ni II $\lambda\lambda$1709, 1741, and Si II λ1808 in an unidentified component at $z_{\rm abs} = 1.05077$, an Al II λ1670 line in the unidentified redshift component $z_{\rm abs} = 1.05135$, Al I λ1765 in the identified component at $z_{\rm abs} = 1.05190$, and Ni IIλ1754 and Al I λ1765 in the identified component at $z_{\rm abs} = 1.05209$. On reflection, the $z_{\rm abs} = 1.05077$ component should have been found, since most of the lines were of reasonable strength and unblended. However, this component was displaced by $\sim 30\,{\rm km\,s^{-1}}$, or ~ 3.5 Å, from its closest neighbour in the complex and there are other lines present in the gaps, breaking up the coherency of the pattern between the transitions. In this case it was simply that more care was required in the identifications than was applied. The other unidentified lines were generally weak and noisy or partially blended, making identifications difficult.
- A C IV $\lambda\lambda$1548, 1550 doublet at a new $z_{\rm abs} = 1.30379$ component of the $z_{\rm abs} = 1.302$ complex. The λ1548 component is heavily blended with a Lyman α line, and no other ions are visible in this component, so identification would have been extremely difficult based on the lone λ1550 line.
- Four C IV λ1548 lines in the complex at $z_{\rm abs} = 1.523$. Another λ1548 line at $z_{\rm abs} = 1.52335$ is noted as a tentative identification in Table B.8 and Table B.9 because the corresponding λ1550 line was blended in a strong feature and could not be identified clearly. The four extra components are all weaker, and their corresponding λ1550 lines fall in an inter-order gap, making the identifications impossible to confirm with the available data. The strongest unidentified component also contained an unidentified Si IV λ1393 line.
- Three lines in an unidentified component of the $z_{\rm abs} = 1.790$ complex. These were C II λ1334 and Si IV $\lambda\lambda$1393, 1402 at $z_{\rm abs} = 1.78850$. These three lines are not blended but, as in the $z_{\rm abs} = 1.05077$ component of the $z_{\rm abs} = 1.052$ complex, they were separated from the main complex by a considerable velocity difference ($\Delta v \sim 20\,{\rm km\,s^{-1}}$) and so overlooked. A fourth unidentified line in this complex was Si IV λ1402 in the known $z_{\rm abs} = 1.78958$ component. This line was heavily blended with other lines.
- Two Si IV λ1402 lines in the $z_{\rm abs} = 1.864$ complex. The Si IV λ1393 lines corresponding to these lines are in an inter-order gap. The redshifts of the components in this system were set by blended and uncertain Si III λ1206 lines, so confirming the identifications of the Si IV lines was difficult.
- A single Si II λ1193 line in the $z_{\rm abs} = 2.160$ complex. This was a weak line at $z_{\rm abs} = 2.16018$, a redshift not seen in any other line in the complex because other lines at this redshift were either blended or too weak to be detected.

- One N V λ1238 line in the $z_{\rm abs} = 2.202$ complex. This line was initially identified correctly, but ultimately rejected as N V because of a poor wavelength match with other lines in the system. The poor match was caused by the blending of two N V lines, which made confident identification of this line difficult.
- A single Si III λ1206 line in the $z_{\rm abs} = 2.229$ complex. This line was heavily blended with Lyman α features. No other lines were visible in this component, including Lyman α (in an inter-order gap), so there was no way of confirming the identification.

Overall, seven of the 25 unidentified lines in identified systems could have been identified with a little more care in the analysis. The other 18 lines were difficult to identify because of either blending or a lack of any confirming lines at the same redshift.

Unidentified Lines in Unidentified Systems

The 12 unidentified metal lines which were members of unidentified heavy element absorption systems consisted of:

- A Mg II λ2796 line at $z_{\rm abs} = 0.41139$. This line was blended with a Lyman α line, as was the corresponding Mg II λ2803 line. No other lines in the redshift system are visible in the spectrum. Identifying the system from this blended doublet would have been extremely difficult.
- Four C IV $\lambda\lambda$1548, 1550 doublets at $z_{\rm abs} = 1.34984$, 1.38641, 1.39625, and 1.42553. The first system is unblended and should have been recognised. The other systems each have one member of the doublet strongly blended with other features, making them more difficult to recognise. A second unidentified λ1548 component was present in the $z_{\rm abs} = 1.425$ system.
- A single C IV λ1550 line at $z_{\rm abs} = 1.56769$. The corresponding λ1548 component was misidentified as Si II λ1526 in the $z_{\rm abs} = 1.604$ system, but the possible C IV doublet should still have been recognised.
- A C II λ1334 line at $z_{\rm abs} = 1.72775$. No other lines in this system are easily identifiable. The Si IV $\lambda\lambda$1393, 1402 lines are blended in the Mg II λ2796 complex at $z_{\rm abs} = 0.364$ and the damped Lyman α line at $z_{\rm abs} = 2.160$ respectively. Si II λ1304 is blended in the C IV λ1550 complex at $z_{\rm abs} = 1.302$. O I λ1302 is blended in the C IV λ1548 complex at $z_{\rm abs} = 1.302$. Si II λ1260 is too weak to be detected. No other lines expected to be seen fall within the wavelength coverage of the spectrum.

A total of three of the 12 lines in unidentified systems could reasonably be expected to have been identified in an even more exhaustive search, while the remaining 9 were difficult to identify.

3.6.2 Discussion of Metal Line Identifications

It is a reasonable assumption that most of the lines of H_2, C I*, and C I** in the CNQ spectrum would have been identified if those lines had been in the search list. If this is the case, then the rate of contamination of the Lyman α sample by unidentified metals is $\sim$ 20%. In a totally exhaustive search, perhaps ten more metal lines could have been identified, leaving the claimed Lyman α sample with $\sim$ 15% metal lines which cannot be identified because there is no more than one unblended line available in each unidentified redshift system.

It is important to recall that the CNQ spectrum seems to contain an unusually large number of heavy element absorption systems, with few identifiable metal lines per system. The high density of metal lines in the spectrum makes it difficult to identify redshift systems because of the high likelihood of line blending. In a real QSO spectrum, containing fewer metal systems, with more identifiable lines per system, one would expect to have a higher success rate at identifying the metal systems and their lines, so the contamination rate of 15% is most likely an upper limit, provided adequate care is taken in the search for heavy element systems.

The study of the spectrum of Q2206−199N by Pettini *et al.* (1990) (PHSM) focused attention on the possible existence of Lyman α lines with very low velocity dispersions, including some lines with $b < 10\,\mathrm{km\,s^{-1}}$. Such lines pose serious problems for theoretical models of the Lyman α clouds, as discussed in detail in Chapter 6. However, in the extreme limit that 15% of the narrowest lines identified as Lyman α are actually unidentified metal lines, then most of the lines in PHSM's study with $b < 10\,\mathrm{km\,s^{-1}}$ could in principle be produced by metals[3]. This rate of contamination does not, however, account for the large number of lines seen with $10 < b < 20\,\mathrm{km\,s^{-1}}$.

Although it is possible for minor improvements to be made in the rate of correct metal identifications over what was achieved here with the CNQ spectrum, it should be borne in mind that a Lyman α contamination rate of up to 15% unidentified metal lines may be present in an analysis of the Lyman α forest at S/N ratios of $\sim$ 8–12. At significantly higher S/N ratios, a better success rate may be expected, because blended lines will be easier to deconvolve and their wavelengths easier to measure accurately.

3.7 Analysis of the b–N Correlation

In order to see how significantly the fitting process alters (or indeed induces) the apparent b–N correlation, all of the true lines associated with the detected lines shown in Figure 2.2 were analysed. Since some lines identified as single lines are in fact blends, this sample contains 325 lines, significantly more than the 205 lines identified as Lyman α. Also, several lines in this sample were either metals misiden-

[3]In fact, Rauch *et al.* (1993) managed to identify some of the low b lines from PHSM with metals.

tified as Lyman α lines or metals blended with Lyman α lines. In order to compare these with the line parameters of the assumed Lyman α identifications, the $\log N$ values for these metal lines in the sample were adjusted according to Equation 2.3. These metal lines were left in the analysis initially, so the overall behaviour of any correlation under the effect of the fitting errors could be investigated, regardless of the identifications. This sample of lines will be referred to as the "True" sample in the Sections below.

3.7.1 Production of the Lyman α Cloud Parameters

After first assigning $\log N$ values for Lyman α lines in the CNQ spectrum according to the distribution function given in Equation 3.6, Jenkins produced b parameters by applying the formulae:

$$b_0 = A_0 + B_0 \log N; \tag{3.7}$$

$$b = b_0 + g(0.4\, b_0), \tag{3.8}$$

where $A_0 = -49.176$, $B_0 = 5.765$, and $g(0.4\, b_0)$ is a gaussian random variable with standard deviation $\sigma = 0.4\, b_0$.

Jenkins deliberately chose this method of generating b values so that lines with high $\log N$ and low b would not readily be produced, in accordance with the non-detection of such lines in observations of real QSOs (Pettini *et al.*, 1990; Carswell *et al.*, 1991, for example). The dispersion, however, was chosen to be large enough to populate the low $\log N$, high b region of the parameter space—where lines probably exist in real QSO spectra but are difficult to detect because of the noise.

The CNQ spectrum contained 727 simulated Lyman α lines, most of which were not detected because of line-blending or S/N effects. The correlation coefficient of b and $\log N$ for these 727 lines was calculated to be $r = 0.316$. A formal calculation showed that a correlation coefficient as high as this had a probability of only 7×10^{-19} of occurring in a sample of 727 uncorrelated data points, indicating significant correlation between the parameters, as expected.

3.7.2 Comparison of Measured Correlation with True Correlation

A comparison plot between the fitted b and $\log N$ values and the corresponding true values is shown in Figure 3.12. The distortion of the b–$\log N$ distribution can be seen in this Figure as an apparent shift in the points to a more concentrated range of $\log N$ values and a slightly increased range of b values. The apparent loose correlation of b with $\log N$ remains, although its slope is increased significantly. Correlation analyses of the 325 lines corresponding to Lyman α identifications (the "True" sample) were done, fitting to Equations 2.4 and 2.5. The results are shown in Table 3.6.

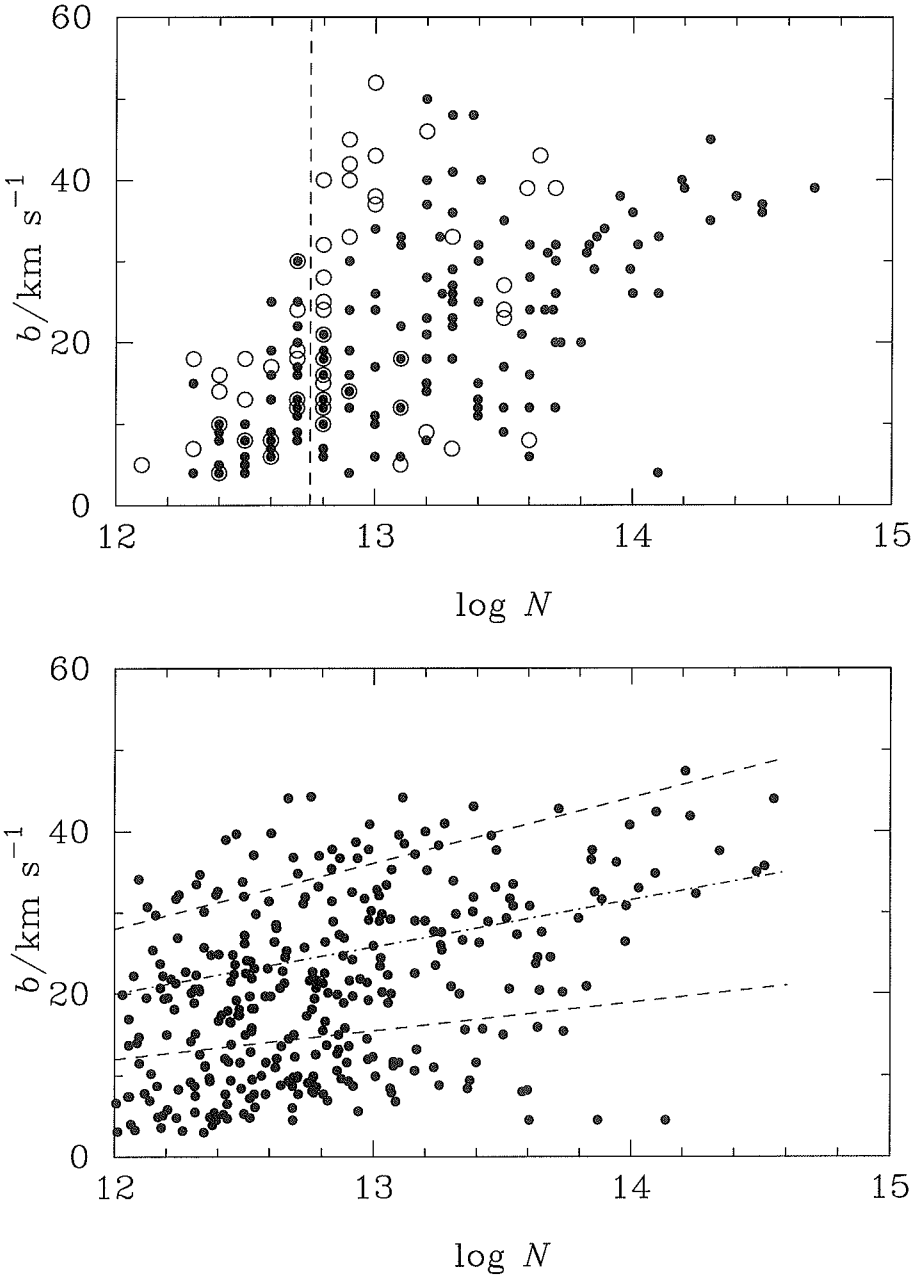

Figure 3.12 Comparison between the fitted line parameters and the corresponding true parameters in the b–$\log N$ plane for lines identified as Lyman α in the CNQ. The upper panel shows the fitted values, with open circles indicating fits marked as uncertain in Table B.3. The lower panel shows the true values for lines associated with the fitted lines. The dot-dash line in the lower panel shows the mean b value of Lyman α lines as a function of $\log N$ in the simulation and the dashed lines show 1σ deviations. The lines end at $N = 4 \times 10^{14}\,\mathrm{cm}^{-2}$, where the Lyman α distribution is truncated. The excess of points below the lower dashed line is caused by metal lines misidentified as Lyman α.

Table 3.6 Results of regression analyses between the measured and true b and $\log N$ values for the Lyman α lines in the CNQ spectrum.

$b = A + B \log N$				
Sample	No. lines	A	B	r
1[1]	205	-207 ± 7	17.2 ± 0.6	0.52 ± 0.06
2[2]	146	-216 ± 9	17.8 ± 0.7	0.62 ± 0.06
True[3]	325	12.4[5]	0.0198[5]	0.42 ± 0.07
Input[4]	727	-49.2[6]	5.76[6]	0.31 ± 0.08

$N = A' b^{B'}$				
Sample	No. lines	A'	B'	r
1[1]	205	$(5 \pm 2) \times 10^8$	3.34 ± 0.15	0.51 ± 0.07
2[2]	146	$(8 \pm 5) \times 10^8$	3.23 ± 0.17	0.59 ± 0.06
True[3]	325	1.25×10^{-5}	5.60	0.38 ± 0.07
Input[4]	727	—	—	0.26 ± 0.07

Notes:
[1] Sample 1 is identical to Sample 1 in Table 2.2, namely all the lines identified as Lyman α in the CNQ spectrum.
[2] Sample 2 is identical to Sample 2 in Table 2.2, namely the lines identified as Lyman α in the CNQ spectrum and with well-determined b and $\log N$ parameters.
[3] The "True" sample is as defined in Section 3.7, namely all the true lines in the CNQ spectrum associated with lines identified as Lyman α.
[4] The "Input" sample contains all the Lyman α lines actually in the CNQ spectrum.
[5] The A and B values for the "True" sample have no uncertainties quoted because the true values contain no inherent uncertainty.
[6] The A and B values shown for the "Input" sample are not fits to the 727 input lines, but the actual values from the relation used by Jenkins in generating the Lyman α cloud parameters before applying a gaussian dispersion to the b values. This equation is plotted as the dot-dash line in Figure 3.12.

Another striking feature of Figure 3.12 is the existence of a large group of points in the lower panel well below the mean b value of the input Lyman α lines. Most of these points are due to unrecognised metal lines, either fitted individually or blended with other lines. The effect of these lines is to produce an apparent population of narrow Lyman α lines and increase the apparent slope of the fitted correlation, although their effect on the sample of 325 true line parameters is to increase the apparent dispersion in b of the Lyman α lines, resulting in a flatter slope. These changes in apparent slope are clear from the numbers shown in Table 3.6.

Bootstrap Calculation of Correlation Coefficients

A bootstrap resampling method (Efron and Tibshirani, 1986, and references therein) was used to calculate the correlation coefficients and their uncertainties for each sample. This was done in order to be able to compare directly the strengths of the correlations between the various samples, without being concerned with the number of lines in each sample (since they are all different).

The bootstrap method involves forming "resamplings" of a data set by choosing members of the set at random, with replacement (*i.e.* each data point may be chosen multiple times). The parameter of interest is calculated for each of many different resamplings, and its distribution over the ensemble of resamplings provides an estimate of the parameter and its variance. The advantages of using this method to calculate correlation coefficients is that it provides 1σ uncertainty estimates for r, and that the number of data points in each resampling can be identical, regardless of the number of points in the samples being analysed.

In this case, four samples, all of different sizes, need to have the correlations between their parameters quantified so they can be compared. To do this, r was calculated for each sample by taking 1000 bootstrap resamplings, with each resampling equal in size to the number of data in the smallest sample, namely 146 points for Sample 2. A small sample size rather than a large one was chosen to avoid any possible effects from gross resampling of the same data points in any one sample. Using the same sample size eliminated any effects on r that might stem from different sample sizes—so the resulting values of r become directly comparable measures of the strength of the correlation of the data.

Run several times with different pseudo-random number generator seeds, the mean correlation coefficient for each sample was reproducible to < 0.01, much smaller than the estimated 1σ uncertainties. The results of the bootstrap calculations for r are shown in Table 3.6. It should be noted that formal calculations give exceedingly low probabilities ($\ll 10^{-10}$) for any of the measured correlation coefficients to have arisen in uncorrelated data of the same number of points. The samples are certainly all correlated; it is a question of comparing the strengths of the correlations.

The method of bootstrap resampling is used again in a different context in Section 7.2.2.

Discussion of Correlation Coefficients

The correlation coefficients listed in Table 3.6 show that the measured line parameters imply a stronger correlation than actually exists. In the linear case (Equation 2.4), the value $r_1 = 1.25\, r_{\rm True}$, and in the power law case (Equation 2.5) $r_1 = 1.36\, r_{\rm True}$. This demonstrates clearly that the fitting procedure can strengthen the appearance of a b–log N correlation. Even if there was no correlation intrinsic to the data, the migration of points in the b–log N plane, as discussed in Section 3.3.3, would produce an apparent correlation of significant strength.

Also, the value $r_{\rm True} = 1.33\, r_{\rm Input}$ in the linear case and $r_{\rm True} = 1.42\, r_{\rm Input}$ in the power law case. This shows that, disregarding the fitting procedure, the selection effects produce a stronger b–log N correlation than is actually present. The selection effects alone produce an increase in the correlation coefficient of a similar magnitude to the increase caused by the fitting procedures. Even without the metal lines spuriously identified as Lyman α lines, the depopulation of the upper left region of Figure 3.12 causes a highly significant increase in the measured slope, as well as an increase in the measured correlation coefficient.

At the S/N ratios being studied (7–20), it therefore appears that the two independent processes of line selection and line fitting both contribute approximately equally to the apparent increase in significance of any correlation between b and $\log N$. It is also interesting to note that the correlation coefficients for Sample 2 are higher than those for Sample 1. Since Sample 2 was defined by removing lines with ill-determined parameters from Sample 1, this shows clearly that selecting only well-determined lines also tends to induce a correlation between b and $\log N$.

This result, combined with the work by Rauch *et al.* (1993), throws some doubt on the strength of the b–log N correlation reported by PHSM, although it does not necessarily invalidate it completely. Since relatively few points were presented there, and some points were deliberately not included in the analysis (those corresponding to saturated lines and complex blends), it is difficult to estimate by how much the reported value of $r = 0.77$ may have been overestimated. If the correlation is much weaker than this, it would also weaken their argument for most of the velocity dispersion seen in high-b clouds being due to bulk motion.

The results of this Section are discussed further in Section 6.4.3, where the correlation of b and $\log N$ in real data is addressed.

3.8 Summary

Thorough comparisons were made between various measurements taken from the Cloudy Night QSO spectrum and their true values, as revealed by Professor Edward Jenkins.

It was found that the continuum levels fitted to the CNQ data were almost invariably too low, usually by 5–10%. Low S/N ratios and the presence of absorption features were seen to have a dramatic effect on the accuracy of the continuum fitting.

The accuracy of the Voigt profile fitting to the absorption lines in the CNQ spectrum was analysed. The quoted uncertainty estimates for both b and $\log N$ values were found to be good approximations to 1σ uncertainties, although the distribution of fitting errors for both parameters was found to be more peaked and to contain more outliers than a normal distribution.

The measured wavelengths and column densities of absorption lines were found to be unbiased, but the measured velocity dispersions were found to be systematically biased in the sense that b was often underestimated. The bias was greater for lines which were intrinsically broader, but no meaningful calibration correction for this

effect could be made because of the large scatter in the measured values.

The shape of the column density distribution of the CNQ Lyman α lines was found to have been recovered with high accuracy, considering that a single power law was fitted to a distribution which was actually a sum of two power laws. This was despite incompleteness at low $N(\mathrm{H\,I})$ caused by S/N effects and the line blanketing effect.

The identifications of heavy element absorption lines were analysed to determine their reliability. A total of 18% of the lines were found to have been identified incorrectly. Taking into account oversights in the metal line search list, it was determined that $\sim$ 20% of the lines claimed as Lyman α were in fact unidentified metal lines, despite careful searches. If this is a typical rate of misidentification, great care must be taken when interpreting the results of published work on the Lyman α forest.

The correlation between velocity dispersion and column density for the Lyman α lines was investigated. The effects of metal line contamination, line detection thresholds, and fitting errors combined to strengthen the existing correlation. The b and $\log N$ values of the Lyman α lines in the CNQ spectrum were correlated, so the strength of the induced correlation in uncorrelated data could not be determined. Nonetheless, the enhancement of the correlation was revealed and partially quantified, and the results can be used to better assess claims of a real correlation.