Maximum entropy methods applied to NMR and mass spectrometry

Hughes, Leslie Peter

MAXIMUM ENTROPY METHODS APPLIED TO NMR AND MASS SPECTROMETRY

By

LESLIE PETER HUGHES BSc, MRSC

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Durham.

2001.

MAXIMUM ENTROPY METHODS APPLIED TO NMR AND MASS SPECTROMETRY

By

LESLIE PETER HUGHES BSc, MRSC

Maximum Entropy data processing techniques have been widely available for use by NMR spectroscopists and mass spectrometrists since they were first reported as a tool for enhancing damaged images. However, the techniques have been met with a certain amount of scepticism amongst the spectroscopic community; not least their apparent ability to get something for nothing.

The aim of the work presented in this thesis is to demonstrate that if these techniques are used carefully and in appropriate situations a great deal of information can be extracted from both NMR and mass spectra. This has been achieved by using the Memsys5 and Massive Inference algorithms to process a range of NMR and mass spectra which suffer from some of the problems which are commonly encountered in spectroscopy, i.e. poor resolution, poor sensitivity, how to process spectra with a wide range of peak widths.

The theory underlying the two algorithms is described simply and the techniques for selecting appropriate point spread functions are outlined. Experimental rather than simulated spectra are processed throughout.

Throughout this work the Maximum Entropy results are treated with scepticism. A pragmatic approach is employed to demonstrate that the results are valid.

It is concluded that the Maximum Entropy methods do have their place amongst the many other data processing strategies used by spectroscopists. If used correctly and in appropriate situations the results can be worth the investment in time needed to obtain a satisfactory result.

MEMORANDUM

The research presented in this thesis has been carried out predominantly at Unilever Research, Port Sunlight Laboratory, between April 1996 and March 2001. It is the original work of the author unless stated otherwise. None of this work has been submitted for any other degree.

ACKNOWLEDGEMENTS

I would like to extend my thanks to my two supervisors, Dr. Kenneth S. Lee (Unilever Research, Port Sunlight Laboratory) for many useful discussions, and Prof. Robin K. Harris (University of Durham), not only for his guidance but also for his patience and understanding whilst I have been conducting this research.

I would also like to thank my wife, Judith, and our children, Matthew and Amy. They have had to put up with a great deal over the past years and without their encouragement I would not have reached this stage.

Finally, I would like to mention my friend, Dr Edward Smith, who sadly died whilst I was conducting my research. Ed introduced me to NMR and without his guidance and encouragement during the early years of my career I would not have grown to realise the wonders of this technique. We always enjoyed our work.

The financial support of Unilever Research, Port Sunlight Laboratory, is acknowledged.

CONTENTS

CHAPTER 1: INTRODUCTION

Probabilistic processing of spectroscopic data is a relatively new extension of the principles learnt and applied in such diverse fields as image analysis and telecommunications[1,2].

Conventional data processing relies on linear techniques to improve one particular facet of some damaged dataset. The procedures usually involve a compromise between, for example, resolution and signal: noise ratio (S/N). The research described in this thesis employs the Maximum Entropy algorithm (Memsys5), developed by Skilling and Bryan[3], and the later Massive Inference algorithm (MassInf), developed by Skilling and Sibisi[4], to demonstrate the level of extra information that can be recovered from spectroscopic data. In the context of this thesis the term MaxEnt will be used to infer Bayesian analysis based on either the Memsys5 algorithm or the MassInf algorithm.

It has become clear that the few spectroscopists who actively use the various MaxEnt algorithms are more concerned with tackling the fundamental problems of data processing, e.g. truncation artifacts, rather than applying the principles learnt to real systems. It is perhaps this lack of industrial application, which has led to the scepticism surrounding these techniques, that was the main driving force for undertaking this research. The following chapters will describe the application of the principles of Bayesian analysis to a range of spectroscopic problems, e.g. the determination of polymer microstructure from severely overlapped NMR data.

The preliminary results describing the determination of polymer microstructure by NMR and MaxEnt data processing were presented in the form of a poster at the 16th International Conference On Maximum Entropy and Bayesian Analysis (MaxEnt '96), Kruger National Park, South Africa, August 1996. (See Appendix 1).

1.1 Introduction to Maximum Entropy Techniques

The key difference between the algorithms used in this research and conventional data processing is that the new algorithms use a probabilistic approach to quantifying the position and intensity, along with error bars, of any feature in a spectrum. As will be discussed in Chapter 2, both

the algorithms described in this thesis have their foundations in Bayesian analysis, a general method for inferring the form of a probability distribution.

Conventional data processing techniques try to filter the damaged spectrum somehow in order to recover, in one-step, what the original spectrum must have been. These linear methods, e.g. line-broadening, are limited because there is always a compromise between resolution (the ability to separate peaks) and the signal:noise ratio (S/N) found in the spectrum, i.e. as the resolution increases so the S/N decreases. Furthermore, any artifacts or spikes present in the raw data are likely to be enhanced with traditional processing methods.

In contrast, the MaxEnt algorithms never actually process the experimental spectrum. [5] The experimental spectrum is used only as a reference with which to compare the MaxEnt result at any point on its iteration cycle. The algorithms sample the probability cloud of possible trial spectra and compare each with the starting data. The difference between each trial spectrum and the original, taking into account the noise, is used to guide the algorithm to its next better choice of possible candidates. For the Memsys5 algorithm, the optimum candidate is chosen as the one with the Maximum Entropy, i.e. that trial spectrum which best fits the experimental spectrum but contains the minimum structure. For MassInf the concept of optimum candidate is discarded and the full probability distribution of plausible spectra are considered, and if required the user can take an arithmetic mean of a number of plausible spectra to obtain a spectrum for comparison. Apart from a few standard instructions the only input to either Memsys5 or MassInf is a peak profile or point spread function (PSF) and an estimate of the RMS noise level in the data (Sigma). When necessary the PSF may be optimised from the program's output diagnostics.

In summary, the MaxEnt result is a synthetic reconstruction of the experimental spectrum. It leads to quantification of position and intensity for any feature in a spectrum, accompanied by probabilistic error bars. The MaxEnt reconstruction is largely free from noise and artifacts. Simultaneous improvements in both resolution and S/N are possible.

Whilst Memsys5 will be shown to produce perfectly acceptable results, in practice it suffers from an internal mathematical inconsistency. MassInf avoids this difficulty and gives a new method

of probabilistic data processing, which has no internal mathematical contradictions; i.e. it uses the full probability distribution of plausible spectra. The differences between the two algorithms will be discussed in Chapter 2.

1.2 This thesis

This thesis falls into two main sections: techniques (Chapters 2 and 3) and application (Chapters 4-7). Chapter 2 describes the necessary mathematical background of Bayesian data processing as applied to NMR spectroscopy and Mass spectrometry. It is complete in terms of providing all the necessary mathematical concepts. However, the detailed mathematics of the algorithm is beyond the scope of this work and is excluded. All the NMR techniques employed in this report are well understood both in terms of theory and application. Accordingly, following a brief introduction to the principles of NMR, Chapter 2 describes only those particular aspects of NMR theory which give rise to the underlying problems that this thesis has tried to address, e.g. line-broadening. Chapter 3 describes how the Maximum Entropy techniques employed in this work are used in practice. It includes a description of the Maximum Entropy hardware and software and describes the preliminary data processing required to change the experimental data into a form appropriate for data analysis. The basic operation of the Memsys and Massive Inference software is discussed and the link between the software inputs and the appropriate theory is highlighted.

Chapter 4 introduces the chemical characterization of sodium carboxymethyl cellulose as an area which may benefit from a more rigorous approach to spectral data processing. The NMR spectra of these systems suffer from very poor signal:noise ratio, especially if the spectra are of commercially available systems, and normally rely on hydrolysis methods in order to realise useable NMR spectra. This chapter describes how the MaxEnt algorithms can be used to 'de-noise' such spectra providing valuable information without the need for chemical modification of the sodium carboxymethyl cellulose.

Chapter 5 describes the application of MaxEnt techniques to the ^{13}C NMR spectrum of a styrene / maleic anhydride copolymer. It describes how the techniques can simultaneously improve spectral resolution and the signal:noise ratio in the spectrum.

Chapter 6 is concerned with the application of these techniques to NMR spectra which contain peaks of significantly different linewidth. The concept of using Linear Prediction techniques as a pre-processing stage is introduced.

Chapter 7 describes the application of MaxEnt techniques to the electrospray mass spectra of complex dye mixtures. The techniques are used to overcome some of the problems associated with assigning peaks in such mixtures and the measurement of the peaks areas is shown to be consistent with the current models for the electrospray ionisation process.

Chapter 8 summarizes the principal conclusions from the work described and highlights areas where, in the future, these techniques may become a viable alternative to conventional processing methods.

References:

1. B R Frieden, *J. Opt. Soc. Am.*, **62**, 511, (1972)

2. C E Shannon, *Bell System Tech. J.*, **27**, 379, (1948)

3. J Skilling, R K Bryan, *Mon. Not. R. Astr. Soc.*, **211**, 111, (1984)

4. Private Communication 'A Massive Odyssey', J Skilling

5. MaxEnt User's Manual

CHAPTER 2: THEORY

This chapter is split into two sections that outline the theory behind the Maximum Entropy algorithms, i.e. Memsys and MassInf. The theory of NMR is also outlined in a third section.

Section 2.1 describes the Memsys algorithm in a form which can be applied to NMR spectroscopy and mass spectrometry and which has the following properties:

- It is consistent in terms of nomenclature.
- It is complete in terms of providing all the necessary mathematical concepts. However, the detailed mathematics of the algorithm is beyond the scope of this work and is excluded.
- It describes the origins of the parameters used in the practical application of the Memsys algorithm.

The theoretical problems associated with the Memsys algorithm are discussed.

Section 2.2 presents the MassInf algorithm in a manner that fulfils the above criteria and highlights how this algorithm overcomes some of the difficulties associated with Memsys. Again, the detailed mathematics of the algorithm is beyond the scope of this work. The differences between the two algorithms are highlighted.

All the NMR techniques employed in this report are well understood both in terms of theory and application.[1,2] Accordingly, following a brief introduction to the principles of NMR, section 2.3 describes only those particular aspects of NMR theory which give rise to the underlying problems that this thesis has tried to address, e.g. line-broadening.

2.1: Memsys theory

This chapter presents an outline of the theory underlying the Memsys algorithm and is largely based on a collection of expository essays by Davies et al.[3], Daniell[4], Skilling[5], and Hore[6]. Wherever possible, the mathematics is deliberately kept to a minimum. For a review of the use of Maximum Entropy methods in NMR spectroscopy see Stephenson[7].

2.1.1 The problem

Davies[3] states that spectroscopic data are affected by a variety of imperfections which include:

- data truncation,
- distortion introduced by the spectrometer detection system,
- corruption due to thermal and digital noise,
- the finite bandwidth caused by the finite data sampling rate.

He concludes that any data processing method, which acts on the data directly, will incorporate these imperfections into the resultant spectrum. An example of such a direct manipulation technique is the discrete Fourier transform (DFT) used in modern pulsed NMR spectroscopy. When using the DFT, if a spike is observed in the time domain it will be translated into the frequency domain as a sine function. On the other hand, the direct approach to data processing does have advantages that include speed and linearity, i.e. the relative intensities of resonances of different widths, shapes and frequencies are not distorted. The DFT is not model-dependent. Ferrige and Lindon[8] have reviewed conventional data processing methods involving the Fourier transform. They conclude that with suitable data filtering, e.g. multiplication of the time domain data with a mathematical function, improvements in either spectral resolution or signal-to-noise ratio can be achieved. However, it is difficult with Fourier techniques to achieve both enhancements simultaneously. One method that claims to offer both improvements simultaneously

is the TRAF function developed by Traficante and Nemeth.[9] The TRAF function was derived by examining the resultant Free Induction Decay (FID) achieved by adding one FID to another FID that had been reversed in the time-domain, whilst simultaneously taking advantage of the matched filter condition, i.e. multiplying the FID with an exponential function that decays at exactly the same rate. Traficante points out that in cases where the signals do not have a perfect exponential decay, application of the TRAF function will not yield the expected improvement in both sensitivity and resolution. This is because the derivation of the function is based on purely exponential decays. Most experimental FIDs are not purely exponential. Furthermore, an FID that is composed of signals that have different spin-spin relaxation times (T_2) will produce baseline distortions with the TRAF function in the form of sinc (sine x / x) sidelobes.

Maximum Entropy methods, such as that implemented in the Memsys algorithm, do not act directly on the experimental data but calculate a theoretical spectrum. As the experimental time domain data are used for comparison only, any imperfections are not translated into the frequency domain. Further, as will be demonstrated later in this thesis, simultaneous improvements in spectral resolution and signal-to-noise ratio can be achieved using Maximum Entropy methods.

A theoretical trial spectrum should incorporate any prior knowledge of the experimental conditions, e.g. in NMR spectroscopy if the peak shape is known this can be modelled into the estimate of the spectrum. Davies [3] notes that the theoretical data are then fitted to the experimental data via some appropriate constraint, yielding either model parameters or an idealized spectrum. He describes the problem facing the spectroscopist in a mathematical form, i.e.

$$d_i = \sum_{k=1}^{N} O_{ik} f_k + \sigma_i \qquad \text{.....(2.1)}$$

where:

d_i are the imperfect experimental data

f_k represent the 'perfect' spectrum of the sample

the matrix $\boldsymbol{O}$ represents the instrumental response or blurring function

σ_i represent the noise on each digitized point.

The aim of all data-improvement methods is to obtain the best possible estimate of $\boldsymbol{f}$, using all the available information about the causes of likely imperfections in the data. This prior information is used to build a model of the blurring function, $\boldsymbol{O}$. As Davies points out, $\boldsymbol{f}$ cannot be recovered by inversion of equation (2.1), even if the matrix $\boldsymbol{O}$ is known, owing to the presence of the noise and the finite nature of $\boldsymbol{d}$. The situation is further complicated because there may be a large number of spectra, $\boldsymbol{f}$, which is consistent with the experimental spectrum, $\boldsymbol{d}$, given the noise, $\boldsymbol{\sigma}$. In the absence of extra knowledge about the spectrum there would be no reason to pick one of these trial spectra over the others. This selection problem is fundamental to Maximum Entropy data processing.

2.1.2 Summary of the constraints

There are two opposing constraints in trying to find the best $\boldsymbol{f}$. Firstly, maximising the fit to the experimental data and secondly constraining the selection of $\boldsymbol{f}$. Maximum Entropy data processing attempts to overcome this selection problem by choosing the $\boldsymbol{f}$ that contains the minimum structure, and yet is consistent with $\boldsymbol{d}$ given the noise, $\boldsymbol{\sigma}$.

2.1.2.1 Minimise the structure, maximise the entropy

Hore[6] argues that, for comparison, if the theoretical trial spectra, $\boldsymbol{f}$, are inverse Fourier transformed, the vast majority can be immediately rejected in that they bear no resemblance whatsoever to the inverse Fourier transform of $\boldsymbol{d}$, i.e. the comparison is made in the time domain. However, there is a reasonable number that will match the experimental FID closely, and in the absence of extra knowledge about the spectrum, there is no reason to prefer one of these trial spectra

rather than the others. Given that, the Maximum Entropy choice is to select that f with the minimum information content, or equivalently the maximum entropy, i.e. the maximum disorder. This choice is the most easily defended because it is the least likely to lead to over-fitting d. The selection problem is now more rigorously defined: select that f, with the greatest entropy, from all the possible f s that are consistent with d. The entropy, S, is measured in the frequency domain but σ is measured in the time domain.

The most commonly used entropy in spectroscopic applications is defined by:

$$S = -\sum_{k=1}^{N_f} \frac{f_k}{\delta_k} \ln \frac{f_k}{\delta_k} \qquad \text{.....(2.2)}$$

where:

- f_k is the (real, positive) weighted NMR intensity at the kth point in a trial spectrum that is digitised at N_f regular intervals in the frequency domain.
- δ_k is a weighting function applied to each frequency, k. It is known as the regularization parameter and, according to Hore[6], for spectroscopy it serves to incorporate prior knowledge about the spectrum baseline. Setting δ_k equal to a number much smaller than the expected peak intensities ensures that the baseline has a low intensity and any noise frequencies are not reconstructed as peaks in the final Maximum Entropy reconstruction. An alternative definition of entropy has Σf_k in place of δ_k, such that the spectral intensities are normalised. This has the disadvantage of pulling up the baseline to an unsatisfactory level. See section 2.1.4.1 for a more detailed description of the regularization parameter.

This definition of entropy is analogous to that using in statistical thermodynamics. Here, the entropy of a system is related to the number of states that are thermally accessible.[10] (see section 2.1.4.1)

Davies[3] notes that the definition of entropy in the form p ln p enforces positivity on the Memsys result. Clearly, the logarithm term cannot take a negative value. Therefore, this approach for deciding which is the **best** f can not be used when the collection of plausible results contains signals of varying phase, i.e. all the spectra must be correctly phased and contain only positive going peaks.

Thus, for a correctly phased NMR spectrum, maximising the entropy of the trial spectra enables the selection of f to be constrained to that containing the minimum structure. However, opposing this is the desire to maximise the degree of fit to $\boldsymbol{d}$.

2.1.2.2 Minimise the residuals, maximise the fit

According to Davies[3], the procedure adopted to calculate the maximum entropy spectrum is to start with a trial spectrum that is usually flat and featureless. This avoids any bias. If prior knowledge is available, e.g. the number of peaks expected in the result, this can be incorporated into the model at this stage, but it is more usual not to impose any such constraint. It has not been used in any of the examples shown in the following chapters.

The Memsys algorithm compares f with $\boldsymbol{d}$ in the time domain. Therefore, this first trial spectrum, f_1, is inverse Fourier transformed to give a mock FID, using a form of $\boldsymbol{O}$ appropriate for NMR spectroscopy, e.g. a baseline correction function. The ability to control $\boldsymbol{O}$ makes this a general technique. The mock FID is then compared with the experimental FID using an appropriate consistency test. **For NMR spectroscopy, it is assumed that the measured data are corrupted solely by additive Gaussian noise with constant variance, σ^2.** Given this approximation, a normalised chi-squared measure, χ^2, is an appropriate test. The constraint, $C(f)$, that the reconstructed spectrum must be consistent with the measured data takes the form $C(f) \leq C_0$, where C_0 is an upper bound on the allowed error. Given a prior estimate of the amount of noise in the data, C_0 should be comparable to the power of the noise, i.e. $C_0 = \Sigma|\sigma_i|^2$. Other forms of the constraint, $C(f)$, are possible, but χ^2 is the one generally used for NMR spectroscopy, i.e. $C(f) = \chi^2$.

The convex chi-squared surface, χ^2, has the following form:

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N_t}(D_i - F_i)^2 \qquad \text{.....(2.3)}$$

where:

F_i is the ith complex intensity in the mock FID

D_i is the ith complex intensity in the experimental FID

σ is the root-mean-square noise amplitude, which is assumed to be constant for all F_i.

The aim is to match D_i and F_i to within a certain tolerance specified by the noise level. This results in an essentially noise-free F_i.

2.1.3 The Memsys solution

The aim of Maximum Entropy methods is to solve this constrained optimisation problem. Hore[6] states that the Memsys task is to maximise $S(f)$ with respect to f_k, where $k = 1,2,\ldots N_f$, subject to the constraint $C(f) \leq C_0$.

Converting the above into an equivalent unconstrained optimization problem often solves such constrained problems. For Maximum Entropy reconstruction, the equivalent problem is to maximize the function:

$$Q(\lambda, f) = S(f) - \lambda C(f) \qquad \text{.....(2.4)}$$

where:

λ is a Lagrange multiplier. This is a standard mathematical method for finding the maximum of a function of several variables if the relations among the variables are known. (See the footnote at the end of this chapter for a numerical example of Lagrange multipliers).

The required solution corresponds to a critical point of Q, i.e. a point where $\nabla Q = 0$. There is no general analytical solution so numerical techniques are used. The Memsys algorithm attempts to solve for Q iteratively using a procedure developed by Skilling and Bryan[12]. This procedure comprises three search directions within the N_f -dimensional space:

- The gradient of the entropy in the spectral domain
- The gradient of the consistency test, $C(f)$, in the spectral domain
- A combination of the two in the form of the gradient of Q.

The angle between the gradient of the entropy and the gradient of $C(f)$ is an important diagnostic in determining if the algorithm is following the optimum iterative path, i.e. the Memsys trajectory. If the algorithm is following the optimum trajectory, the gradient of the entropy and the gradient of the consistency test should be parallel. Therefore, 1-cos(angle between the two gradients) = 0. This is the Memsys 'Test' parameter described in section 3.2.3.1. A mathematical description of the iterative techniques is beyond the scope of this thesis.

While a numerical solution is required in the case of the Memsys algorithm, there is a special case of Maximum Entropy reconstruction that has an analytical solution. It arises when the number of points in the reconstructed spectrum is equal to the number of experimental data points, i.e. the relationship between the trial spectrum and the mock FID is given simply by the inverse Fourier transform. Forcing the chi-squared measure to be equal to the number of observations is not the definitive rule. Alternative possibilities are discussed by Gull.[11]

2.1.4 The origin of the two constraints

As indicated earlier, the trial spectrum should incorporate any prior knowledge of the experimental conditions and the iterative search procedure should 'learn' from previous estimates. This is essentially a Bayesian probability approach to data analysis. Probability theory allows us to assign a numerical code that expresses our belief as to what the trial spectrum should look like. Furthermore, to asses the reliability of a maximum entropy (or other) selection, a probabilistic

description is necessary. There are two basic rules of probability theory, from which Bayes Theorem is derived. These rules form the basis of Maximum Entropy data processing and are illustrated below with everyday examples.

1. The sum rule

$$P(A) + P(\overline{A}) = 1 \qquad \text{.....(2.5)}$$

where:

P(A) is the probability of A, i.e. a proposition is **true.**

$P(\overline{A})$ Is the probability of not A, i.e. a proposition is **not true.**

Equation (2.5) is read as "the sum of the probability of event A being true and the probability of that event being false is unity." This is intuitively correct given that *P* is defined as a fractional probability. For example, consider tossing a coin. For an unbiased coin the probability of a head is ½ and that of a tail is ½. This is a normalised probability, i.e. fractional.

2. The product rule

For two events, A and B, given that propositions A and B can each be either true or false,

$$P(A,B) = P(A)P(B|A) \qquad \text{.....(2.6)}$$

where:

P(A,B) is the joint probability that proposition A and B are both true.

P(B|A) is the conditional probability that proposition B is true given that proposition A is true.

Equation (2.6) is read as "the probability that the compound event A and B will happen is the product of the probability that A will happen and the probability that B will happen if A does." A good description is given in Boas[12] and is reproduced here for clarity.

Visualisation of the Product Rule[12]

Consider, two successive events A and B.

Let P(A) = probability that A will happen

Let P(A,B) = probability that both A and B will happen

Let P(B|A) = probability that B will happen once A has happened

Let N = Total number of sample points in a uniform sample space

Let N(A) = Number of sample points corresponding to event A

Let N(B) = Number of sample points corresponding to event B

Let N(AB) = Number of sample points corresponding to the compound event A and B

This can be pictorially represented in a N x N matrix as shown in Figure 2.1 where the

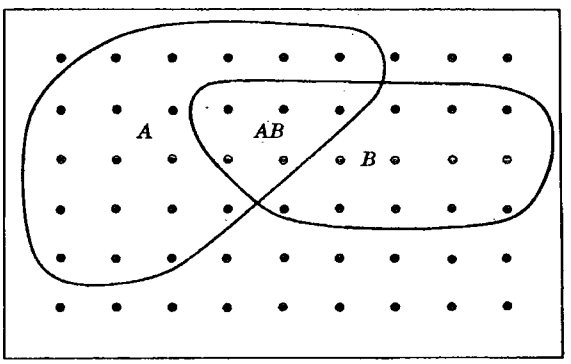


Figure 2.1

probability of the different events is represented by the asymmetric shapes, i.e. all the points which correspond to A happening are encircled and labelled appropriately. Then,

$$P(A,B) = \frac{N(AB)}{N}$$

$$P(A) = \frac{N(A)}{N}$$

According to Boas, N(A) is the number of sample points corresponding to event A; the N points in the original sample space all had the same probability so we can assume that if we cross off all the points corresponding to A not happening, the remaining N(A) points also have equal

probability. Thus we have a new uniform sample space consisting of N(A) points. N(AB) of these points correspond to the event B assuming A. Thus,

$$P(B|A) = \frac{N(AB)}{N(A)}$$

From these three probabilities,

$$P(A,B) = \frac{N(AB)}{N} = \frac{N(A)\,P(B|A)}{N} = \frac{P(A)\,N\,P(B|A)}{N} = P(A)\,P(B|A)$$

Alternatively, $P(B,A) = P(B)\,P(A|B)$

(c.f. equation 2.6)

If prior information, I, is available, this can be incorporated into the two probability rules in the following way:

$$P(A|I) + P(\overline{A}|I) = 1 \qquad \text{.....(2.7)}$$

$$P(A,B|I) = P(A|I)P(B|A,I) \qquad \text{.....(2.8)}$$

The product rule allows two alternative factorisations. Thus equation (2.8) can be written in two equivalent forms, i.e. equations (2.8 and 2.9):

$$P(A,B|I) = P(B|I)P(A|B,I) \qquad \text{.....(2.9)}$$

Rewriting equations (2.8) and (2.9) in a form that is appropriate for our spectroscopic problem gives:

$$P(f,D|I) = P(f|I)\,P(D|f,I) = P(D|I)P(f|D,I) \qquad \text{.....(2.10)}$$

$P(f|I)$ is defined as the **prior** and is the probability assigned to the mock spectrum before the actual data are acquired, i.e. what the probability is of finding structure in the spectrum given only our background information. This could take the form of expert knowledge of the system.

$P(D|f,I)$ is the **likelihood** and is a measure of how likely our damaged spectrum is given our estimate of the mock spectrum and our background knowledge of the problem.

$P(D|I)$ is the **evidence** and quantifies how well the actual data could have been predicted in advance given our background knowledge of the problem.

Finally, $P(f|D,I)$ is the **posterior** and is the result we require. It quantifies our inferences about the mock spectra, f. Memsys attempts to find the most probable f by maximising $P(f|D,I)$, i.e. the probability of finding the perfect spectrum, f, given the measured spectrum, $\boldsymbol{D}$, and our background information, $\boldsymbol{I}$. It also gives a measure of the reliability of this choice. This is derived from the spread of reasonably plausible f. This distribution encapsulates all that we know about the mock spectra, f. From this complete distribution, we may want to extract the most probable value, the mean, and error bars on the distribution.

Rearranging equation (2.10) gives an expression for the posterior in the form of an expression known as **Bayes' theorem**:

$$\boxed{P(f|D,I) = P(f|I)\,\frac{P(D|f,I)}{P(D|I)} \qquad \text{.....(2.11)}}$$

Bayes' theorem tells us how to update prior probabilities in the light of experimental data. This gives a posterior probability that includes all relevant information. From Bayes' theorem it is

apparent that the measurements which define $P(D|f,I)$ do not fully describe our result. The prior, $P(f|I)$, also needs to be assigned. Much of the skill of Bayesian data processing lies in developing priors that reflect the experiences and assumptions of scientists, i.e. building the model. It is in this area that most of the current research in Bayesian methods lies.

2.1.4.1 The prior

As always, in probabilistic analysis, the prior must be assigned first, i.e. how we tell the analysis what sort of f we expect to find before factoring in our data. Different priors will result in different posterior probabilities; hence the choice of prior is not arbitrary. The use of the entropy prior can be justified in several ways. One method uses the so-called Monkey Model described by Jaynes[14] and summarized by Daniell[4].

Daniell considers that the problem is to determine some function, F(x). x-space is then divided into cells of width Δx. Δx is much smaller than the experimental resolution, so no approximation is involved. If the intensity of F(x) in a cell, i, is f_i and this intensity is quantized such that $f_i = n_i\delta$, where the quantum δ is so small that it again is not an approximation and the numbers n_i are all large, then the function F(x) can be represented by a set of integers, $n_1, n_2, \ldots$

Daniell then imagines the classic problem of a team of monkeys, throwing balls into boxes. The boxes correspond to the cells of F(x). This is a completely random process, with no regard for any experimental data that is collected as the monkeys continue to throw the balls. The probability that the monkeys produce $n_1, n_2, \ldots n_n$ is the combinatorial expression:

$$P(f_1, f_2 \ldots) \propto 2^{-M} \frac{M!}{n_1! n_2! \ldots n_n!} \qquad \ldots\ldots\ldots\ldots\ldots(212)$$

where:

M is the total number of balls.

The Monkey Problem exhibits the important constraint of **independence**. Independence refers to the distinct boxes. Learning about F(x) through knowledge of the number of balls in one box does not help to predict how many balls will be placed in another box without knowing the total number of balls to be thrown.

Expression (2.12) is effectively the number of different ways of throwing the balls to give the same result and is analogous to the weight of a purely random configuration found in statistical thermodynamics.[10] A general configuration (n_1, n_2 ....n_n) can be achieved in W different ways, where W is the weight of the function, i.e. the number of distinguishable ways in which M balls can be sorted into boxes with n_i in bin i.

Expanding equation (2.12) gives:

$$log(P) = log(2^{-M}) + log(M!) - log(n_1!) - log(n_2!) - log(2^{-n1}) - log(2^{-n2}) \quad ...(2.12a)$$

for n_n configurations.

Daniell then notes that since n_i are large, Stirling's approximation gives:

$$log(P) = log\left(\frac{2^{-M}}{(2^{-n1}.2^{-n2})}\right) + (MlogM - M) - \Sigma(n_i log(n_i) - n_i)) \quad .....(2.12b)$$

$$log(P) \propto const. + \Sigma_I\, n_I - \Sigma(n_i log(n_i) \quad .....(2.12c)$$

$$P \propto const.\ exp(\Sigma(n_i log(n_i)) \quad .....(2.12d)$$

---

[a] Stirling's approximation simplifies to $\ln x! = x \ln x - x$.

In the case of spectroscopy,

$$P(f_1, f_2 ....) \propto \exp(\alpha S) \qquad .....(2.13)$$

where:

$$S = -\Sigma f_i \log \frac{f_i}{\delta} \qquad .....(2.14)$$

and α and δ are constants depending on M.

The quantity S is called the entropy of the function F(x) and is used as the **Prior** equation 2.11 and is normalised by the function α. This normalisation parameter is related to the amount of material in the balls and is called the *regularization parameter*. In practical spectroscopy α is unknown and is determined along with F(x) and defines the relative importance of the entropy and the data in obtaining a solution.

Equation (2.14) is analogous to equation (2.2) and is one of the constraints on finding that spectrum which is consistent with the data but which contains the minimum structure.

A variant of the classic maximum entropy prior, described by equation (2.13), is the so-called pre-blur maximum entropy technique. The classic prior tends to produce irregular results by amplifying the noise level found in the estimates of f. This is particularly true for mass-spectrometry, which exhibits Poisson noise statistics. Consequently, the classic prior was altered to make it spatially smoother without destroying the entropy maximization.

Skilling[15] states that this was achieved by assuming that the spectra we seek, f, are a convolution of some hidden spectrum, $\boldsymbol{h}$, and some 'intrinsic correlation function' (ICF).

$$f = ICF \otimes h \qquad .....(2.15)$$

The ICF is smooth and broad, which results in correspondingly smooth f. Placing an entropic prior on $\boldsymbol{h}$ gives:

$$P(h) = \exp(\alpha S(h)) \qquad \text{.....(2.16)}$$

The experimental data, $\boldsymbol{D}$, are related to $\boldsymbol{h}$ via a composite convolution with both the ICF and the instrumental response, $\boldsymbol{O}$. (C.f. Equation(2.1))

$$D = O \otimes f + noise = O \otimes ICF \otimes h + noise \qquad \text{.....(2.17)}$$

Skilling notes that instead of reconstructing the experimental spectrum directly, the pre-blur method uses classic maximum entropy techniques to find the hidden spectrum. This is then blurred to give f. He continues that when using the pre-blur, the functional form of the ICF is required, i.e. its shape and width. This is the point-spread-function (PSF) parameter used as an input to Memsys.

Properties of the Memsys Prior

Memsys is the MSL implementation of the general Maximum Entropy method. In order to find the 'best' maximum entropy solution a number of constraints were assigned to the best solution and to the Memsys prior. Firstly, the best f must be positive. The spectrum we seek must have a positive intensity value at each point. Secondly, the best f must be additive. The additivity refers to the total intensity residing in a specified area. Skilling illustrates these points with the example of light intensity being both positive and additive, with its sum representing a physical energy flux, whereas the amplitude of incoherent light is not additive. Secondly, it was hoped that the Memsys prior would be independent and divisible. Independence refers to the distinct domains of the best f; learning about our best f in one domain does not help to predict the best f using another domain. Divisibility means that the cells (data points describing the best) f can be divided arbitrarily finely.

Problems with the Memsys Prior

The Memsys prior proved not to be divisible, and as the best f was divided into finer and finer cells the early Memsys algorithms did not behave sensibly. This can be illustrated by considering equation 2.14. As the best f is divided onto a finer grid it becomes clear that:

$$N \text{ Log } (N) \neq 2 \, (N/2 \text{ Log } (N/2))$$

This problem was due to the Gaussian approximation. This was needed to practically compute the probability maximum. Furthermore, the Gaussian approximation destroyed the other property of positivity. Samples for f could and did go negative. Although, as will be seen in this thesis, the Memsys prior still works well for most spectroscopic problems, these internal mathematical inconsistencies led to the development of the Massive Inference (MassInf) prior. (see section 2.2)

2.1.4.2 Likelihood

$P(D|f,I)$ is the likelihood. It is the conditional probability of acquiring that particular spectrum, D, given f and our background understanding of the problem, I. Skilling observes that the form of the likelihood ought, in an ideal world, to be provided by the instrument manufacturer. It could be observed over many samples. This would allow the manufacturer to derive an idea of what the output, D, would be for any given input. Of course, this prediction of D will be imprecise and any variation in D is allowed for in the form of $P(D|f,I)$.

The normal method for determining the likelihood is to assume a linear experiment with Gaussian noise (see equation 2.1). Skilling observes that knowledge of the normal distribution gives the likelihood as a function of the usual chi-squared statistic. (c.f. equation 2.3)

$$P(D|f,I) = \text{Func. } (C(f)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.18)$$

If the data do not follow normal Gaussian statistics then the likelihood should be changed to accommodate the new circumstances. For example, mass spectrometry exhibits 'shot' or Poisson noise. A small change in the likelihood function will accommodate this change in the noise characteristics. The form of the likelihood function is well known for both Gaussian and Poisson noise and will not be described here.

The likelihood function can always be determined, apart form some ignorance of the exact shape of the point-spread-function (hidden spectrum).

2.1.4.3 Evidence

P(D|I) is the evidence and in the Memsys algorithm this is measured in decibels:

$$Log\,\mathrm{Pr}\,ob \sim \log_{10} P(D|I) = \log_{10} Evidence \qquad \text{.....(2.19)}$$

The *Evidence*, as used by this algorithm, is described in Chapter 3.

The evidence depends on the theorist who is advising on the form of the prior for f and is defined as:

$$P(D|I) = \sum_f P(f, D|I) \qquad \text{.....(2.20)}$$

From equation (2.20), evaluating the evidence involves a sum over all possible f and cannot be calculated directly. Skilling approximates the evidence.

The evidence is a crucial safety guard against arbitrary misuse. As there is complete freedom in the choice of prior, the Memsys result could, in principle, take almost any form. The saving grace is that an inappropriate prior will almost never manage to predict the data well, so that it will have only a low numerical value for the evidence. Hence, the evidence, through a series of trials, allows the Memsys operator to refine his choice of prior (hidden spectrum).

2.1.4.4 Posterior Inference

$P(f|D,I)$ is the posterior inference and quantifies the plausibility of the different spectra, f, after the data, D, have been taken into account. This is the main objective of the Bayesian approach to data analysis. The aim is to maximise the posterior and hence maximise the entropic prior.

In NMR spectroscopy, where any f can be described by thousands of data points, the only feasible way of presenting the inference is directly, as a list of 'typical' f, randomly sampled from the posterior inference. For clarity, Memsys augments this inference with the most probable f. Fortunately, all the major properties of f, e.g. location and intensity of lines, can be found after only a dozen or so random samples.

2.2: MassInf theory

As described earlier, the Gaussian approximation of the posterior inference detroys the property of positivity and the entropic prior proves not to be divisible. Although the Memsys algorithm works well in practice and has been used regularly in this thesis, the mathematics behind the Memsys prior were flawed.

The clue as to how to develop a new prior comes from the divisibility argument. Skilling[16] describes this problem by giving a simple example. Consider our prior knowledge about some quantity **F**. Let the intensity of **F** in one domain, **A**, be one unit with a standard deviation of one unit. If the domain **A** is divided into **N** independent cells, then the total intensity should be divided into N^{-1} equal parts, each with a standard deviation of $N^{-0.5}$. Therefore, the intensity in each cell, i, is:

$$F_i \approx N^{-1} \pm N^{-0.5} \qquad \text{.....(2.21)}$$

If **N** is large, the standard deviation greatly exceeds the mean, i.e. positivity is violated. Positivity must be enforced.

No F_i can be negative, so the standard deviation can only be reached with a prior that usually makes F_i very small, but just occasionally gives it a substantial value.

This suggests that the prior on **F** must be **atomic**. This implies that, no matter how many cells are used to describe the original problem, all of the intensity is almost certainly contained in a bounded number of cells.

The prior that is actually assigned in the MassInf kernel uses a distribution that has a finite number of atoms that are distributed at random. For completeness, the prior is assigned on a small cell of importance, **ε**, and is:

$$P(F_i) \propto \delta(F_i) + (\varepsilon / \alpha q)\exp(-F_i / q) \qquad \text{.....(2.22)}$$

The parameter α is similar to the parameter α found in the Memsys5 prior (see equation (2.13)). In both cases, it represents the expected degree of macroscopic uniformity. In MassInf, uniformity arises by adding more atoms, in Memsys by making the prior increasingly peaked around the global maximum at S = 0.

The parameter q governs the scale of the expected quantities F.

One feature of MassInf is that there is no most probable spectrum. The conventional reconstruction for display is the mean <**f**> of the posterior.

Skilling summarizes the differences between the two algorithms thus:

	Memsys	MassInf
Displayed result	Max S(f)	Mean <f>
Positive	Yes	Yes
Additive	Yes	Yes
Sample *f* positive	No	Yes
Sample *f* additive	Yes	Yes
Divisible prior	No	Yes
Approximation	Gaussian	None

2.3: NMR Theory

All the NMR techniques employed in this report are well understood both in terms of theory and application.[1,2] Accordingly, following a brief introduction to the principles of NMR spectroscopy, this section describes only those particular aspects of NMR theory which give rise to the underlying problems that have been addressed in this thesis, e.g. line-broadening.

2.3.1 Introduction

Many common nuclei have a 'motion' which gives them angular momentum. This motion is called spin, and according to quantum mechanics it is quantized with a quantum number, I, having possible values of positive multiples of 1/2. When a direction in space is specified by the application of an external magnetic field the orientation of the angular momentum also becomes quantized. The relevant quantum number is M_I and can have values between -I and +I in unit steps. The removal of the energy level degeneracy (Zeeman splitting) results in:

- unequal Boltzmann populations in the energy levels
- the possibility of transitions between the energy levels.

In order to maximise the population differences, strong, static external magnetic fields (B_0) are employed. Typical magnetic field strengths put the transition frequency (Larmor frequency) in the radiofrequency region of the electromagnetic spectrum.

At equilibrium, the nuclear spins of a particular isotope, say ^1H, have a net magnetisation vector parallel to the applied magnetic field, say along the z-direction of a Cartesian co-ordinate system. Following a radiofrequency pulse, of amplitude B_1, in the xy- plane perpendicular to this, the bulk magnetisation is induced to precess about B_1. This complex precession has two time-dependent terms (motion about B_0 and about B_1). To overcome the visualisation of such a complex motion, a rotating frame of reference is adopted, i.e. a frame rotating about B_0, at the same frequency as the carrier radiofrequency. In this frame of reference, the bulk magnetisation is tipped towards the xy-plane, a motion termed nutation. Any component of this bulk magnetisation that lies in the xy-plane when the RF pulse is switched off can be detected by the signal it induces in a receiver coil placed in this plane. The spin-system, in the absence of B_1, is able to relax to equilibrium, and the induced signal is called the free induction decay (FID). The detected FID contains characteristic frequencies of all the protons contributing to the net magnetisation vector. For solutions ,these frequencies vary depending on the local magnetic environment of a nucleus (chemical shift), and interactions between two or more nuclear spins (spin-spin coupling).

The acquired FID is then subjected to a Fourier Transform to give the characteristic NMR frequency spectrum.

2.3.2 Sensitivity

The energy difference between the Zeeman energy levels is very small compared to thermal energy and so the population difference between the spin states is minute. This leads to small NMR intensities and correspondingly small S/N ratios, which restrict the accuracy of intensity and frequency measurements.

In considering ways to improve the sensitivity of the NMR spectrum by data processing techniques alone it is necessary to distinguish the differences between genuine signals and noise. In

the frequency domain, genuine signals tend to be smooth and usually sharp while noise is random and has a broad frequency distribution. These differences may be used to discriminate between noise and signal by applying a digital filter in the time-domain. This linear smoothing is normally applied before Fourier transformation, by multiplying the FID by a weighting function, e.g. a decaying exponential. This process gives greater weighting to the time domain points with the highest S/N ratio, and lesser weighting to those points with the lowest S/N ratio. This has the corresponding effect of broadening lines in the frequency domain. This compromise between sensitivity and linewidth is common to all linear processing techniques. A paper by Ferrige et al.[17] describes most of the available linear filters.

The Maximum Entropy approach to sensitivity improvement is somewhat different. As described above, genuine signals tend to have finite width. This would normally distinguish them from random noise. This difference is exploited by designing the PSF such that it is characteristic of the genuine signal alone. In Maximum Entropy data processing this has the effect of essentially 'de-noiseing' the spectrum with no loss of resolution. Any signals that do not match the applied PSF are treated as noise by the algorithm. However, this approach is limited because the algorithm is based on one input PSF. In general, NMR signals tend to have variable width. Therefore, the data have to be treated either by regridding, if possible, such that all the peaks have the same width but maintain their areas, or by treating only that part of the NMR spectrum which contains peaks of similar width.

2.3.3 Resolution

The observed NMR lineshape can be considered to be composed of two parts. Firstly, the natural linewidth, which in itself suffers from several possible origins of line-broadening, and secondly, the magnetic field distribution function which arises from inhomogeneities in the applied magnetic field. Of the several possible sources of line-broadening the most important are highlighted below:

- **Shielding anisotropy**

In liquids magnetic dipole-dipole interactions are mainly responsible for the exchange in energy between the lattice and the nuclear spin system and is characterised by the spin-lattice relaxation time (T_1). The shielding effect of a cloud of electrons around a nucleus alters the magnetic field experienced by that nucleus and hence its resonant position. The magnitude of this effect depends on the orientation of the molecule relative to the external magnetic field. In a solid, the molecules, are in fixed alignments and a range of possible frequencies is observed which contributes to broad lines. In liquids molecular tumbling of the molecules averages out this spread of resonant frequencies and a relatively sharp line is observed. This tumbling effect can be simulated in solids by rapidly rotating the sample when inclined at the magic angle to the applied magnetic field, a technique known as magic-angle spinning.

- **Spin-spin coupling**

Nuclear spins that are J- coupled affect the magnetic field strength that they each experience since they each have a magnetic moment. For any given nucleus, this will experience an interaction with the magnetic moments of all other magnetic nuclei in the immediate vicinity. The size of this interaction will depend primarily on the through bond distance for so-called scalar coupling and the through space distance for dipolar coupling. Other influences on the size of this effect depend on their relative orientations and the nature of the bonding present. This spin-spin interaction manifests itself in the NMR spectrum as peak slitting or as a spread of resonant frequencies, i.e. broadened peaks. In the work described in this thesis, a system containing aluminium and hydrogen nuclei, heteronuclear coupling is observed between protons and aluminium. This may be removed with heteronuclear decoupling techniques. Further, homonuclear decoupling may exist between aluminium nuclei. Homonuclear decoupling is much more difficult to achieve. For the spectra shown in this thesis no decoupling techniques have been employed.

- **Chemical shift dispersion**

If the nuclei of interest are in a range of similar environments there will be slight differences in chemical shifts that will produce a broadening of the lines in the spectra.

- **Electric quadrupole moments.**

Nuclei, such as ^{27}Al ($I = 5/2$), with spin I greater than ½ have an electric quadrupole moment. This quadrupole moment can interact with any gradient in the electric field surrounding the nucleus. This results in a quadrupolar splitting of the enery levels which is supplementary to the Zeeman splitting created by the application of the external magnetic field in NMR spectroscopy. This quadrupole splitting results in a change in the observed transition frequency and hence a broadening of the NMR signal. In the examples presented in this thesis, all the ^{27}Al spectra recorded are of aqueous aluminium chlorohydrate solutions. In this case, the averaging effects of molecular tumbling help to reduce the quadrupolar effect. Further, the size of this residual effect will depend on the asymmetry of the ligand field. For example, the resonances of low symmetry Al^{3+} species can be so broad (linewidth greater than a few thousand Hz) that they are lost in the baseline, whereas when the ligand symmetry is high, e.g. Al hexahydrate, the Al resonance is sharp (a few Hz). Therefore, even in solutions the electric quadrupole moment can interact with the electric field gradient of the nucleus and so introduce further relaxation mechanisms. For example, in chlorine this relaxation mechanism is so effective, that it is practically non-magnetic as far as NMR measurements are concerned.

Footnote: Numerical Example Of Lagrange Multiplier[18]

This numerical example of the use of a Lagrange multiplier is taken directly from ref. [18].

By Lagrange's method, if among all points (x, y) satisfying the constraint g(x, y) = 0, the function $f(x, y)$ takes on its greatest or least value at (x_0, y_0), then there is some number λ such that

$$f_x(x_0, y_0) = \lambda\, g_x(x_0, y_0) \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

and

$$f_y(x_0, y_0) = \lambda\, g_y(x_0, y_0) \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where λ is called a Lagrange multiplier and f_x, f_y, g_x, g_y are partial derivatives.

We want to minimise the function:

$$f(x, y) = x^2 + y^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

subject to the constraint:

$$g(x, y) = 3x + 4y - 15 = 0 \quad \ldots\ldots\ldots(4)$$

Solution: We begin by calculating partial derivatives:

$f_x(x, y) = 2x \qquad g_x(x, y) = 3$

$f_y(x, y) = 2y \qquad g_y(x, y) = 4$

By Lagrange's method, we seek a number λ such that equations (1) and (2) are satisfied.

By Equation (1)

$2x = \lambda 3$, i.e. $x = 3\lambda / 2$ .....(5)

and by equation (2)

$2y = \lambda 4$, i.e. $y = 2\lambda$ .....(6)

Substituting (5) and (6) into (4) gives

$3(3\lambda/2) + 4(2\lambda)-15 = 0$

and hence $\lambda = 6/5$ .....(7)

Substituting (7) into (5) and (6) gives

$x = 9/5$

$y = 12/5$

Hence, the only point on the line $3x + 4y = 15$ at which $f(x, y)$ can be a maximum or minimum is (9/5,12/5).

REFERENCES

1. J. K. M. Sanders, B. K. Hunter, Modern NMR Spectroscopy, Oxford University Press (1988).
2. R Freeman, A Handbook Of Nuclear Magnetic Resonance, Longman Scientific and Technical, (1988).
3. S. Davies, K. J. Packer, A. Baruya, A. I. Grant, *Enhanced information recovery in spectroscopy using the maximum entropy method*, in *Maximum Entropy in Action*, Oxford University Press, 1994, Edited by B. Buck, V.A. Macaulay.
4. G. J. Daniell, *Of maps and monkeys: an introduction to the maximum entropy method*, in *Maximum Entropy in Action*, Oxford University Press, 1994, Edited by B. Buck, V.A. Macaulay.
5. J. Skilling, *Fundamentals of MaxEnt in data analysis*, in *Maximum Entropy in Action*, Oxford University Press, 1994, Edited by B. Buck, V.A. Macaulay.
6. P. J. Hore, *Maximum entropy and nuclear magnetic resonance,* in *Maximum Entropy in Action*, Oxford University Press, 1994, Edited by B. Buck, V.A. Macaulay.
7. D. S. Stephenson, *Progress in NMR spectros.,* **20**, 515, (1988).
8. J. C. Lindon, A. G. Ferrige, *Progress in NMR spectros.,* **14**, 27, (1980).
9. D. D. Traficante, G. A. Nemeth, *J. Mag. Res.*, **71**, 237, (1987).
10. P. W. Atkins, *Physical Chemistry*, 4th ed. Oxford University Press, (1990).

11. S. F. Gull. *Developments in maximum entropy data analysis* in *Maximum entropy and Bayesian methods,* Cambridge, England, Kluwer, Dordrecht, Edited by J. Skilling, 53, (1989)

12. M.L. Boas, *Mathematical Methods in the Physical Sciences,* John Wiley &Sons, Inc., (1966).

13. J. Skilling, R. K. Bryan, *Maximum entropy image reconstruction: general algorithm* in *Monthly Notices of the Royal Astronomical Society*, **211**, 111, (1984).

14. E. T. Jaynes, *Monkeys, kangaroos and N* in *Maximum entropy and Bayesian methods in applied statistics: proceedings of the fourth maximum entropy workshop,* University of Calgary, Cambridge University Press, Edited by J. H. Justice, 26, 1984.

15. J. Skilling, *Foundations of maximum entropy*, in University of Cambridge Programme for industry, 1992.

16. J. Skilling, *A Massive Odyssey,* Private communication, 1996.

17. A. G. Ferrige, J. C. Lindon, *J. Magn. Reson.*, **31**, 337, (1978).

18. H. Anton, B. Kolman, B. Averbach, C. G. Delinger, *Mathematics with applications for the management, life and social sciences*, 3rd Edition, Harcourt Brace Jovanovich, 1988.

CHAPTER 3: PRACTICAL METHODS

This chapter describes how the Maximum Entropy techniques employed in this work are used in practice. The NMR or mass spectrometry methods will be discussed as part of the appropriate chapters. Section 3.1 includes a description of the Maximum Entropy hardware and software and describes the preliminary data processing required to change the experimental data into a form appropriate for analysis. In section 3.2, the basic operation of the Memsys and Massive Inference software is discussed and the link between the software inputs and the appropriate theory is highlighted. Standard methods which have been used in this work for optimizing the software inputs, e.g. determination of the optimum point spread function, will be demonstrated. The output from the two algorithms will be demonstrated.

3.1: Hardware, software and pre-processing

All the Memsys and Massive inference data processing described in this thesis has been carried out using proprietary software developed and sold by MaxEnt Solutions Ltd.[1]. The software package comprises:

Memsys5 Kernel, Spec5 v.2.25, 1995

MassInf Kernel, Deconvolve v1.10, 1996

Electrospray kernel, Spray5 v2.21, 1995

MaxInt2 graphics interface, v2.46, 1994

Various plotting and utility programs.

This runs on a computer system comprising a 90 MHz Pentium personal computer, with 32 MB RAM and a 1 GB disc drive. The computer uses a Sun operating system, Solaris v2.4 for X86. The total cost of this package was ca. £25,000 and is considered to be beyond the cost of most

academic / small industrial NMR groups unless a particular application has been identified. This may explain the limited exposure the techniques have received in the literature for the applications described in this work.

Both Memsys and MassInf can accept data from any spectroscopic technique provided the following minimum requirements are met:

- the experimental spectrum must be real, with no imaginary components from Fourier transforms. Ideally, the data should be correctly phased.
- the spectrum must be unfiltered, i.e. not convolved with other lineshapes such as exponential line broadening.
- the peak widths present in a spectrum should not vary. The half-height peak widths should, ideally, be within a range of 50% of each other for the optimum results.
- the best deconvolutions will be achieved if the digital resolution is sufficient to give a minimum of four data intervals at the half-height-width of peaks.
- the baseline should be uniformly spread about zero.

For most NMR spectra the above criteria can be easily met. The main limitation is that of peak-width variability. This limitation will be discussed in more detail in Chapter 6, where the combination of linear prediction and Maximum Entropy data processing is used to derive additional information from the ^{27}Al NMR spectra of aluminium-chlorohydrate systems. The practical implications of the above are that a FID must be Fourier transformed without the application of window functions. The subsequent spectrum should be correctly phased and baseline corrected. The real part of the spectrum is then converted into ASCII as two columns, one of intensity and the second the corresponding data-channel number. The Maximum Entropy software does not recognize the δ chemical shift scale. The ASCII file is then transferred to the PC and converted into a binary format that can be read by the program kernels. This use of a data-channel scale and the

ASCII format makes the software general and removes the need for third-party data conversion programs. It relies on the instrument manufacturers to provide routines for conversion to ASCII.

The above only describes how to get a spectrum into a format ready for further processing. It does not attempt to establish if Maximum Entropy data processing is appropriate, or if the desired information can be extracted from the spectrum. Ferrige et al.[2] have developed an empirical relationship, the deconvolution criterion (**D**), which enables the analyst to assess if the required information can be extracted from a spectrum using deconvolution techniques. The deconvolution criterion is given by:

$$\mathbf{D} = \mathbf{P}^2 . (\mathbf{W.S})^{0.5}$$

where **W** = half-height peak width in data intervals

S = S/N of weakest peak of interest (measured peak to peak)

P = separation between peaks measured in units of peak width at half -height.

For **D** > **2.5** good separation is expected in the Maximum Entropy result.

For **D** < **2.5** poor separation is expected.

Ferrige[3] has noted that the error on **D** is about 1.2. Accordingly, when **D** is greater than 4 peak separation is likely to be baseline resolved. Similarly, when **D** is less than about 1.3 it is most unlikely that any degree of peak separation can be achieved. Furthermore, by fixing **D**, the smallest peak separation that can be separated by a deconvolution technique can be estimated from values of **W** and **S**. If the required degree of separation cannot be obtained, the data must be improved by re-designing the experiment such that **W** or **S** is increased. Ferrige continues that the deconvolution criterion provides a good estimate of data quality for virtually all situations. However, the empirical relationship breaks down under the combination of very high **S** and **W** and a small value for **P.**

Where appropriate in the following chapters, the use of the deconvolution criterion will be demonstrated. However, it is worthwhile illustrating the usefulness of the relationship for a known system. The selected spectrum was a ^{1}H NMR spectrum of glucose penta-acetate. This system proved useful because the five acetyl peaks are equally intense, which makes assessment of a successful deconvolution more straightforward. The system could also be used to validate the software for GLP compliance in line with the policy of Unilever Research, Port Sunlight Laboratory.[4]. All the research presented in this thesis was conducted within a GLP environment. It is worth noting, at this point, that the software has never been subjected to modern quality assurance standards, e.g. ISO 9000, and cannot be validated *per se*.

Approximately 10mg of β-D-glucose penta-acetate were dissolved in approximately 1 cm^3 of deuterochloroform. The subsequent ^{1}H spectrum, Figure 3.1, was acquired on a Bruker AM360 spectrometer fitted with a four-nucleus QNP probe using a pulse with a 30° flip-angle and an inter-

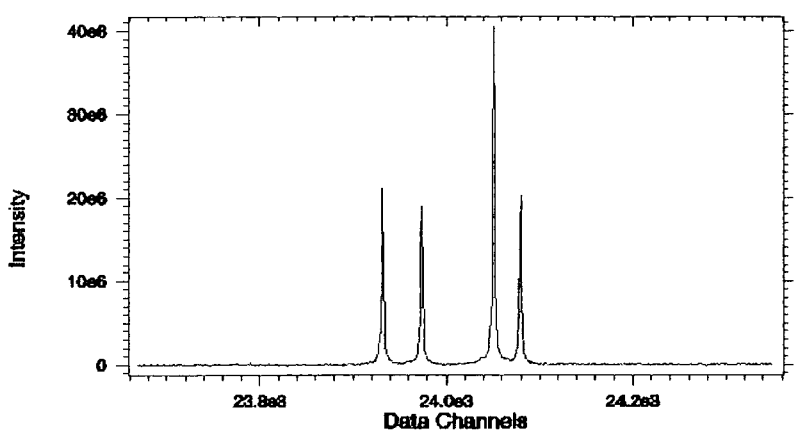


Fig. 3.1 ^{1}H NMR spectrum of glucose penta-acetate.

pulse delay of 20 seconds, i.e. sufficient for full spin-lattice relaxation.

The two peaks to be separated (24.05 x 10^3 data channels) are, to the naked eye, coincident. The value of **P** cannot be easily measured. Furthermore, the digital resolution of the spectrum is limited, with only 2.5 data channels at half-height peak width. The S/N ratio of the peaks has been measured at approximately 420:1. Assuming **P** to have a value of approximately 0.1, i.e. almost coincident, the deconvolution criterion gives a value of **D** of 0.3. This is considerably less than the value of 2.5 required for a successful deconvolution.

Applying the Memsys algorithm to this spectrum produces the results shown in Table 3.2. The algorithm has failed to separate the two overlapped peaks. The deconvolution criterion could have saved considerable time and effort by indicating that the required peak separation was not possible. A Lorentzian PSF of width 3.2 data channels was used.

Table 3.2 Memsys: Table of Peak Areas for Glucose Penta-acetate

Peak Position / data channels x 10^3	Cumulant*	Error*
23.93	81.295	0.456
23.97	82.135	0.555
24.045	162.313	0.695
24.07	83.496	0.712

*Expressed as a percentage of the total spectral intensity

This system will also be used to demonstrate the standard methods that have been used in this work for optimizing the software inputs, e.g. determination of the optimum point spread function. If the glucose penta-acetate solution is mixed with deuterobenzene ($CDCl_3 : C_6D_6$ = 2:1 (v/v)) a partial separation of the two overlapped peaks can be achieved. (See Fig.3.3).

Applying the deconvolution criterion to this system gives a value of **D** of approximately 32, assuming a peak separation equivalent to one peak width. This is well in excess of the 2.5 that Ferrige indicates is required for a good separation. The Maximum Entropy deconvolution is expected to achieve baseline resolution for this system. This is easily achieved and, as shown in Table 3.4, the error bars associated with overlapped peaks are slightly larger than those on the

discrete peaks. Even the area of the peak at 24.36 data channels carries a slightly larger error than is expected for a discrete peak. This is due to the small degree of overlap with the other peaks. This is intuitively correct, but the errors are still below the 1% level, which is indicative of a very successful deconvolution. The size of these errors is also considerably smaller than would be expected for the normal integration methods associated with NMR spectroscopy.

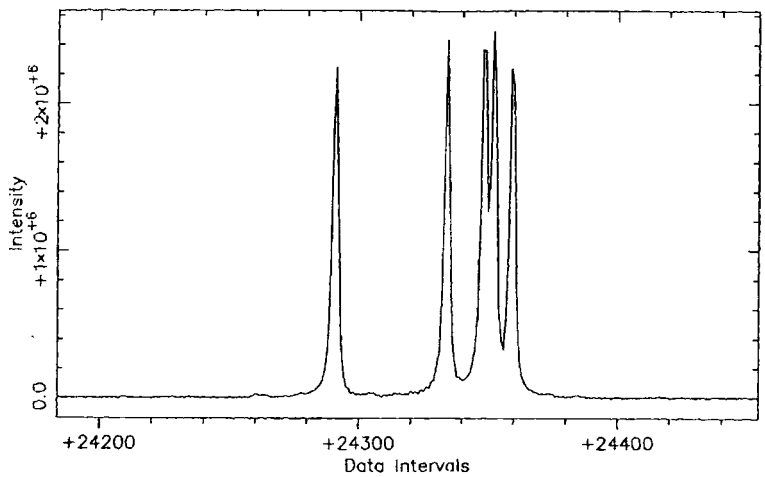


Fig. 3.3 Glucose penta-acetate in $CDCl_3$: C_6D_6 2:1 (v/v)

Table 3.4 Memsys: Table of Peak Areas for Glucose Penta-acetate

Peak Position / data channels x 10^3	Cumulant[1]	Error[1]
24.29	18.89	0.19
24.33	19.52	0.20
24.352	19.25	0.24
24.353	19.31	0.25
24.36	19.75	0.26

[1]Expressed as a percentage of the total spectral intensity
[2]The position of these peaks is quoted to 5 significant figures to aid assignment

It is clear from the above examples that the deconvolution criterion can provide a useful decision-making tool for estimating whether or not a successful deconvolution can be achieved. It is not limited to Maximum Entropy methods but can be applied before any deconvolution technique.

Once it has been decided that a successful deconvolution can be achieved, there are two key inputs to both the Memsys and Massive Inference algorithms, i.e. point spread function and sigma. These program inputs will now be discussed.

3.2: Operation of Software

Once the spectrum to be processed is in the correct form and the Deconvolution Criterion suggests that the desired peak separation can be achieved, a number of measurements have to be taken from the spectrum before the algorithms can be started.

3.2.1 Point Spread Function

The point spread function (PSF) is a key input to both the Memsys and MassInf algorithms if a successful deconvolution is to be achieved. The PSF is an estimate of the underlying peak width and shape. There are a number of different methods for determining the optimum PSF parameters.[2,3] All the methods have been developed by MSL Ltd.[1] Only those methods employed in this work will be described here.

3.2.1.1 Evidence matrix: Design of a parametric PSF

If the desired PSF can be adequately modeled in terms of a mixture of a Lorentzian and Gaussian peak-shape, and the level of noise in the spectrum can be measured, an evidence matrix is one of the most useful methods for determining the PSF parameters. This process uses either algorithm to refine the PSF parameters based on a number of trial deconvolutions. It can be a laborious process. For a Gaussian peak the program input, wing, is set to a value of zero and for

Lorentzian peak a value of one is used. Mixed Gaussian / Lorentzian peak shapes can be achieved by adjusting the value of the wing parameter.

PSF optimisation is achieved by recording a series of trial deconvolutions for, say, a range of peak widths, whilst keeping the PSF shape parameters (wing) constant. By observing the algorithm's evidence diagnostic (see Chapter 2) it is possible to construct an evidence matrix of left width against right width for a parametric curve, providing the estimate of the noise input to the algorithm is kept constant throughout any one series of trials. The more positive the evidence value, the greater the likelihood that the chosen PSF parameters fit the data. Table 3.5 gives the width evidence matrix for the glucose penta-acetate spectrum shown in Figure 3.1. The columns represent the change in evidence for the PSF right width for the range 0.75-1.2 data channels in increments of 0.05 channels, and the rows the change in evidence for PSF left width for the same range.

$$\text{left} = \begin{bmatrix} -4266.10 & -4266.83 & -4266.17 & -4265.23 & -4264.23 & -4262.73 & -4261.18 & -4259.92 & -4260.87 & -4263.96 \\ -4267.13 & -4260.12 & -4260.75 & -4260.49 & -4259.80 & -4258.92 & -4257.46 & -4256.47 & -4258.11 & -4260.84 \\ -4267.26 & -4261.39 & -4254.79 & -4255.82 & -4255.54 & -4254.68 & -4253.48 & -4253.05 & -4255.02 & -4257.11 \\ -4267.19 & -4261.78 & -4256.16 & -4249.12 & -4250.42 & -4249.68 & -4248.80 & -4249.23 & -4251.24 & -4254.53 \\ -4267.40 & -4261.97 & -4256.67 & -4251.04 & -4244.33 & -4245.31 & -4244.86 & -4246.45 & -4248.68 & -4251.70 \\ -4267.64 & -4262.45 & -4256.97 & -4251.39 & -4245.87 & -4240.78 & -4241.91 & -4243.44 & -4246.17 & -4249.88 \\ -4267.88 & -4262.50 & -4257.05 & -4251.91 & -4246.82 & -4242.01 & -4238.08 & -4241.22 & -4244.72 & -4249.07 \\ -4268.53 & -4263.10 & -4258.07 & -4253.73 & -4248.96 & -4244.11 & -4241.11 & -4241.09 & -4245.26 & -4250.13 \\ -4270.53 & -4265.55 & -4261.18 & -4256.80 & -4252.27 & -4247.83 & -4244.53 & -4245.09 & -4247.87 & -4245.13 \\ -4274.54 & -4269.96 & -4265.43 & -4261.43 & -4257.34 & -4253.64 & -4250.67 & -4249.86 & -4252.82 & -4258.26 \end{bmatrix}$$

Table 3.5. Glucose penta-acetate: Width evidence matrix for a purely Lorentzian Line

The above matrix can be better represented as a contour plot of left width against right width in which the optimum PSF parameters become more obvious to the eye. For this system the optimum PSF width parameters indicate a symmetrical line of half-width 1.05 data channels. The single maximum value indicates that there is only one peak width present in the data. If a range of peak widths were present this would be apparent by a number of local maxima in the evidence values. The system can only be described by one PSF. For a spectrum with different peak widths,

the evidence values will indicate a maximum at the position corresponding with the average peak width.

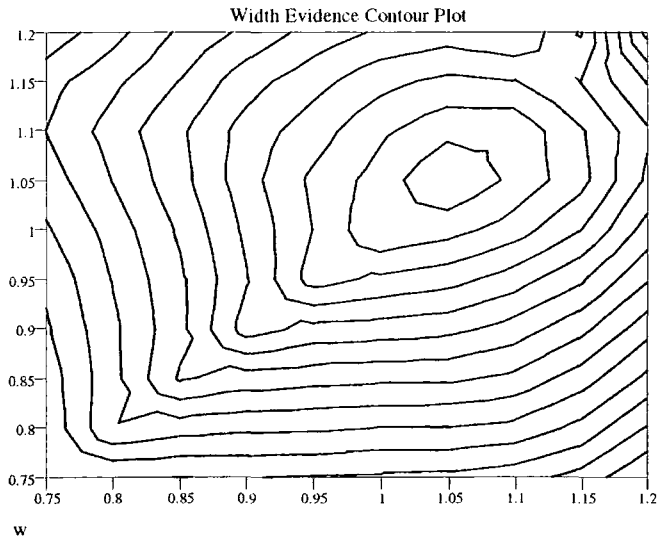


Fig. 3.6 PSF width contour plot with left width represented on the horizontal axis and right width on the vertical axis in units of data channels.

In principle, some knowledge of both peak width and shape can be obtained from such evidence plots. However, in practice, the evidence values for optimizing the peak-shape parameters tend to be much less informative than the corresponding width trials. For example, in the case of a noisy spectrum, where the peaks are known to be Gaussian, an input width that is in error by a few percent is sufficient to prevent the input PSF from fitting the data. There will also be a corresponding reduction in the evidence value. However, within the noise level, virtually any shape can be used and will fit the data adequately, resulting in little change in the evidence diagnostic. The input PSF shape parameters only become important when the signal:noise ratio is high.

There are two exceptions to the above observations. Firstly, for a particularly noisy spectrum it is possible to obtain an enlarged evidence value for an unreasonably narrow PSF. This corresponds to the optimum PSF for the noise frequencies and should not be confused with the optimum PSF for the genuine peaks. This may become a problem for solution-state ^{13}C NMR spectra where the peak widths can be very similar to the noise frequencies. Secondly for peaks

which are severely overlapped, if, within the freedom provided by the noise, there is greater evidence for fewer peaks than the number known to be actually present, there will be a greater evidence value corresponding to an unreasonably wide PSF. With these caveats in mind, it is possible to use the evidence diagnostic to accurately design parametric PSFs for an unknown spectrum.

3.2.1.2 Sigma Profile: If the noise level is in doubt

This is the most general method for determining the optimum PSF because it also determines the optimum value for sigma (noise level : see chapter 2). If the measured noise level in a spectrum is unreliable, through, for example, the application of a window function, the estimate of sigma will cause an error in the determination of the PSF parameters. In these circumstances, additional trial runs are necessary in which the value of sigma is varied along with the PSF parameters of width and shape. For each sigma there will be an optimum PSF with the greatest evidence value. A plot of PSF width against input sigma produces a curve with one or more points of inflection. At these points a relatively large change in sigma corresponds to a small change in PSF width. The optimum width and sigma are measured from the point with the minimum slope.

This method of PSF optimization has been used for the styrene / maleic anhydride spectrum described in Chapter 5.

3.2.2 Other Program Inputs

Apart from the PSF parameters of left width, right width, left shape and right shape, there are a number of other inputs available. These are well described in the Memsys manuals but the most useful are listed below:

- LEVEL. This controls the extent of the output diagnostic information (see section 3.2.3.1). When the default value of ten is used, diagnostic information is output at the end of each iteration. With a value of zero, only the final convergence criteria are output at the end of a deconvolution.

- MOVIES. A positive value determines the number of samples that are taken from the probability cloud of plausible results. These samples are used to calculate the error bars on any spectral feature. The default is fifteen.
- RATE. This parameter controls the rate of change between each step of the iteration. If rate is set to too high a value, the algorithm will fail to follow the Memsys trajectory and may not converge. The default setting is one.
- SIGMA. This value is measured directly from the spectrum and, as shown above, is used in PSF optimization. The algorithms are provided with graphical facilities for measuring the noise level in a spectrum
- UDEF and UTOL. For spectra with a very high S/N ratio and severely overlapped peaks the algorithms may fail to converge for a given PSF. One method for overcoming this difficulty is to force the algorithm to follow the Memsys directory in smaller steps. This can be achieved by relaxing the values of UDEF and UTOL. The default values are each 0.1.
- AIM. The default value of aim is 1.0. Setting this parameter to a smaller value forces the algorithm to proceed further down the trajectory towards fitting the data more closely.

3.2.3 Program Output

The primary output is a table of peak position against peak intensity, each presented with the appropriate standard errors determined from the movie samples. The program is also capable of generating a range of different graphical outputs as an aid to visual interpretation of the deconvolution results. The following are examples of the different types of output available.

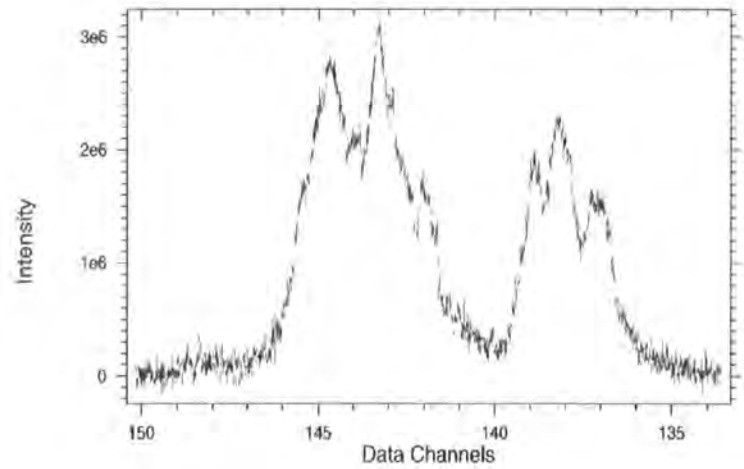


Fig.3.7 A Typical unprocessed spectrum

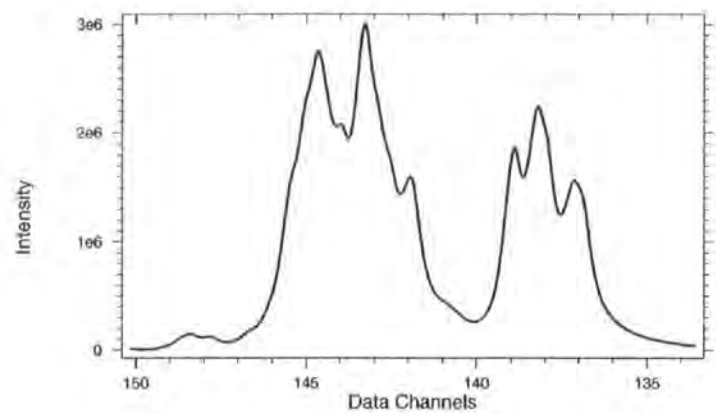


Fig.3.8 Mock Data

The mock data are determined during the final Memsys iteration and are the convolution of the Memsys result and the applied PSF, i.e. the algorithm's reconstruction of the data, and are essentially noise free. When the applied PSF accurately fits the raw data peak-width and peak-shape, the mock data will be identical with the raw data apart from the noise. The mock data are useful for assessing the quality of the MaxEnt result, since for an ideal fit the difference between the mock data and the raw data should give residuals within the noise. This method for de-noising a

spectrum will be utilized in Chapter 4 in the analysis of the NMR spectrum for sodium carboxy methyl cellulose.

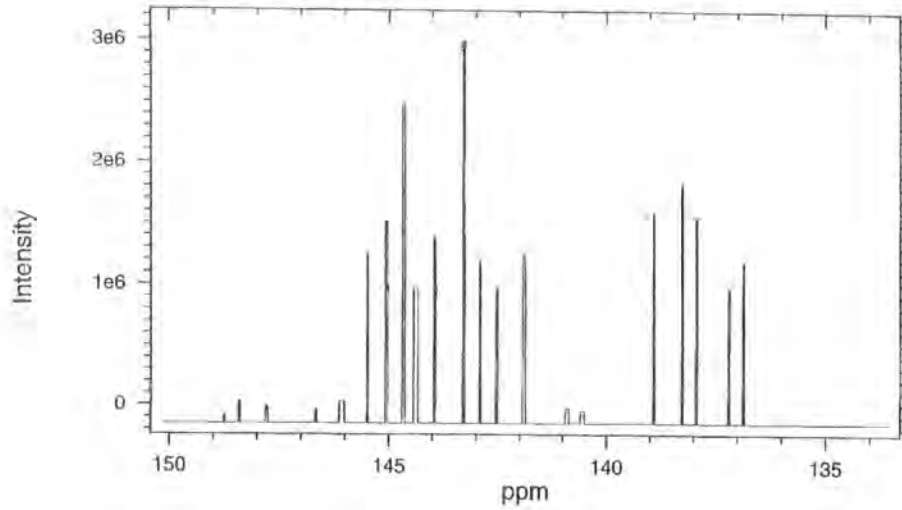


Fig. 3.9 Spike Plot with Errors

For the 'spike plot with errors' the program has condensed the total intensity of the peaks found in the Memsys result into their median positions. In these plots the absolute peak areas are directly proportional to the spike heights. The user may observe the true intensity ratios directly from the plot. The width of the spikes indicates the assigned positional errors as reported in the table of quantification. This facility is particularly useful for examining NMR spectra to establish if multiplet peak separations are the same within the standard error.

The MassInf algorithm produces the same graphical output as Memsys. However, because MassInf uses the full probability distribution, which is usually sampled fifteen times to produce the movies samples, it is possible to display each movie sample. (See Fig.3.10).

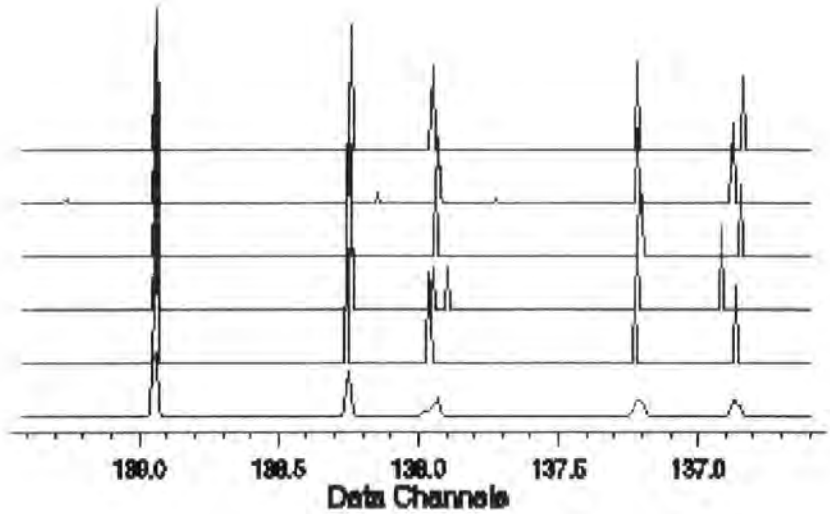


Fig. 3.10 Typical MassInf output for 5 movie samples

As can be seen in Fig.3.10, each of the five movie samples has evidence for peaks of slightly different intensity and position. The lower trace represents the arithmetic mean of all the movie samples.

3.2.3.1 Diagnostic Information

As well as the graphical information presented above, the algorithm also outputs an array of diagnostic information that describes how well the algorithm is proceeding with a deconvolution. The most important of these are:

- TEST. This is (1-cosθ), where θ is the angle between the gradients of entropy and χ^2. Test is 0 on the Memsys trajectory and less than 1 whenever the angle is acute. (See section 2.1.3).
- CHISQ. This is the normalised chi-squared test of how well the Memsys result fits the experimental spectrum. It assumes Gaussian statistics.
- OMEGA. This is the algorithm's stopping criterion for reaching convergence. At covergence omega = 1.

REFERENCES

1. MaxEnt Solutions Ltd., 9 Church Street, Isleham, Ely, Cambridge CB7 5QS.

2. *A users's manual for running 1-D MaxEnt*, MaxEnt Solutions Ltd., 1994.

3. *MassInf user manual, 1-D Deconvolution & Electrospray,* MaxEnt Solutions Ltd., 1996.

4. OECD Series on Principles of GLP and Compliance Monitoring, 1999.

CHAPTER 4: THE APPLICATION OF MAXIMUM ENTROPY DATA PROCESSING TO SPECTRAL DE-NOISING

SODIUM CARBOXYMETHYL CELLULOSE

Sodium carboxymethyl cellulose (SCMC) is a chemically modified natural polymer and is the most widely used water soluble derivative of cellulose. Its applications include use as an anti-redeposition agent and an emulsifier in the detergent, food and textile industries.

SCMC is made by the reaction of cellulose with monochloroacetatic acid. The resulting linkage of the carboxymethyl (- CH_2COO^-) groups with the free hydroxyl functions of the cellulose is used to achieve water solubility. This is a major benefit for the above industries[1]. For industrial applications, the stoichiometry of the reaction is normally chosen such that the cellulose does not react completely, i.e. not all the hydroxyl groups on the anhydroglucose ring undergo carboxymethyl substitution. Typically, this results in a product with an average degree of substitution (ds) in the range 0.4 - 1.3 (ds attains the value of 3 for complete reaction of the hydroxyls). The three possible sites for substitution are shown in Figure 4.1.

By limiting the reactivity of the available hydroxyl groups the resulting polymer can consist of up to eight different monomers:

- one unsubstituted glucose residue
- three monosubstituted glucose residues, i.e. 2-, 3-, and 6- carboxymethyl glucose
- three disubstituted glucose residues, i.e. 2,3-, 2,6-, and 3,6- carboxymethyl glucose

- one trisubstituted glucose residue, i.e. 2,3,6- carboxymethyl glucose.

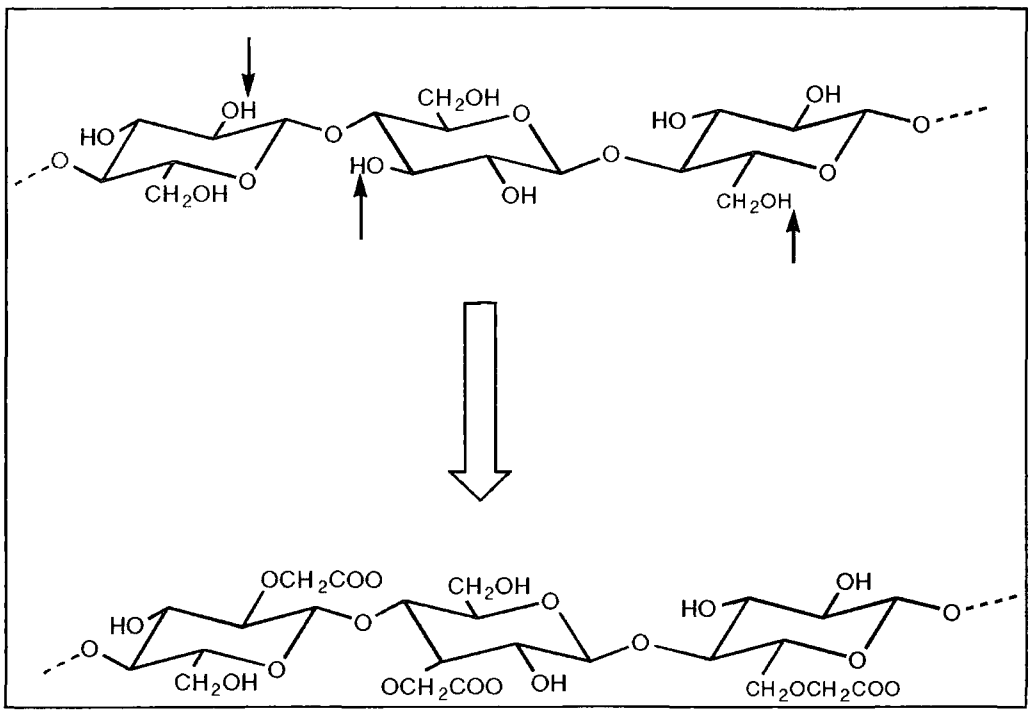


Figure 4.1 Section from cellulose molecule showing the three different hydroxyl groups before and after substitution. (ds = 1.0).

Figure 4.2 shows one of the eight possible monomers, the trisubstituted SCMC, and indicates the carbon numbering that will be adopted in this thesis. Taking into account the α and β anomeric glucose units the situation is complicated further; there is the possibility of sixteen different monomers.

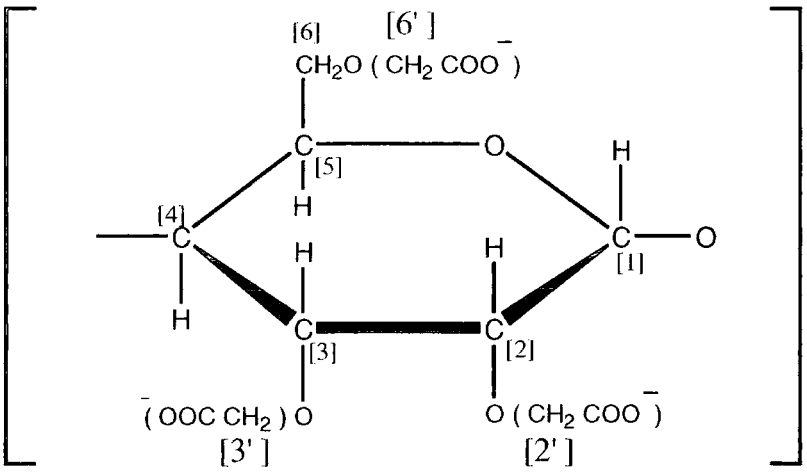


Figure 4.2 2,3,6-carboxymethyl glucose.

The NMR spectra of SCMCs can consist of a large number of resonances which are usually broad, prone to severe overlap and very poor signal : noise ratio. The situation is further complicated because the industrial grade raw material is normally less than 70 % pure SCMC. The remainder consists of water, various inorganic salts, glycollic acid derivatives and unreacted cellulose fibre.[2]

Section 4.1, describes the characterization of SCMC as presented in the literature with supporting NMR spectra acquired during the course of this research. The aim of this review is to illustrate the limitations of the NMR technique for these systems as a point of comparison with the Maximum Entropy processed spectra which are shown in section 4.2.

4.1 SCMC Characterization : No data processing

As noted by Chaudhari et al.[3] SCMCs have remained poorly characterized materials and Baar et al.[1] comment that the chemical characterization of CMC has been limited to determination of the average ds. Traditional analysis has relied upon classical chemical methods for determination of ds. These have included conductometry[15], gravimetry[16], and colorimetry[17]. More recently, ^{13}C NMR spectroscopy has been used to characterize cellulose ethers[14] and the use of this technique to study the microstructure of carboxymethyl derivatives has been reported [8,14].

The industrial characterization of SCMC has been limited to the determination of some non-specific relationship between ds and improved product performance. In the detergents industry, if a direct correlation could be established between, e.g. ds and improved soil anti-redeposition, this was normally sufficient to ensure the 'correct grade' of SCMC was used. More refined measurements are needed if industry is to tailor SCMCs to meet specific product requirements. For example, it may be that substitution at only one position, say C-6, is required in order to achieve an improvement in polymer functionality and this structural characteristic would not necessarily manifest itself in the measurement of a mean ds.

Information about microstructure can also give important insights into polymer functionality. For example, are all the di-substituted glucose residues together in a block ? Baar et al. have reported a ^{13}C NMR method involving sample degradation and spectral deconvolution techniques for determining the distribution of substituents over the three possible positions [1].

The aim of the work presented in this thesis is to establish if more information could be extracted from the NMR spectra of a variety of SCMCs. Samples from different manufacturers, with different molecular masses and different degrees of substitution have been analysed. Whilst this would appear to be a modest aim, the level of information that can be extracted directly from a typical spectrum is small.

Typical fabric washing powder formulations indicate SCMC levels of only ~0.5wt% (quoted as 100% SCMC). However, SCMC raw materials contain significant levels of impurity, e.g. glycollates, so the level of 'active' SCMC measured is likely to be lower than the 0.5% quoted. Furthermore losses, which may be associated with any pre-concentration stage, e.g. ultrafiltration, result in the level of SCMC measured in the NMR experiment being lower than that actually present in the powder.

Given the above, it is difficult to determine accurate SCMC levels in a fully formulated product, although approximate values tend to suffice for competitor product screening, and so a spectral fingerprinting method is likely to be of most use for these systems.

4.1.1 NMR Spectroscopy Of SCMC: Sample degradation

The NMR spectrum of polymeric SCMC usually suffers from very poor signal : noise ratio due to the limited solubility of the SCMC and the high viscosity of the resulting polymer solutions. The spectrum is further complicated because the natural NMR line-width of polymers is large and the spectral bands overlap considerably, yielding little quantitative information.

The difference between the NMR line-width of polymers and those of low molecular weight molecules is due to the T_2 contribution to the line-width. T_2 is the time constant that describes the decay of transverse magnetization as a result of a loss of phase coherence between the nuclear spins. The T_2 value can provide information about the distribution of resonant frequencies and about the local fields experienced by the magnetic moments of the nuclei. The local fields are related to the structure and nature of the local magnetic environment around the nucleus.

If molecular tumbling is fast compared with the Larmor frequency the proportion of molecules tumbling at the resonant frequency is low, i.e. a low spectral density function at the resonant frequency, and so relaxation is not particularly efficient. This results in long T_2 values and hence narrow resonant lines. This is the case for the SCMC monomers.

In highly viscous polymer solutions molecular tumbling is slow compared with the resonant frequencies, i.e. they have a large spectral density function at the precessional frequency. Relaxation is efficient resulting in a rapid loss of phase coherence between the spins. Hence, the T_2 values normally observed in polymers are very short and the resonant lines broad. As the chemical shift difference between most of the bands observed from SCMC is very small the resultant spectrum is severely overlapped.

One method of overcoming peak overlap in polymer systems is molecular degradation. The high molecular weight polymer can be degraded either completely to monomeric units or partially to small oligomers. Whilst the subsequent spectra are often better resolved, due to the sharper spectral lines, interpretation of these spectra does not describe the polymer system as used in practice. There is often less information available about microstructure and position of substitution.

The literature[1,4] describes three methods as being suitable for SCMC degradation:

- acidic hydrolysis
- sonication
- enzymatic degradation

As described by Adams et al.[5] other methods of degradation are generally avoided. These include:

- oxidative attack. This can result in a variety of reactions other than chain scission; generally oxidizing agents are unspecific in their action upon cellulose and consequently they lead to a variety of products. Given the already complex nature of the SCMC spectrum this type of degradation is not suitable for characterizing SCMCs.
- alkaline degradation. Under rigorous conditions, i.e. normal alkali at 170^0C, alkaline hydrolysis of the glycoside linkages results in random chain scission.[18] These conditions do not lend themselves to easy sample preparation for NMR analysis and are considered to be unsuitable for the work associated with this degree.

4.1.1.1 Sample Preparation

- **Acid Hydrolysis**

For the work associated with this research typical hydrolysis conditions are: ca.20 mg SCMC added to 1 cm^3 solution of 20% DCl in D_2O and heated for 1hour 15minutes at 70^0C.

Acid hydrolysis of cellulose proceeds by random scission of the glycoside linkages until the monomer, glucose, is produced. As described in the introduction to this chapter, it is possible to generate eight different substituted monomers following acid hydrolysis and, taking into account the α and β anomeric glucose units, sixteen different anomeric glucose units are possible. Whilst Ho et al.[6] describe a method for determining both the average ds and the order of reactivity of the hydroxyls in cellulose, Baar comments that the sixteen different units cannot be completely resolved by ^{1}H NMR[1]. Furthermore, following acid hydrolysis to the monomers it is not possible to determine the distribution of substituents on a macromolecular scale or indeed derive any microstructural information.

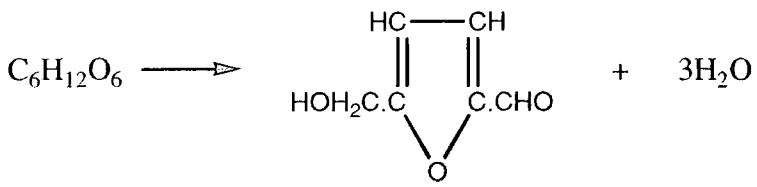
- **^{1}H NMR Of Acid Hydrolyzed SCMC**

A ^{1}H NMR spectrum of an acid hydrolyzed SCMC ex Aldrich Chemical Co. (ref.: 41,927-3) is shown in Figure 4.3, with expansions shown in Figure 4.4 and Figure 4.5. According to Aldrich, the sample has a weight average molecular mass of 90 kDa and a ds of 0.7. This spectrum was acquired on a Bruker DRX500 spectrometer following acidic hydrolysis of the SCMC as described above. The spectrum was acquired with a 30 degree pulse of 4 μs duration and an interpulse delay of 5 s. The probe temperature was 27 °C.

The molecular mass of this sample is relatively low compared to other commercial SCMC samples (typical values for a commercial SCMC are ~200 kDa). Apart from intensity differences and any impurities present the spectrum of the acid hydrolysed SCMC is independent of the starting material, i.e. glucose is always produced. Nevertheless, this low molecular weight SCMC was chosen as a standard because it was hoped that this sample could later be used to produce a ^{13}C spectrum free from severe band overlap. Unfortunately, as will be shown later in this chapter, the main spectral bands are still clearly overlapped even at 125 MHz for ^{13}C and spectral deconvolution is required.

Samples with a high ds value may require a longer period of heating than indicated above, although care should be taken to avoid decomposition of the glucose monomers. Decomposition is

known to accompany acid hydrolysis[7] and may change the spectral intensities sufficiently to produce errors in a calculation of ds. One of the products of the decomposition gives rise to doublets near 7.0 ppm and 7.8 ppm. These doublets are due to hydroxymethylfurfural which is formed as an intermediate during the acid decomposition of glucose to laevulic acid.[12]



Hydroxymethylfurfural

These bands provide a useful internal control to establish if the hydrolysis has proceeded too far. If they are particularly intense, the hydrolysis should be repeated using a lower temperature and/or shorter hydrolysis time.

According to Clemett and Wright[2], apart from decomposition there are a number of other possible sources of error in calculating the average ds from the ^1H spectrum of the acid hydrolyzed monomers.

Firstly, specific hydrogen / deuterium exchange may occur as a result of using DCl for the hydrolysis. If this occurs preferentially at one site, the spectral intensities would be distorted, and the calculation of average ds would be in error. Clemett notes that if the mild hydrolysis conditions described above are used, the effect of isotopic exchange on the calculation of average ds is likely to be small.

Secondly, interfering species such as glycollates may be observed in the spectral region of interest. Clemett comments that sodium glycollate and disodium diglycollate are present in varying amounts in most samples of crude SCMC. Sodium ethoxyglycollate is present in the product of some manufacturers only, suggesting that ethanol may be used in the manufacturing process.

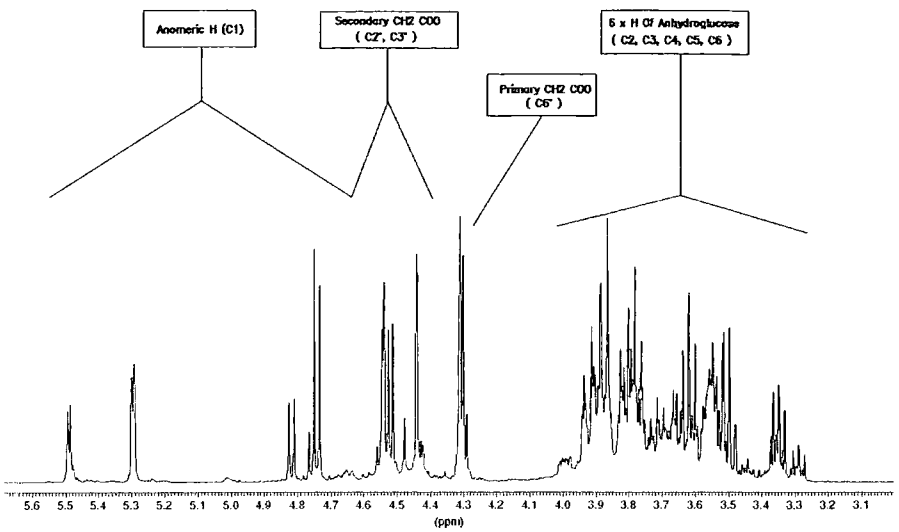


Figure 4.3 ^{1}H NMR spectrum of acid hydrolyzed SCMC

$HO\ CH_2\ COONa$	$O\ (CH_2\ COONa)$	$CH_3\ CH_2\ O\ CH_2\ COONa$
Sodium Glycollate	Disodium diglycollate	Ethoxyglycollate

Unfortunately, the ^{1}H NMR resonance of glycollic acid, derived from sodium glycollate, is almost coincident with that of the primary substituted carboxymethyl group and severely affects the calculation of ds. This problem can be overcome by either further sample preparation involving a methanol wash to remove the glycollates or by measuring the ds based on the intensity of the secondary alcohols, which do not suffer from interference.

Figure 4.4 is an expansion of the spectral region 4.6 - 5.6ppm. The progress of the hydrolysis can be monitored by observing an increase in the intensity of the NMR signals from protons at the reducing end [C1] of the degraded sugar, i.e. the anomeric protons. This group of

signals arises as a result of the cleavage of the glycoside linkage in cellulose, and is observed as two sets of doublets.

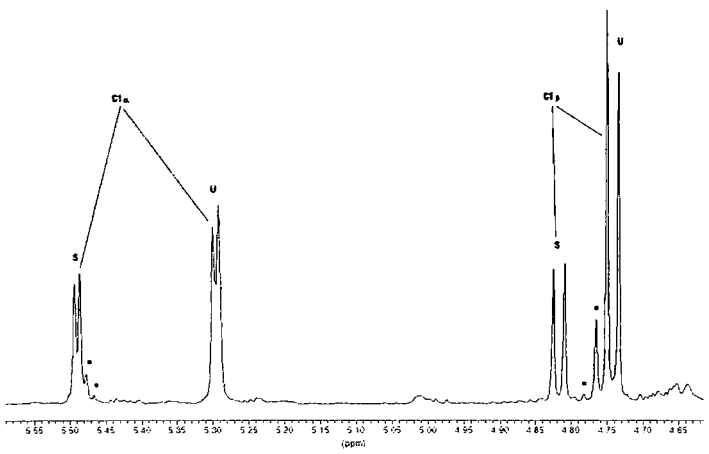


Figure 4.4 C1 protons of hydrolyzed SCMC

The doublet splitting is caused by coupling with the single proton at C2. Ho has assigned the set of high frequency doublets to the C1 proton of the α-anomer ($J_{axial\text{-}equitorial}$ = 3.8Hz), while the set of low frequency doublets is due to the proton at C1 of the β-anomer ($J_{axial\text{-}axial}$ = 7.9Hz). The S and U identifications of the doublets refer to substituted and unsubstituted hydroxyl group at C2. There are also other unresolved minor bands, possibly more doublets (marked with an asterisk), which have not been identified in the literature, but may be due to long-range coupling.

Based on the identified bands, it is possible to calculate the relative amounts of:

1. α-anomer with the C2 hydroxyl substituted
2. α-anomer with the C2 hydroxyl unsubstituted
3. β-anomer with the C2 hydroxyl substituted
4. β-anomer with the C2 hydroxyl unsubstituted.

If absolute amounts are required, an internal intensity standard such as Analar Acetic Acid must be incorporated into the system. The quantification may be compromised by the overlapping unidentified resonances; some form of deconvolution would be required for the most accurate values. There are other possible sources of error.

Firstly, the different species may have different T_1 s and so the experimental conditions need to be set for quantitative analysis. T_1 is the time constant that describes the spin - lattice relaxation of the nuclear spins. [19]

Secondly, care must be taken to ensure that there is no preferential hydrolysis of any one moiety. This can be checked by comparing the intensity of the bands in the spectral region 4.6 - 5.6ppm (the anomeric protons) with the intensity of the bands in the region 3.2 - 4.1ppm (the other CH protons of the glucose unit). For complete, quantitative hydrolysis the intensities should be in the ratio of 1:6.

It is also worth noting that as the ds increases the intensity of the C1αS proton increases at the expense of the C1αU intensity.

For samples free from glycollate impurities the ds can be calculated from the following[2]:

$$ds = (6B-A) / 2A \qquad \text{.....}\{1\}$$

where A is the intensity of the region 3.0 - 4.2 ppm, corresponding to the 6 protons of the anhydroglucose unit, and B is the intensity of the region 4.2 - 5.5 ppm, corresponding to the anomeric protons.

For samples of SCMC contaminated with glycollates the 'total ds' is obtained via the secondary hydroxyl ds, i.e.

$$ds\ (secondary) = C/2 \times 6/A \qquad \text{.....}\{2\}$$

where C is the intensity of the region 4.4 - 4.6 ppm and A is as before. Allowance for the contribution of ethoxyglycollate to A must be made if it is present. Clemett then shows a plot of total ds versus secondary ds for a number of SCMC samples free from glycollates. These data fit a straight line described by equation {3} with a coefficient of correlation of 0.992.

$$ds\ (total) = 0.1 + 1.38ds\ (secondary) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \{3\}$$

Hence, if the secondary ds can be measured, the total ds can easily be found.

For the Aldrich SCMC sample the ds has been calculated as 0.84 using equation {1} and standard integration techniques. This value is higher than that quoted by Aldrich.

As stated earlier, Baar comments that the 16 different α and β anomeric glucose units cannot be separately resolved by ^{1}H NMR Spectroscopy. However, Ho et al[6] describe a ^{1}H NMR method for determining the distribution of the substituents over the positions C2, C3 and C6 following complete hydrolysis. Furthermore, from the peak intensities in the ^{1}H spectrum, information about the relative reactivity of the three hydroxyl groups in the anhydroglucose unit can be deduced.

Ho calculates the distribution of substituents from the spectral region 4.0 - 4.5ppm. He has assigned the intense peaks in this region to carboxymethylation of hydroxyl groups at C3, C2α, C2β and C6 going from low to high field. Ho continues with the argument that assignment of the C2α and C2β peaks is difficult at 90 MHz, and differentiates between the two by using *a priori* knowledge that the latter should be the larger of the two signals.

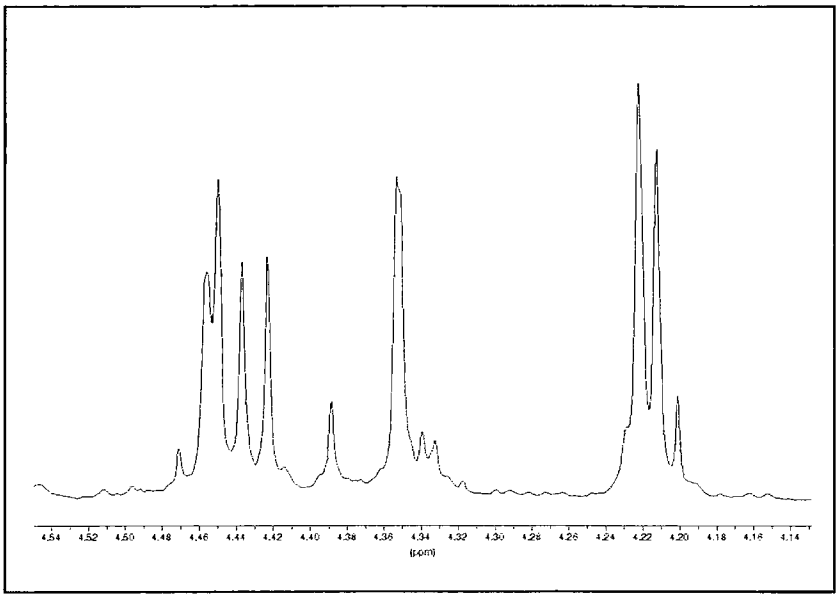


Figure 4.5 Acid hydrolyzed Aldrich SCMC, 41,927-3.

Ho's samples were acquired at 90 MHz and consequently were poorly resolved. Figure 4.5 shows a spectrum of the Aldrich SCMC acquired at 500 MHz. The improved resolution at the higher field would tend to indicate that Ho's assignment of C2 and C3 should be reversed. The coupling constants for the peaks at 4.452ppm and 4.43ppm are 3.2Hz and 6.9Hz respectively, indicating C2α and C2β. There are clearly other bands present in the spectrum with sodium glycollate present as a large singlet next to the C6 resonance at 4.21ppm.

Despite some difficulty in assigning the spectrum, Ho used the intensities of the signals to deduce the relative reactivity of the hydroxyls in cellulose toward carboxymethylation varied in the order :

$$C2 > C6 > C3$$

This result is somewhat surprising given that according to Ekkundi et al.[20] the primary OH group on C6 is expected, theoretically, to be the most reactive in alkali catalysed condensation with chloroacetic acid. Ho explained his experimentally derived order by the higher acidity and greater accessibility of the hydroxyl at C2 compared to C6, even though the latter is a primary OH group.

- **^{13}C NMR of acid hydrolysed SCMC**

The failure to achieve separation of all 16 different α and β anomeric glucose units by ^{1}H NMR, led other research groups to consider ^{13}C NMR spectroscopy as a method for characterizing hydrolysed SCMC. Reuben and Connor[8] achieved separation with ^{13}C NMR spectroscopy for an acid hydrolysed SCMC. Like Ho et al., Reuben and Conner concluded that the order of reactivity of the hydroxyls is:

$$C2 > C6 > C3$$

Their results relied on spectral deconvolution. Lorentzian lines were constructed and matched with the experimental peaks until the difference between the two spectra was minimized. Reuben and Conner note that the r.m.s. deviations between the experimental and calculated spectra were less than 1%. The integrals of the curve-resolved spectrum were then printed out. The authors estimate that based on the integrated areas of the monoprotonated carbons, C1, C2, C3, C4 and C5 of glucose, this approach is accurate to +/- 3%.

As Baar points out, this is an extremely laborious measuring procedure, followed by an equally time consuming evaluation over the 120 resonances observed in the aliphatic region of the ^{13}C spectrum. There is clearly scope for a maximum entropy approach in this sort of spectral analysis.

It is worth re-emphasizing that acid hydrolysis reduces the SCMC to the monomer. As a consequence it is not possible to examine the distribution of the substituents in the macromolecule, i.e. in the form the polymer is used in practice.

- **Sonication**

Baar continues with the argument that a mechanical method of degradation fails to break the polymer down completely to the monomer. Hence, the occurrence of end groups and their associated effects on the chemical shifts of adjacent carbons is avoided. This is achieved because the molecule is always undergoing main chain cleavage in the centre of the macromolecule. His ultrasonication based results rely heavily on spectral deconvolution, using the WINNMR program[9]. They enable the determination of the composition of the eight monomers of SCMC.

Baar's sonically derived spectra are severely overlapped and would benefit from a more rigorous approach to spectral deconvolution. Baar relies on the differential curve, i.e. experimental spectrum minus simulated spectrum, as a guide to the quality of the spectral deconvolution. He concludes that the deviations are of the same order of magnitude as the baseline noise, and hence there is a high degree of fit. The difficulty with the WINNMR approach to curve fitting is that there is a danger of over-fitting the spectrum, i.e. fitting the data to too many lines to ensure good quality residuals. One advantage of the MaxEnt approach is that the data will be fitted with the minimum number of lines that match the data and the number of lines chosen is independent of the operator.

- **Enzymatic Treatment**

Enzymatic (cellulase) hydrolysis has also be used for SCMC degradation. Gautier and Lecourtier[4] comment that acidic hydrolysis of SCMC leads to poor results due to a strong alteration of the sample. Enzymatic hydrolysis leads to short polymer chains and hence ^{13}C NMR spectra can

be obtained with relatively good sensitivity and resolution. The authors compare the ^{13}C NMR spectra of a commercial SCMC (Relative molecular mass = 1.1×10^{6}) and observe that a better spectrum quality is obtained for the enzymatically degraded sample, especially concerning:

- the spectral resolution. Gautier and Lecourtier comment that this is due the greater efficiency of the enzymatic process in degrading the polymer.
- the signal to noise ratio. The authors claim that because the enzymatic hydrolysis is more efficient, the viscosity of the solution is no longer a problem, and it is then possible to use a more concentrated sample solution.
- the number of resolved singlets is higher. Anomeric carbon atoms at the end of the polymeric chains were detected following enzymatic hydrolysis indicating shorter polymeric chain lengths, i.e. many more chain ends.

Gautier and Lecourtier found that the relative intensity of the $C_2(s)$, $C_3(s)$ and $C_6(s)$ were in the order:

$$C3 > C2 > C6$$

This is clearly not in agreement with the other literature and Baar et al conclude that enzymatic treatment leads to the occurrence of oligomers and monomers which introduce spectral overlap and hence increase the errors on the measurement of peak intensities.

4.1.2 NMR Spectroscopy Of Intact SCMC

Based on the viscosity and polymer linewidth arguments described earlier, the NMR spectrum of an intact SCMC is likely to be poorly resolved with poor S:N ratio. Chaudhari et al [10] describe how the ds and relative reactivity of the three hydroxyls has been determined directly from the ^{13}C NMR spectrum (125 MHz for ^{13}C) of an intact SCMC recorded at $70^{0}C$. Assignments are presented

(see Table 4.1) and the intensity of the peaks assigned to C2, C3 and C6 are used to determine the relative reactivity of the three possible sites of substitution.

Table 4.1 ^{13}C chemical shifts and assignments of CMC[10] (Shifts in ppm).

C_6 (us)	C_6 (s)	CH_2COO^-	C_2 (us)	C_5 Carbons *	C_3 (us) * + C_4*	C_4*	C_2 (s)	C_3 (s)	C_1 anomeric	Carbonyl
62.96	71.56	73.34, 73.85	74.06	75.86, 76.13	76.53, 76.70, 76.91	77.68	81.41	87.21	104.54, 104.99, 105.30	87.84

* Tentative Assignment, (us) Unsubstituted, (s) Substituted

The sensitivity and resolution of Chaudhari's spectra are poor and the spectra are not fully assigned. It is difficult to establish how accurate values for the peak areas could be determined using conventional integration methods and how chemical shifts can be quoted to 2 decimal places from spectra of such poor quality. Nevertheless, Chaudhari concludes that, based on the intensities of the three peaks, the relative reactivity order of the hydroxyls is:

$$C2 > C6 \sim C3$$

Chaudhari comments that whilst these results are in reasonable agreement with Reuben's ^{13}C study and Ho's ^{1}H work on acid hydrolysed systems they are different from results presented by Parfondry and Perlin [14] who used enzymatic hydrolysis to improve resolution. Parfondry and Perlin found:

$$C2 > C6 >> C3$$

Chaudhari continues with the arguement that these discrepancies are due to the majority of the literature work being based on depolymerised SCMC whereas his results are for the intact

polymer. There is still clearly a discrepancy here and a more definitive piece of work is required to determine the order of reactivities with more confidence.

Chaudhari's assignments, combined with the resolution enhancement and sensitivity improvements from modern data processing techniques, offer an opportunity to improve on the level of information that can be extracted from the NMR spectra of SCMC.

4.1.2.1 Solid State NMR of Intact SCMC

Hoshino et al[11] describe the use of ^{13}C CP/MAS NMR spectroscopy for the study of cellulose derivatives. A spectrum of CMC is presented (ds = 1.35) and the carbon peaks have been assigned as shown in Table 4.2.

Table 4.2 ^{13}C CPMAS chemical shifts of CMC (Shifts in ppm)

C=O	C_1	$C_{2,3,4}$	$C_{5,6}$ (OCH_2)	C_6
177.18	103.73	81.40	73.78	61.87

In this work, Hoshino's experiments have been extended to include a range of SCMCs of different molecular weights and different ds. The spectra, presented in Figure 4.6, were recorded at 75 MHz on a Bruker DSX300 using a 4mm High Speed MAS probe. The contact time was 2ms with a 10s recycle delay. Typical spin rates were 4kHz. (Spectra courtesy of Jeff Rockliffe, Unilever research)

The spectra are consistent with the single spectrum of SCMC presented by Hoshino. There is clearly little difference between the spectra. Molecular weight or ds have little effect on the subsequent CPMAS spectrum.

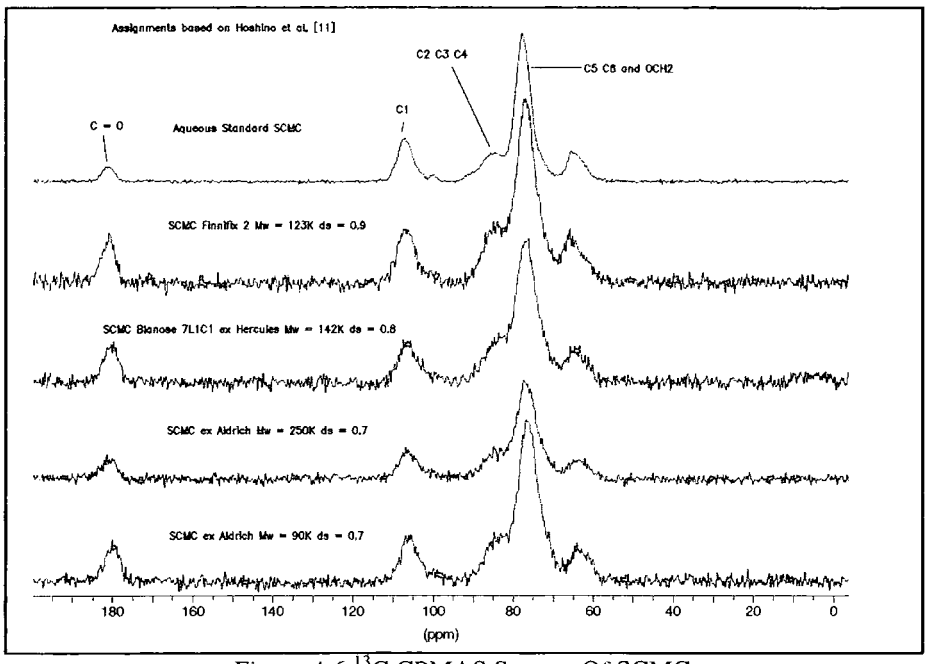


Figure 4.6 ^{13}C CPMAS Spectra Of SCMC

It is concluded that, apart from confirming the CMC structural fragments, little information about ds or microstructure can be extracted from these systems by studying them in the solid-state.

4.2 Data Processing

The following section describes how Maximum Entropy techniques have been applied to the NMR spectra of SCMC. The ability of the techniques to successfully de-noise spectra is shown to be an extremely powerful method for improving the identification of SCMC in a commercially available detergent powder.

4.2.1 ^{13}C NMR spectroscopy of intact SCMC and MaxEnt data processing.

The system chosen for this study is a SCMC from Aldrich Chemical Co. (Aldrich Ref. 41,927-3) with a Molecular Weight of 90,000 Da and an average ds of 0.7 (Aldrich figures). A 10%

(w/w) solution in D_2O was prepared and the ^{13}C spectrum recorded at 75 MHz on a Bruker DRX 500. The spectrum was recorded with inverse gated ^{1}H decoupling to suppress Nuclear Overhauser effects and with a relaxation delay of 1s, i.e. the same relaxation delay as used by Chaudhari. The short relaxation delay was chosen to provide adequate sensitivity in as short a time as possible. As only peaks from ^{13}C nuclei in very similar environments are to be deconvoluted, any differences in intensity due to T_1 are considered to be small.

Figure 4.7 is the ^{13}C spectrum of the intact Aldrich SCMC at 300 K. Expansions of the regions 55 - 95 ppm (C2 - C6) and 177 - 185 ppm (carbonyl carbons) are shown above the main spectrum. The spectrum has been processed with 10 Hz line broadening (decaying exponential) to improve the apparent sensitivity but at the expense of line-width. The spectrum is referenced to 3-(Trimethylsilyl)-1-propane sulfonic acid, sodium salt at 0 ppm. The 10 % aqueous solution was very viscous. Consequently, the mean line-width in the above spectrum is large and the degree of band overlap makes integration difficult. An attempt to narrow the lines was made by increasing the probe temperature and so the mobility of the aqueous solution. The corresponding spectrum at 343 K is shown in Figure 4.8.

Despite the relatively large 10 Hz line-broadening the resolution at 343 K is improved and the spectrum is very similar to that of Hercules 7H SCMC presented by Chaudhari et al at 343 K. Note the extra structure in the carbonyl region (~180 ppm) of the spectrum recorded at 343K.

For any probabilistic data processing the raw data must be unfiltered,[13] i.e. without line-broadening. The expanded spectrum (C2 - C6 region), acquired at 343 K, is displayed in Figure 4.9 following direct Fourier Transformation, i.e. without line-broadening. Unfortunately, the Maximum Entropy software can only display the spectra in data channels and not the more usual ppm scale.

The relative noise level and degree of band overlap is such that integration of individual resonances is not possible without spectral deconvolution.

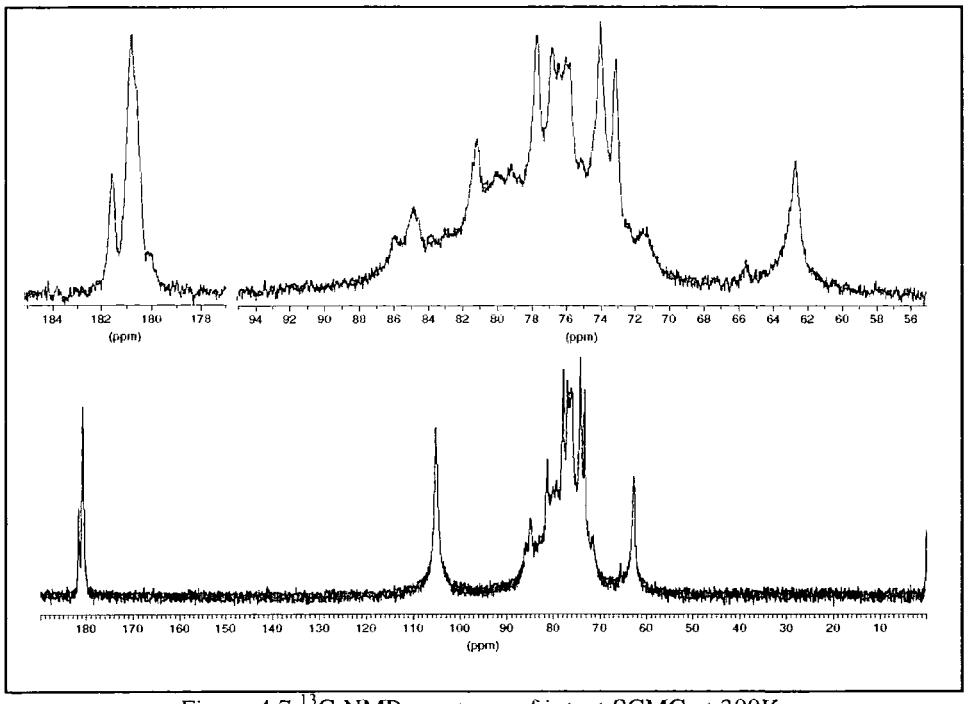


Figure 4.7 ^{13}C NMR spectrum of intact SCMC at 300K

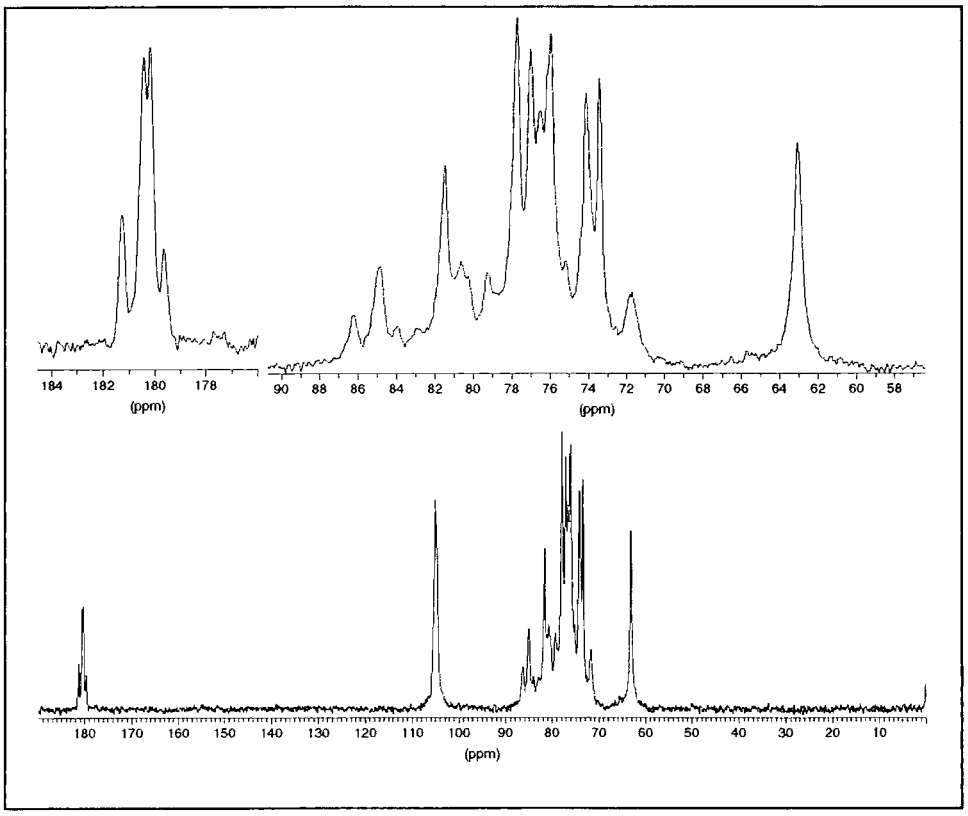


Figure 4.8 ^{13}C NMR spectrum of intact SCMC at 343K

Maximum Entropy deconvolution requires the input of an estimate of the average bandwidth and bandshape present in the spectrum, i.e. the Point Spread Function (PSF). Consequently, peaks which are actually wider than the estimate may be incorrectly split into multiple resonances and peaks narrower than the estimate will carry large error bars.

For data of this quality, three different methods are available for estimating that PSF which will extract the desired information. These methods are discussed separately in the following sections.

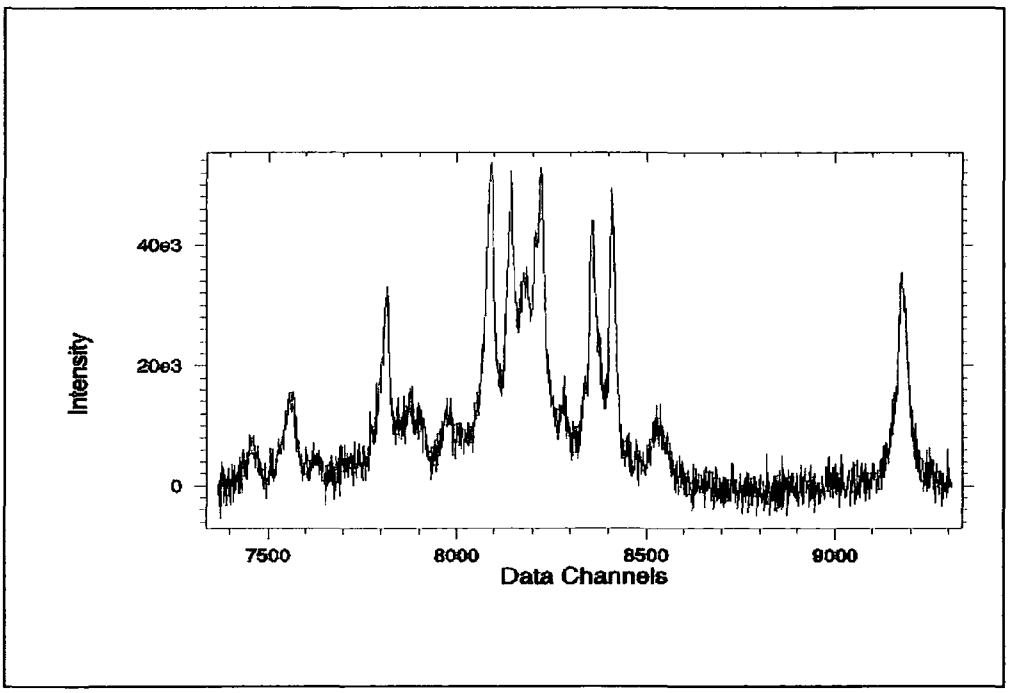


Figure 4.9 Aldrich SCMC 41,927-3 at 343k, no window function.

4.2.1.1 Estimate of PSF By Eye Directly From Raw Data

Initially, a deconvolution was attempted by using the discrete peak at 63ppm, 9178 data channels as an estimate of the PSF. A parametric curve was fitted by eye to this resonance. Although Baar et al assigned this peak to four discrete resonances, i.e. 6_G, 6_2, 6_3, and $6_{2,3}$, where the indices indicate the site of substitution on the anhydroglucose unit and the G stands for the

unsubsitituted glucose unit, the WINNMR deconvolution only found one peak. This will be used as the basis for starting this MaxEnt deconvolution, although this peak will be subsequently processed independently to ascertain if there is any evidence for the presence of more than one peak.

Figure 4.10 shows the C_6 resonance overlaid with the parametric estimate of the PSF. The residuals are also shown and indicate that the PSF is a close fit to the raw data.

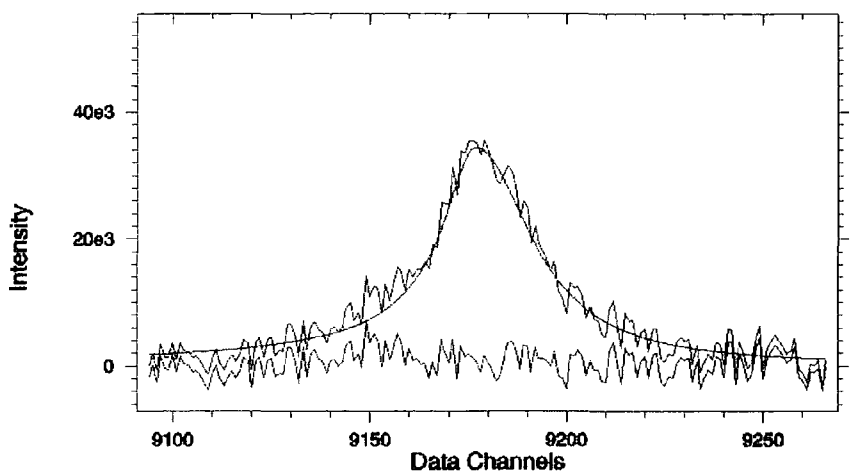


Figure 4.10 Parametric estimate of PSF.

The peak is clearly asymmetric with evidence for a shoulder on the down-field side. Comparison, by eye, of this peak with the rest of the spectrum suggests that it is a good match for some of the peaks, although the peaks at 8359 and 8410 data channels (which Baar has assigned to C_3,C_6 for 8359 and C_2 for 8410 all unsubstituted) are considerably narrower. This would suggest that either there is a number of peaks sitting under the PSF peak or there is a range of peak-widths

present in the spectrum. Nevertheless, the MaxEnt result is shown in Figure 4.11 with the Mock data, i.e. the MaxEnt reconstruction of the raw data, overlaid with the residuals in Figure 4.12.

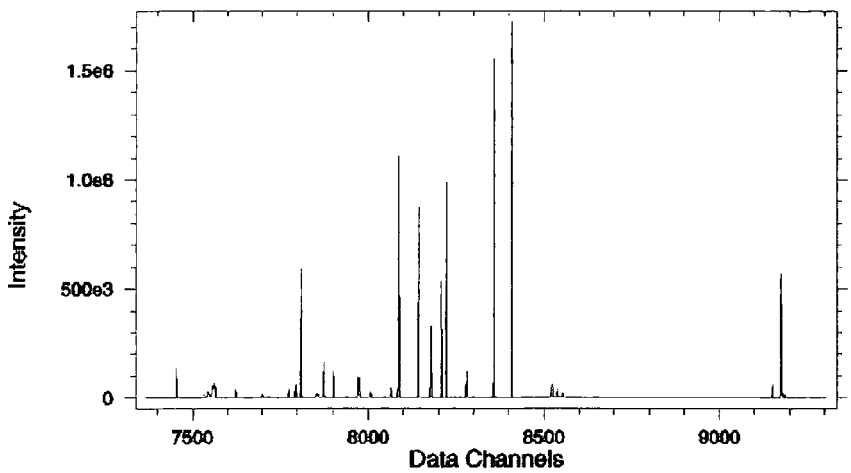


Figure 4.11 MaxEnt result

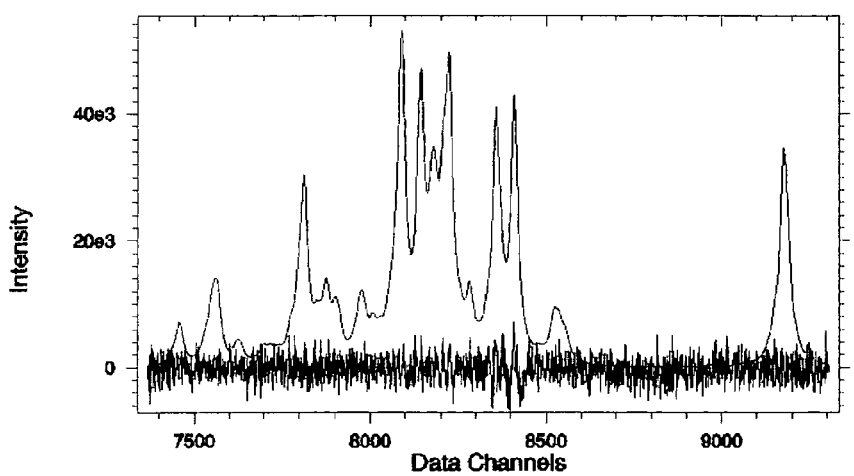


Figure 4.12 Mock Data and residuals

The residuals show structure above that present in the background noise. This is indicative of PSF mismatch. The structure is not present at all positions indicating that there may be a range of peak-widths present. Most of the structure is present on the peaks at 8359 and 8410 data channels which were identified earlier as being narrower than the PSF. If the peak chosen as the PSF standard (9178 data channels) is considered in the MaxEnt result there is evidence for a minor peak down-field of the main resonance. This is further proof that the PSF chosen was not an adequate description of all the peaks present in the data. That said, the residuals are not much larger than the noise level which suggests that a reasonable deconvolution has been achieved. The standard deviation of the residuals is 1.92e+03 compared with 1.84e+03 for the random noise.

If the MaxEnt result is quantified, the intensity of the substituted C_2, C_3, and C_6 peaks can be derived if Chaudhari's assignments are followed. (Table 4.3) It is worth noting that in this deconvolution the C_3 and C_6 peaks are split in the MaxEnt result into a number of peaks.

Table 4.3. MaxEnt intensity of ^{13}C peaks for an intact SCMC

Assignment	MaxEnt Peak Position / Data Channels	MaxEnt Absolute Intensity x 10^{-6}
C_2 (s)	7810	1.20 +/- 0.08
C_3 (s)	7533	0.06 +/- 0.14
	7544	0.18 +/- 0.26
	7557	0.32 +/- 0.32
	7562	0.27 +/- 0.37
C_6 (s)	8523	0.34 +/- 0.11
	8530	0.15 +/- 0.26
	8554	0.11 +/- 0.14

The PSF misfit is also apparent from the error bars associated with the C_3 and C_6 peaks. The error bars for some of the peaks are actually larger than the peak intensity itself suggesting that the peaks have been incorrectly split. Even allowing for these error bars, it is obvious that the most

reactive hydroxyl group is at C_2, but, because of the large error bars, no definite conclusions can be drawn about the C3 and C6 peaks. It is clear that, if quantitative information is to be derived from spectra of this quality, more accurate measurements of the peak intensities are required.

4.2.1.2 Estimate Of PSF Parameters From Evidence Values

As described in ref.[13], at the end of every deconvolution the MaxEnt algorithm outputs an evidence value. This is an internally calculated value which is effectively the logarithm of the probability of finding that PSF in the data. The more positive the number the better that PSF describes the peaks present in the data. This value can be used in a series of trials where, for example, the PSF left width and right width are varied and the effect on the evidence value noted. Thus, it is possible to derive an objective PSF. The main disadvantage with this method is that it does not adequately describe a PSF's parameters if there is a range of bandwidths present in the spectrum, a mean PSF bandwidth will be deduced.

This method has been used, for the SCMC spectrum recorded at 343 K, in an attempt to improve the error bars associated with the deconvoluted peak intensities. The optimised PSF was found to be a symmetric line of half-width 12.4 data channels, compared to the PSF optimised by eye, which was estimated to be an asymmetric line of left half-width 11.9 data channels and right half-width 17 data channels. Despite these width differences the resultant deconvolutions were found to be very similar. Figure 4.13 is the raw data overlaid with the MaxEnt result. For comparison, the corresponding MaxEnt peak intensities are shown in Table 4.4.

Whilst the error bars associated with the probabilistic method of determining the PSF are generally smaller they are not sufficiently small to be able to differentiate between the reactivity of the C3 and C6 hydroxyls. This is further evidence that there is a range of bandwidths present in the data and that the derived PSF is an average for all the peaks.

It is clear that neither of the above methods can fully describe the bandshape present in these spectra. In order to establish if there is any evidence for the peak at 9178 data channels being a composite band this will now be processed independently. The probabilistically optimised PSF will, hopefully, be free of the other line-broadening influences in the spectrum.

Table 4.4 MaxEnt intensity of ^{13}C peaks for an intact SCMC

Assignment	MaxEnt Peak Position / Data Channels	MaxEnt Absolute Intensity x 10^{-6}
C_2 (s)	7810	1.18 +/- 0.05
C_3 (s)	7533	0.06 +/- 0.08
	7544	0.10 +/- 0.14
	7557	0.30 +/- 0.33
	7562	0.39 +/- 0.18
C_6 (s)	8523	0.24 +/- 0.12
	8530	0.16 +/- 0.15
	8554	0.10 +/- 0.08

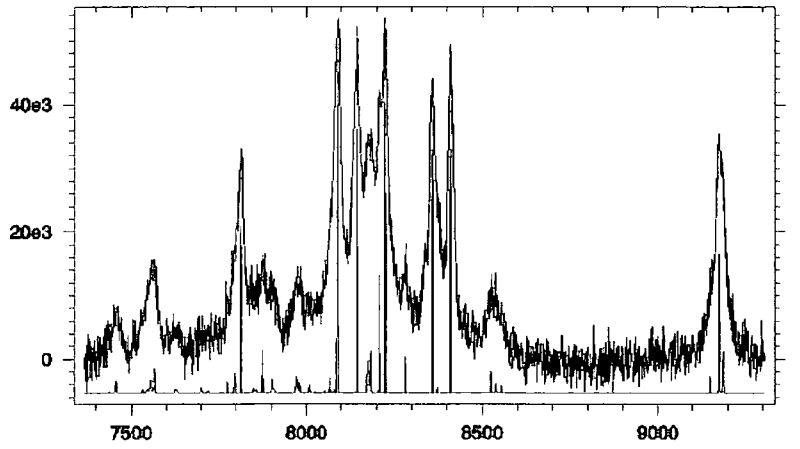


Figure 4.13 MaxEnt result using evidence optimised PSF.

- **Analysis Of Peak At 9178 Data Channels, C6**

This discrete part of the unfiltered spectrum was saved as a new file and processed using the evidence description of the PSF, no prior knowledge of PSF parameters was used. As shown in Figure 4.10, the peak is asymmetric with evidence for a down-field shoulder. The evidence values indicated an asymmetric peak of left half-width 13.5 data channels and right half-width of 18 data channels.

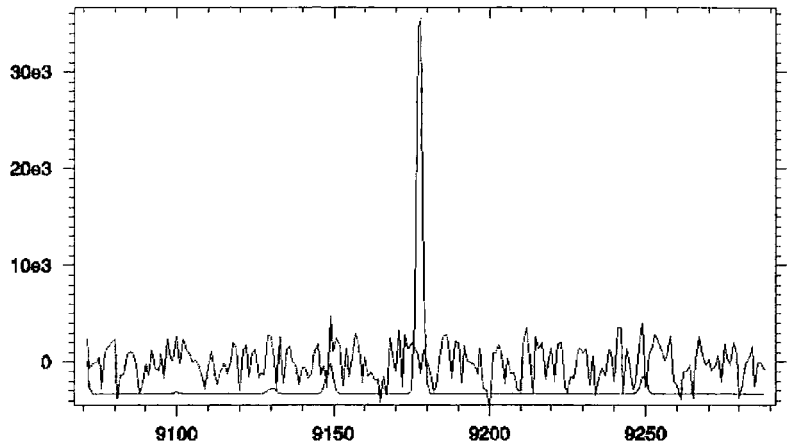


Figure 4.14 MaxEnt result and residuals.

values are quite close to those estimated by eye for this peak but are dissimilar to the parameters reflecting all the peaks in the data. The MaxEnt result and residuals are shown in Figure 4.14.

The residuals are well within the noise level indicating that the PSF is a good description of the bandshape present in the data. Apart from the main resonance there is evidence for additional structure contributing to this band. As described earlier, Baar's assignments indicate that a number species are expected to contribute to this resonance although Baar presents no spectral evidence that this is the case, i.e. deconvolution into a single band. Baar's assignments are based on an incremental calculation which he compares to experimental observation. Accordingly, the following are expected to give rise to resonances in this spectral region (calculated shifts in parentheses) :

C6 not substituted, 60.39ppm (60.8ppm)

C6 with substitution occurring at C2, 60.41ppm (60.2ppm)

C6 with substitution occurring at C3, 60.41ppm (60.2ppm)

C6 with substitution occurring at positions C2 and C3, 60.11ppm (59.6ppm).

Note: The calculated C6 shifts for the moieties substituted at C2 and C3 are coincident.

If the intensity of the MaxEnt derived peaks is considered only three are significant at one standard deviation (Table 4.5). This would be consistent with the three peaks predicted by Baar.

Table 4.5 MaxEnt intensities reported to one standard deviation (Peak 9178).

Peak Position	Peak Intensity / e06	Assignment[1]
9148	0.12	C6 sub. at C2 C6 sub. at C3 (coincident peaks)
9178	1.36	C6 no sub.
9249	0.06	C6 sub. at C2,3

The existence of these peaks has relied on empirical prediction. The observation of these peaks, whether directly or following data processing techniques, has not previously been described in the literature.

4.2.2 Spectral Fingerprinting / De-noising

As described earlier, typical fabric washing powder formulations indicate SCMC levels of only 0.5wt%. Traditional methods of characterizing these systems have relied on the ^{1}H NMR spectrum of the acid hydrolysed sample following ultra-filtration of the powder. The only information reported is ds, and a ds derived molecular weight for the average monomer.

MaxEnt data processing techniques offer a new approach for monitoring competitor products by spectral fingerprinting using the Mock Data facility, i.e. building a database of the range of

SCMCs used in typical competitor products. The techniques could be applied to either the spectrum of the intact SCMC or the acid hydrolysed system.

A major advantage of the Mock data is that, unlike the results obtained from conventional methods, it represents a method for improving the signal:noise ratio of the spectrum without broadening the peaks. Furthermore, MaxEnt trials are unnecessary so the procedure is very fast.

The method involves making an estimate of the width of the narrowest peak in the spectrum and then reducing this value by about 10 %. This is then used as the input to the MaxEnt algorithm; the shape parameters are of little consequence in this application and for most circumstances a gaussian lineshape can be used with confidence. At convergence of the MaxEnt algorithm, there will be too many peaks in the MaxEnt result because all the peaks that are significantly broader than the applied PSF will be spilt into more than one component. The MaxEnt result is irrelevant in this application. Provided the applied PSF is no wider than the narrowest peak, the Mock Data will be faithful to the spectrum. Furthermore, the signal:noise ratio of the Mock Data will be significantly higher than the original spectrum. Care should be taken to ensure that the width of the applied PSF does not approach that of the noise frequencies, otherwise unwanted correlations will be present in the final result.

As an example, consider Tide Ultra fabric washing powder from the Philippines. This is a typical competitor product containing about 0.3wt% SCMC with a ds of ~ 0.3. The problem lies in the relatively low level of SCMC in the powder. The polymer fraction is normally concentrated by ultrafiltration and then the ^{1}H NMR spectrum of the fraction is recorded following acid hydrolysis. Figure 4.15 is an example of such a powder. Note the very poor signal:noise ratio. Accurate integration of this spectrum is extremely difficult using conventional integration techniques, accurate estimates of ds are not possible with such a system. Furthermore, it is not possible to establish if this SCMC is different from one used in, say, a previous formulation. Of course, the signal:noise of the

spectrum could be improved by increasing the acquisition time but this is not always possible given the time pressures on spectrometers in industry.

An estimate of the narrowest linewidth present in this spectrum was made by eye. This value was then used as the basis of the MaxEnt deconvolution shown in Figure 4.16. Clearly, the noise level on the mock data is significantly reduced and the quality of the integral trace is much better. An estimate of the ds could be more easily made with this spectrum. An indication of the quality of the mock data is given by the residuals which are shown in Figure 4.17, i.e. the Mock spectrum has been subtracted from the raw spectrum to leave a band of noise. There is little obvious structure in the residuals indicating that the mock spectrum is a good match for the experimental spectrum. Typical deconvolution times were of the order of two minutes, making this approach much more desirable than having to acquire many more scans to improve the signal:noise ratio.

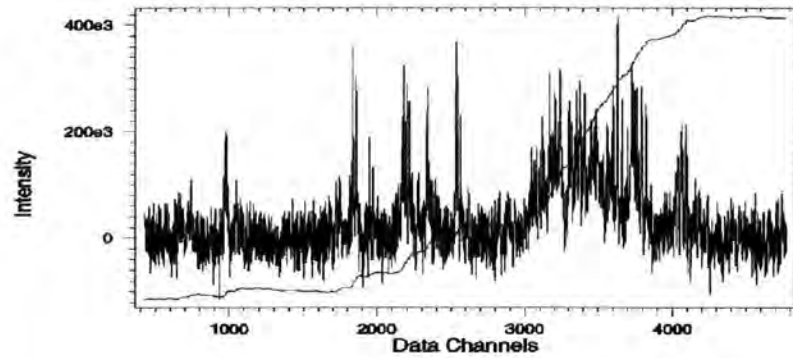


Figure 4.15 SCMC ex Tide Ultra

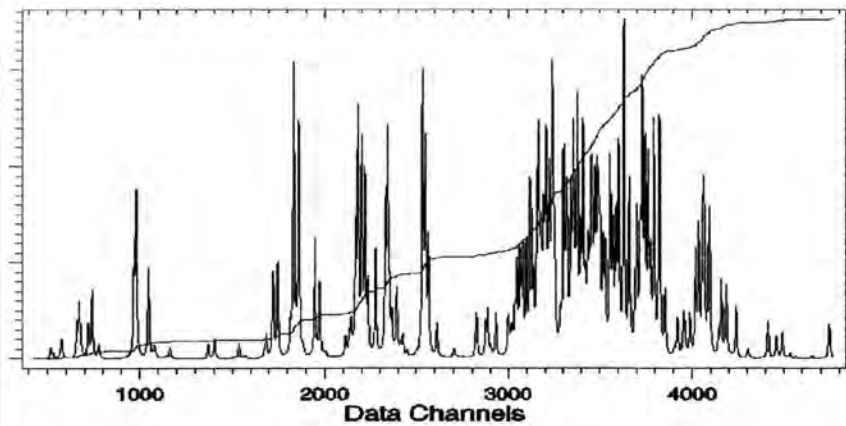


Figure 4.16 Mock spectrum with integral

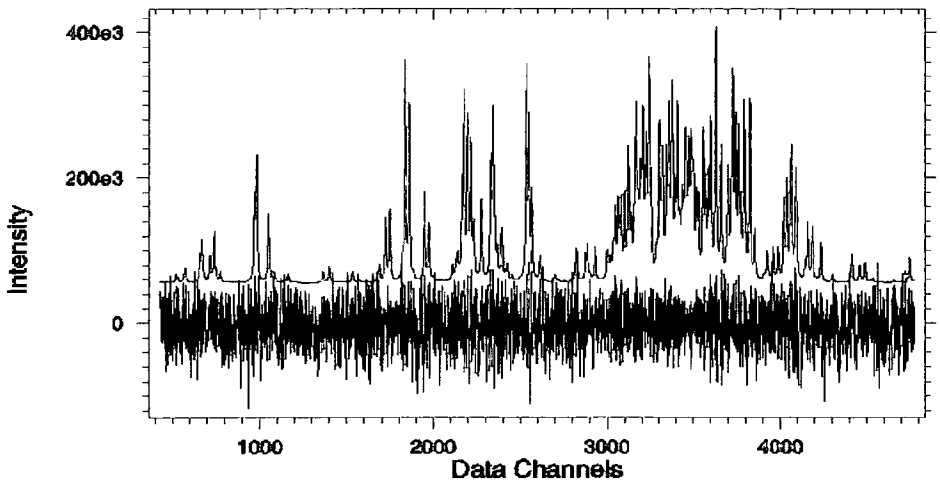


Figure 4.17 Mock spectrum and residuals

The Mock Spectrum offers an opportunity to build a spectral database of the types of SCMC present in competitor products. What happens if the signal:noise ratio is substantially reduced ? This was tested by adding random gaussian noise to the spectrum of Tide Ultra presented above. The signal:noise was reduced until it was no longer possible to make any accurate estimates of ds by conventional methods. The spectrum is shown in Figure 4.18, with the Mock Spectrum shown in Figure 4.19. The PSF parameters used were the same as for the system described earlier.

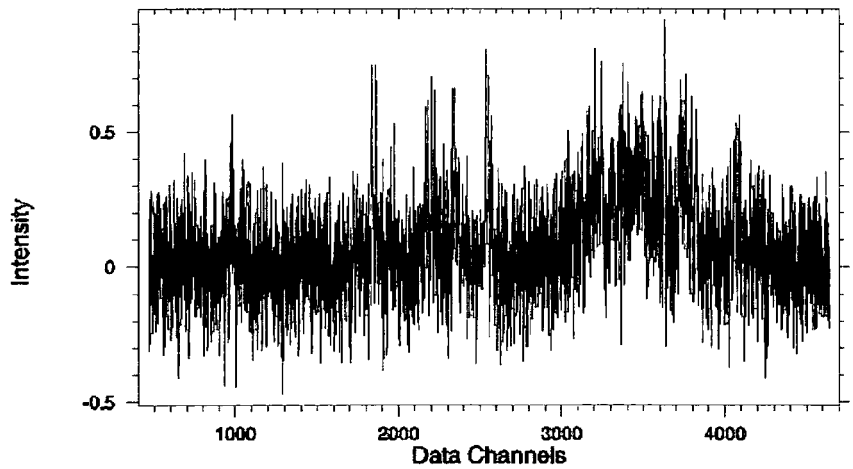


Figure 4.18 Tide Ultra SCMC with poor S:N Ratio

The signal:noise ratio of the above spectrum is such that it is not possible to say with any certainty that this is indeed a spectrum of SCMC. Furthermore, the extra noise actually gives the algorithm more degrees of freedom with which to fit the data and so computation times are significantly reduced. The Mock spectrum took just 30 seconds to produce.

The main spectral features are entirely consistent with those shown in Figure 4.16; the spectrum is unmistakably that of SCMC.

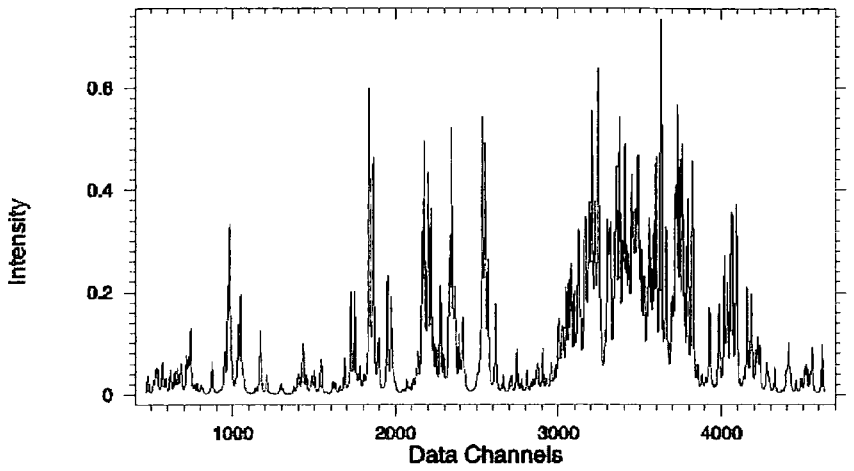


Figure 4.19 Mock spectrum

This type of approach to spectral denoising is not limited to spectra of SCMC but is a general method for improving the signal:noise ratio in all spectra where an estimate of the bandwidth can be made.

References:

1. A. Baar, W.-M. Kulicke, K. Szablikowski, R. Kiesewetter, *Macrom. Chem. Phys.*, **195**, 1483-1942, (1994).

2. C.J. Clemett, P.L. Wright, *Unilever Research Report,* **PPS 75 1128**.

3. S.N.K. Chaudhari, K.C. Gounden, G. Srinivasan, V.S. Ekkundi, *J.Polym.Sci:Part A,* **25**,337-342, (1987).

4. S. Gautier, J. Lecourtier, *Polym. Bull. (Berlin)*, **26**, 457, (1991).

5. R.W.G. Adams, S.C. Bright, R.J. Green, R.S. Johnson, *Unilever Research Report*, **PPS 68 1011**.

6. Floyd F.-L. Ho, D.W. Klosiewicz, *Anal. Chem.*, **52**, 913-916, (1980).

7. D.E. Laberge, W.O.S. Meredith, *Laboratory Practice*, **19**, 1121, (1970).

8. J. Reuben, H.T. Conner, *Carbohydr. Res.*, **115**, 1, (1983).

9. WINNMR, *Bruker Spectrospin Ltd.,* Karlsruhe, Germany.

10. V.S. Ekkundi, K.C. Gounden, G. Srinivasan, S.N.K. Chaudhari, *Hindustan Lever Research Report,* **PIN 85 0053**.

11. M. Hoshino, M. Takai, K. Fukuda, K. Imura, J.Hayashi, *J.Poym.Sci:Part A,* **27**, 2083-2092, (1989).

12. I.L. Finar, Organic Chemistry, Vol.1 The Fundamental Principles, Longmans, Green and Co. Ltd., 441, (1963).

13. L.P. Hughes, 1st Year Ph.D. Progress Report, (1997).

14. A. Parfondry, A.S. Perlin, *Carbohydr.Res.,* **57**, 39, (1977).

15. R.W. Eyler, E.D. Klug, F. Diephuis, *Anal.Chem.,* **19**, 24-27, (1947).

16. A.Z. Conner, R.W. Eyler, *Anal. Chem.,***25**, 941-943, (1953).

17. S. Mukhopadhyay, B.C. . Mitra, S.R. Palit, *Anal.Chem.,***45**, 1775-1776, (1973).

18. O. Samuelsen, *Svensk Papperstidn*, **56**, 779, (1953).

19. A Handbook of Nuclear Magnetic Resonance, Ray Freeman, Oxford University, Longman Scientific and Technical, 1987.

CHAPTER 5: STYRENE / MALEIC ANHYDRIDE

Determination of polymer composition and microstructure.

This chapter describes the application of Maximum Entropy data processing to the ^{13}C NMR spectra of a vinyl copolymer, i.e. styrene-maleic anhydride. The intensities derived from this data processing are used to derive the polymer composition. The results of the data processing are also used as the basis for suggesting a model, based on Markovian statistics, to describe the polymer microstructure. Both the carbonyl and aromatic regions of the ^{13}C spectrum are processed to demonstrate the consistency of the derived results. The results of processing spectra acquired at two different magnetic field strengths are presented; it is demonstrated that for this system the advantages of working at the higher field strength are limited.

An alternative to a literature peak assignment is made on the basis of a stereochemical argument. Whilst the Maximum Entropy results are inconclusive in determining which assignment is correct, the analysis does demonstrate the ability of this type of processing for simultaneously deconvoluting peaks that are severely overlapped, and improving the signal : noise ratio (S/N) of the spectrum. Conventional data processing methods cannot simultaneously improve the S/N ratio and separate overlapped peaks.

5.1 Introduction

Copolymers, such as styrene / maleic anhydride, can exhibit a very large number of different possible structures. The polymer can exist as a random copolymer, a block copolymer or an alternating copolymer. For a copolymer of monomers A and B:

ABAAABBABBA is a random copolymer

AAAABBBBBBB is a block copolymer

and ABABABABABA is an alternating copolymer.

These structural distributions are determined by the nature of the polymerisation process, whether the process is stochastic or chemically controlled in some way. These distributions are known as the polymer microstructure and can have a profound effect on the physical and mechanical properties of that polymer. Therefore, determination of polymer microstructure is an important analytical requirement. Under favourable conditions, it may be possible to probe polymer microstructure by ^{13}C NMR spectroscopy, especially for low molecular weight polymers which tend to produce relatively narrow NMR peaks and so NMR spectra free from peak overlap.

The ^{13}C nucleus is a useful nucleus for probing polymer microstructure. It has a natural abundance of only 1.1% and as a result $^{13}C - ^{13}C$ couplings are not normally observed in the NMR spectrum. $^{13}C - ^{1}H$ coupling can be easily removed by decoupling techniques and the typical ^{13}C solution state NMR spectrum consists of a number of well resolved singlets. The ^{13}C chemical shift range is over 200ppm and minor changes in chemical structure can cause a shift in a ^{13}C resonance due to an interaction with another nucleus as far as five bonds away. This enables the analyst to use the chemical shift to probe polymer microstructure over the triad (three carbon neighbours) to pentad (five neighbours) range, although being able to derive pentad structure is very unusual and model systems are usually employed.

Unfortunately in polymers the situation may be further complicated because the chemical shift is also sensitive to any tacticity present in the polymer. This can result in uncertainty in assignment, particularly if the peaks are overlapped. Tacticity is the type and extent of stereoregularity within a polymer. If all the monomers possess the same enatiomorphic configuration the polymer is described as isotatic, i.e. all the substituents on a polymer chain appear on the same side of the chain. When the substituents alternate regularly from one side of the chain to the other the polymer is said to be syndiotactic, and when a random configuration of substituents is found the polymer is atactic. A ^{13}C nucleus in a sydiotactic environment is likely to resonate at a slightly different chemical shift to a nucleus in a wholly atactic polymer. The extent of this

stereoregularity also effects the properties of a polymer; irregular tacticity can determine the degree of crystallinity present in the polymer and may render it useless for a particular application.

Such an unfavourable case, exhibiting both irregular tacticity and microstructural effects on the NMR spectrum, is seen for the copolymer of styrene and maleic anhydride (see Figure 5.1). The styrene-quaternary aromatic carbons are found in many different environments and in the ^{13}C spectrum the styrene-quaternary aromatic region exhibits such a number of different ^{13}C peaks that the region spans a range as large as 10ppm.

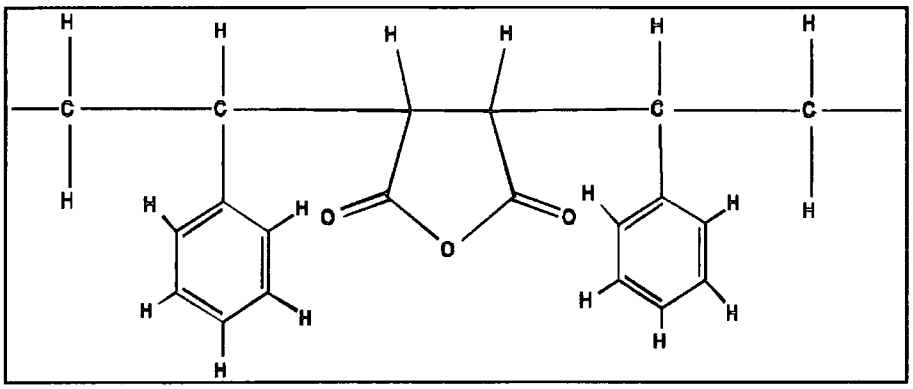


Figure 5.1 Structure of styrene-maleic anhydride-styrene unit

The difficulty in analysing this particular region of the spectrum has been described by Hill et al.[1] However, assignments are presented by Bhuyan[2] and Buckak[3], together with the comment that the monomer distribution calculated using these assignments could have up to 20% error due to poor spectral resolution.

This chapter describes how, using Maximum Entropy data processing, an accurate measure of monomer composition can be obtained from the ^{13}C spectrum of styrene-co-maleic anhydride. A model is presented to describe the polymer microstructure based on the Maximum Entropy derived peak intensities and a literature assignment made by Bhuyan [2] and Buckak [3] is reversed based on a stereochemical argument.

5.2 Experimental

5.2.1 NMR Spectroscopy

The styrene / maleic anhydride copolymer chosen for analysis has a number molecular weight average of 1900, and is described by Aldrich as nominally 75% styrene.[4] It is assumed that the Aldrich figure is based on the known monomer feed ratio. A ^{1}H NMR spectrum was acquired in deuteroacetone –d6 (3% w/w) and is used to calculate the polymer composition as a check for the subsequent work. It is shown in Figure 5.2. This spectrum was acquired on a Bruker AM spectrometer operating at a proton frequency of 360.13 MHz. A 20° flip angle was used and an interpulse delay of 3 seconds. The spectral width was chosen to ensure all the bands of interest were observed.

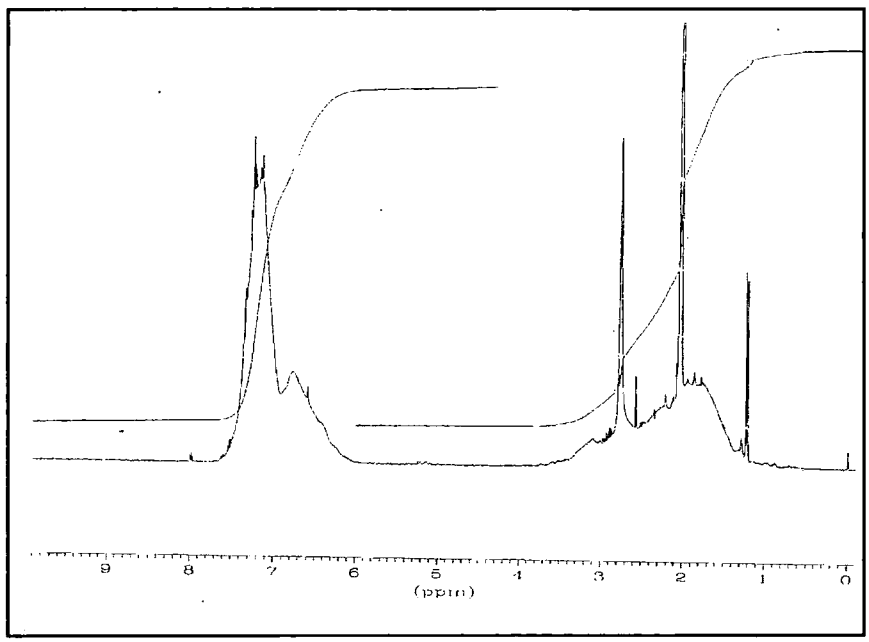


Figure 5.2 ^{1}H NMR spectrum of Styrene / Maleic Anhydride

The spectrum is consistent with a number of severely overlapped peaks of considerable width due to the polymer as well as much narrower peaks from species of lower molecular weight. These narrow peaks include residual acetone, possibly succinic acid and cumene. The cumene is known to be used as a terminator for the free radical polymerisation.[5]

Using this spectrum an approximate polymer composition can been calculated. If it is assumed that the intensity in the aromatic region (6 – 8ppm) is due to the five protons from the styrene and the intensity in the aliphatic region (1 – 4ppm) is due to the backbone from both the styrene and maleic anhydride, i.e. 3 protons from the styrene and 2 from the maleic anhydride, a composition of 73% styrene is calculated, based on the integrals of these two regions. A crude correction for the low molecular weight species is made by ignoring their contribution to the integrals. Although this value is in good agreement with that determined by Aldrich[4] the use of ^{1}H NMR in this context may be limited due the presence of water. Water is often difficult to exclude in such systems.

Further evidence supporting this composition is provided by carbon and hydrogen elemental-analysis[6], from which a figure of 81% Styrene has been calculated (oxygen level determined by difference), and from direct chemical ionisation mass-spectrometry which indicates a styrene content of 78%.[7]

^{13}C NMR spectra of this system were acquired on a Bruker AMX360 spectrometer operating at 90 MHz for ^{13}C and a Bruker DRX500 spectrometer operating at 125 MHz for ^{13}C. Approximately 0.3g of polymer was dissolved in $3cm^3$ deuteroacetone-d6. Both spectra were acquired over a spectral width of 27 kHz, with broad-band ^{1}H decoupling. A 90^0 carbon pulse was used with a 5 second interpulse delay. The spectra were acquired at 27^0C for approximately 28 hours. The spectrum acquired at 125 MHz is shown in Figure 5.3. The most obvious comment to make about this spectrum is that, despite being acquired at a relatively high magnetic field strength, the signal : noise ratio of the polymer peaks is very low. The spectrum is dominated by the solvent resonances. It is very difficult to determine any chemical information from this spectrum without both signal : noise improvement and resolution enhancement. The aromatic quaternary (C1) region of the spectrum is shown in the expansion. This area will be used as the basis for the following data processing because it is likely that these carbon nuclei will be particular sensitive to the polymer microstructure.

The following assumptions have been made before applying Maximum Entropy data processing techniques to any of the ^{13}C spectra:

- The spin-lattice relaxation times, T_1, of each of the C1 aromatic peaks are very similar and unlikely to have any differential effect on the relative intensities of the deconvoluted peaks.
- Any enhancements in signal intensity due to the Nuclear Overhauser Effect are likely to be small. These are quaternary carbons and so are unlikely to be effected by NOE enhancements.

The above assumptions are based on the argument that any deconvolution performed is on a number of carbon resonances in very similar magnetic environments.

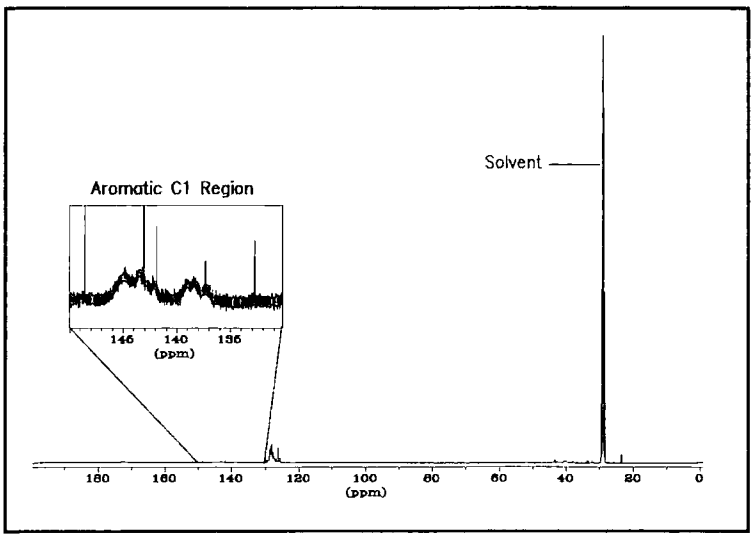


Figure 5.3 ^{13}C NMR (125 MHz) spectrum of styrene-maleic anhydride copolymer

5.2.2 Data Processing

An expansion of the aromatic quaternary region (C1) (134 ppm –150ppm) and the carbonyl region (168 ppm – 176 ppm) is shown in Figures 5.4 and 5.5. Both these spectra were acquired at 90 MHz. There is a strong similarity between these spectra and the spectra of polymer SMA 3000A (75% styrene, 25% maleic anhydride) in reference [3].

Both the aromatic and carbonyl regions of the spectrum clearly contain a large number of unresolved peaks. This lack of any one discrete peak makes optimisation of the Maximum Entropy PSF width and shape parameters difficult – there is no obvious peak on which to base the model. This

difficulty can be overcome by designing a series of trial deconvolutions in which the overall width of the applied PSF is varied for each of a suitable range of values of sigma. A graph of optimum peak width against sigma is plotted, i.e. a sigma profile, and the optimum total PSF width and corresponding value of sigma are then determined by estimating the point of inflection. This method for determining PSF parameters for spectra in which the peaks are severely overlapped or the noise level is in doubt has been described in section 3.2.1.2.

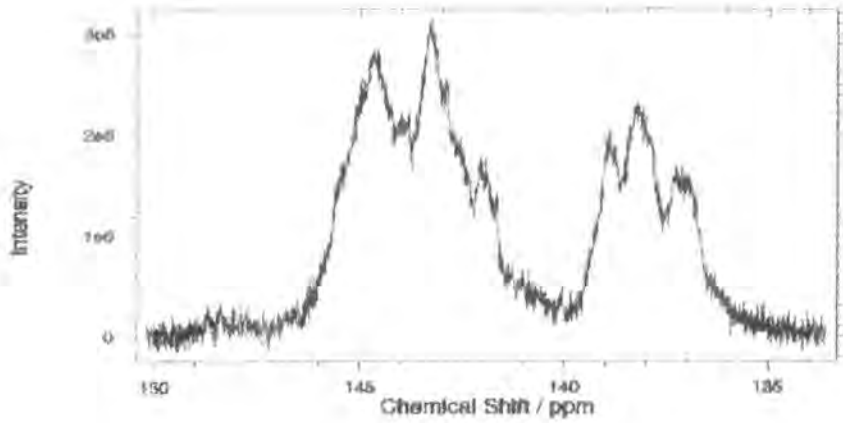


Figure 5.4 ^{13}C NMR spectrum (90 MHz) of styrene / maleic anhydride (aromatic region)

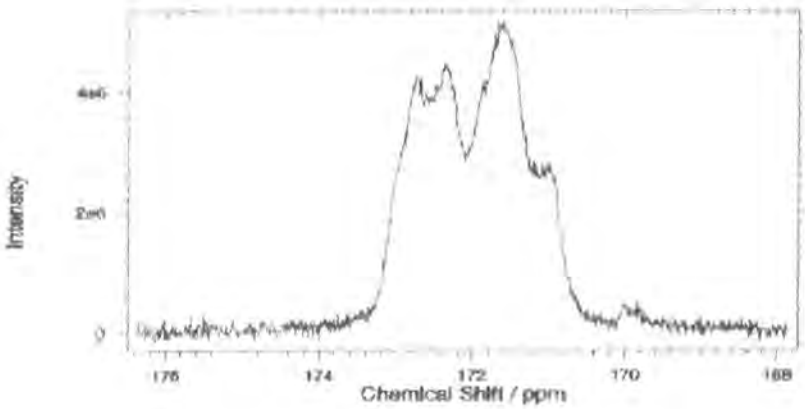


Figure 5.5 ^{13}C NMR spectrum (90 MHz) of styrene / maleic anhydride (carbonyl region)

The sigma profile for the aromatic quaternary region, generated using the Memsys5 algorithm, is shown in Figure 5.6. The overall width determined from the point of inflection was then used as the basis for generating an evidence matrix in width the left width : right width ratio was changed (see section 3.2.1.1 for a description of evidence matrix). The shape parameters were also determined from an evidence matrix.

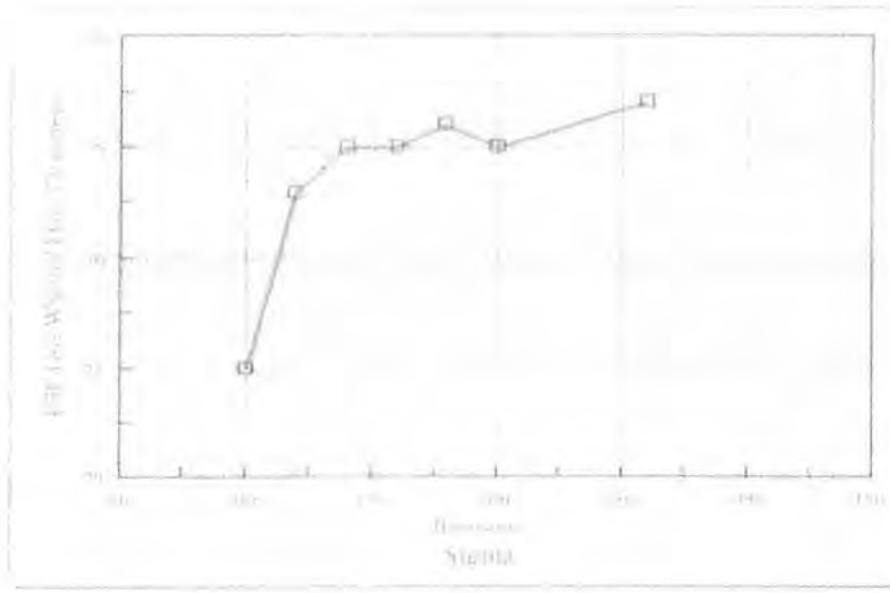


Figure 5.6 Sigma profile – optimum peak width ~ 70 data channels

The sigma profile shown in Figure 5.6 has an ill-defined point of inflection and is indicative of the fact that the this part of the spectrum contains a number of peaks of different width. The point of inflection could lie anywhere between a peak left-width of 34 and 36 data channels. This lack of accuracy was examined by repeating the optimisation process. The newly derived PSF differed slightly from that derived earlier, see Table 5.7, but this difference will be shown to have little effect on the calculation of polymer composition or on the stereochemical conclusions.

Table 5.7 PSF parameters derived from Memsys5

	New PSF	Old PSF
Width (L / R) / data channels	39 / 31	38 / 32
Wing (L / R)	0.7 / 1.4	0.6 / 1.4

where L = left and R = right

From Table 5.7, the overall width of the PSF has not changed and, given the low S/N ratio of the spectrum, the change in ratio (left / right) is unlikely to have an effect on the resultant deconvolutions, i.e. on the number or intensity of the resultant peaks. The peak shape parameters are much less critical to an accurate deconvolution than the corresponding width values so a small change in PSF shape is unlikely to have a significant effect.

Experience, with spectra of this quality, has shown that the width of the input PSF can be within +/- 10% of the optimum without having a detrimental effect on the final deconvolution. This variation in peak width is not surprising. The spectrum is likely to contain a range of peak widths reflecting the different mobilities present within a polymeric chain. It should be noted that each spectral region has been processed independently with sigma profiles being generated, where necessary, for each.

Figure 5.8 shows the Memsys5 mock spectrum overlayed with the residuals (experimental spectrum minus Memsys mock spectrum) for the 90 MHz data. The intensity and randomness of the residual's intensity enable an assessment to be made as to the reliability of any feature in the Memsys5 result. The fact that the intensity of the residuals is very close to that of the noise and very little structure is observed on the residuals suggests that the Memsys5 calculated spectrum is in very close agreement with the experimental spectrum. The Memsys5 result is shown in Figure 5.9 and Table 5.10 gives the Memsys5 derived intensities presented with one standard error. Fourteen main peaks are observed in this aromatic quaternary carbon region of the spectrum together with seven minor peaks. The fourteen most intense peaks fall into four groups, which based on triad structure can be desribed as MSM, SSM, MSS, and SSS, where S refers to one styrene unit and M refers to one maleic anhydride unit. The groups are highlighted in Table 5.10.

The peaks at 144.39 ppm and 144.66 ppm have not been baseline resolved by the algorithm and the intensity of each peaks has a substantial error because of the uncertainty in the absolute intensity of the individual peaks. However, as the Memsys5 errors are negatively correlated, the combined peak intensity is known with a much greater degree of accuracy and hence the errors assigned to the total cumulant for each group are much smaller.

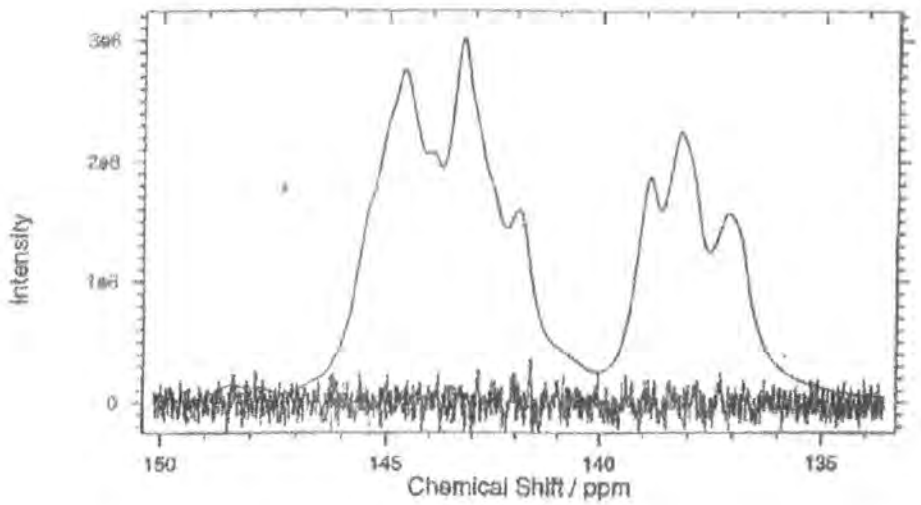


Figure 5.8 Memsys5 mock data and residuals: aromatic region

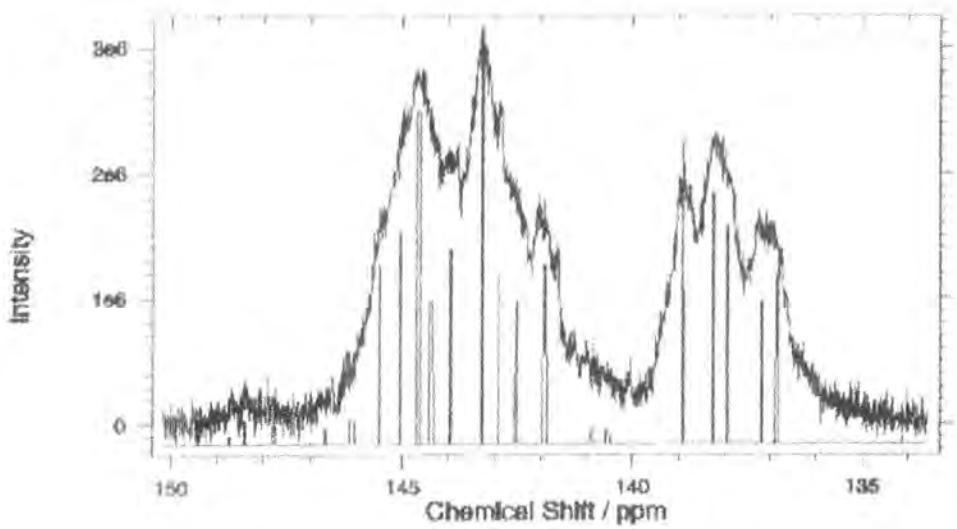


Figure 5.9 Experimental spectrum overlaid with Memsys5 result

Table 5.10 Chemical shift / intensity data based on Memsys5 deconvolution (old PSF)

Group	^{13}C Chemical Shift / ppm	Cumulant (+/- one standard error)	Total Cumulant For Group (+/- one standard error)	Triad Based On Literature Assignment[1]	Triad Assignment Based On this Work[1]
1	136.87	5.27 +/- 0.63	24.21 +/- 0.19	MSM	MSM
	137.20	4.41 +/- 0.57			
	137.94	6.75 +/- 0.75			
	138.26	7.79 +/- 0.93			
2	138.91	9.56 +/- 0.22	9.56 +/- 0.22	SSM	MSS
3	141.91	5.53 +/- 1.79	5.53 +/- 1.79	MSS	SSM
4	142.54	4.44 +/- 0.26	55.61 +/- 0.26	SSS	SSS
	142.92	5.28 +/- 0.46			
	143.29	12.96 +/- 0.47			
	143.95	6.04 +/- 0.35			
	144.39	4.42 +/- 2.27			
	144.66	10.37 +/- 1.63			
	145.06	6.61 +/- 0.92			
	145.49	5.52 +/- 0.31			

[1] where M = maleic anhydride, S = styrene, so MSS = maleic-styrene-styrene triad.

5.2.3 Discussion

The following discussion will desribe how the Memsys5 derived intensities have been fitted to a number of different polymerisation models.

5.2.3.1 Assignment of block ends

Ramey and Buckak [3] and Bhuyan and Dass [2] have made the ^{13}C assignments shown in Table 5.11. The author's claim these assignments are based on substituent additivity effects and statistical considerations. In this work, the block end, styrene-styrene-maleic anhydride (SSM), and maleic anhydride-styrene-styrene (MSS) assignments are reversed.

Table 5.11 Group Assignments

Triad	Group based on literature assignment[2,3]	Group as Re-assigned In This Work
MSM	1	1
SSM	2	3
MSS	3	2
SSS	4	4

There are two main reasons for making this re-assignment. Firstly, the literature assignments are based on the assumption that the β substituent will have the largest effect on the chemical shift of the ^{13}C nucleus of interest. This ignores the through space effect of a γ substituent. The sensitivity of the C1 aromatic carbon of polystyrene to γ effects is noted in reference [8]. The re-assignment presented in this work is more compatible with these steric considerations. Secondly, the statistical considerations presented give no information on the relative assignments of the MSS and SSM triads as the probabilities of both are assumed to be the same.

5.2.3.2 Calculation of polymer composition

It is possible to calculate a polymer composition, based on the Memsys5 derived intensities, which is independent of the block end assignments discussed above. In the following calculations it is assumed that the probability of a maleic anhydride unit joining another maleic anhydride unit is zero.

Total Maleic Intensity = 0.5 x (2MSM + SSM + MSS) .....{i}

Total Styrene = SSS + MSS + SSM + MSM .....{ii}

Substituting the values from Table 5.10, for both PSFs, into the above equations gives the polymer compositions as shown in Table 5.12.

Table 5.12 Determination of polymer composition

	Memsys5 with old PSF	Memsys5 with new PSF
% Styrene	74.9 +/- 1.2	73.8 +/- 1.6

Both deconvolutions lead to a calculated styrene composition which is in very good agreement with the figures quoted by Aldrich, the results from the ^{1}H NMR spectrum and the mass-spectrometry results. This is an extremely encouraging result because, despite the very poor signal : noise ratio and the severe peak overlap in the spectrum, it means that the peak intensities can be trusted for making micro-structural conclusions.

The error bars on the polymer composition have been calculated by taking one standard error on the total cumulants for each group (see Table 5.10) and calculating the maximum range in styrene content. The above results are independent of the assignments of the block ends, MSS and SSM.

5.2.3.3 Polymer Microstructure

General inspection of the Memsys5 results and making the assumption that the structure within each of the groups of Table 5.11 is largely the result of tacticity, leads to the conclusion that there is a strong correlation between microstructure and tacticity. There is only one preferred conformation for the triads MSS and SSM, but four conformations for the triad MSM. This in itself suggests that at least a second order Markovian model is required to describe the microstructure of this polymer. The values of the total group cumulants, given in Table 5.10, can be used to test different polymerisation models.

5.2.3.3.1 First Order Markovian Model

For a first order Markovian process the probabilities $P_{M/SS}$, i.e. the probability that a maleic unit will add to a 'styrene-styrene' block, and $P_{M/MS}$, i.e. the probability that a maleic unit will add to a 'maleic-styrene' block, must be equal. This can be rationalised by the following equations:

$$P_{M/SS} = A_{SSM} / (A_{SSM} + A_{SSS}) \qquad \text{.....(iii)}$$

$$P_{M/MS} = A_{MSM} / (A_{MSM} + A_{MSS}) \qquad \text{.....(iv)}$$

where A = peak area

Substituting the values from Table 5.10, and using either of the block end assignments, it has been calculated that for both algorithms, with either the old or newly derived PSF, the two probabilities are not equal. One example is shown below for the old PSF.

$P_{M/SS} = 5.53 / (5.53 + 55.61) = 0.090$

$P_{M/MS} = 24.21 / (24.21 + 9.56) = 0.72$

Therefore, a first order Markovian model does not adequately describe the polymerisation process.

5.2.3.3.2 Second Order Markovian Model

If a second order Markovian process is assumed, it is possible to calculate the polymer composition for both block end assignments based on:

$$P_S/P_M = 1 + (A_{MSS} / A_{MSS} + A_{MSM})(A_{SSM} + A_{SSS} / A_{SSM}) \qquad \text{.....(v)}$$

where P_M is the proportion of maleic anhydride in the polymer and

P_S is the proportion of styrene in the polymer.

The results of calculating the polymer composition using equation {v} are given in Table 5.13. Presented are the Memsys5 results using both PSFs and also the results from a MassInf deconvolution, again using both PSFs.

Table 5.13. Polymer composition (% Styrene content) based on second order Markovian statistics

Number Of Standard Errors on peak at 141.91 ppm	Memsys5 With Old PSF		MassInf With Old PSF		Memsys5 With New PSF		MassInf With New PSF	
	Literature	Reversed	Literature	Reversed	Literature	Reversed	Literature	Reversed
-2	60	90	71.8	78.2	70.9	75.9	71.9	77.7
-1	66	84	72.0	78.3	71.1	76.1	72.1	77.8
0	69	80	72.2	78.4	71.3	76.3	72.2	77.9
+1	72	77	72.4	78.6	71.5	76.5	72.3	78.1
+2	74	75	72.6	78.7	71.7	76.7	72.5	78.2

From Table 5.13 the Memsys5 deconvolution, using the old PSF with the 'reversed' assignment, gave a calculated styrene content of 75% if two standard errors were allowed on the peak at 141.91 ppm. This agreement with the Aldrich value supported the stereochemical argument for the reversed end-group assignment if a second order Markovian statistics described the polymerisation process. However, the introduction of the MassInf algorithm has resulted in smaller errors for this peak and it is clear that for the old PSF neither the literature nor the reversed assignment are consistent with a second order model.

The introduction of the new PSF has also demonstrated the inadequacy of the second order Markovian model; both algorithms, with either block end assignment, failed to give a calculated composition in agreement with that claimed by Aldrich. It is interesting to note that in all cases the literature block end assignments tended to give a styrene composition less than expected and the reversed assignments a styrene level higher than expected.

In conclusion, apart from the stereochemical argument presented in section 5.2.3.1, the peak intensities derived from the 90 MHz spectrum do not conclusively confirm the reassignment of the polymer block ends. The errors on the Memsys5 and MassInf results are too large to confirm or deny this reassignment. In an attempt to reduce these errors and simplify the spectra a ^{13}C spectrum

was recorded of the same system at a higher magnetic field strength, equivalent to a ^{13}C frequency of 125 MHz. (see section 5.2.3.5).

5.2.3.4 Discussion of the number of peaks in each group

There is only one peak in each of groups 2 and 3 given in Table 5.10. This suggests that the ends of the styrene blocks require a specific stereochemistry. The four peaks in group 1 are easily rationalised in terms of each maleic anhydride unit having two possible orientations with respect to the styrene, each orientation having a similar probability (within a factor of ~1.5) based on the peak intensities. It will be shown in section 5.2.3.4.2 that there is evidence for this assignment in the Memsys5 deconvolution of the carbonyl region of the spectrum. The peaks in group 4 are more difficult to account for.

5.2.3.4.1 Rationalisation of the eight peaks in the SSS triad region of the spectrum

For clarity the eight peaks observed in the SSS region (group 1) are listed in Table 5.14.

Table 5.14 The Eight Peaks in the SSS triads

Peak	^{13}C Chem. Shift / ppm	Memsys5 Intensity	Empirical expression for intensity	Calculated intensity
4a	142.54	4.44	$As^5\{1+s\}$	4.25
4b	142.92	5.28	$As^4\{1+s\}$	5.31
4c	143.29	12.96	$A\{1+s\}$	12.96 (normalised)
4d	143.95	6.04	As	5.76
4e	144.39	4.42	As^2	4.61
4f	144.66	10.37	$As\{1+s\}$	10.37
4g	145.06	6.61	$As^3\{1+s\}$	6.64
4h	145.49	5.52	$As^4\{1+s\}$	5.31

A full assignment of this region is beyond the scope of this work. The aim of this discussion is to show that the structure in this region is not an artifact of the deconvolution programme

Considering the areas A associated with the peaks presented in Table 5.14, the following relationships can be seen:

$A_{4f} / A_{4c} = 0.80$

$(A_{4g} + A_{4h}) / (A_{4c} + A_{4d}) = (0.80)^2$

$A_{4b} / A_{4g} = 0.80$

$A_{4a} / A_{4h} = 0.80$

It would seem unreasonable to assume that the agreement between these figures is coincidental.

If we let u=0.8, then the intensities of all eight peaks can be expressed in terms of simple powers of a constant, s. The 4th column of Table 5.14, presents these empirical expressions for the intensities of the eight peaks. Numerical values, calculated from the expressions in column 4 are presented in column 5. These agree with the experimental values to within ~5%.

This section will show that the structure in this region of the spectrum can be described in terms of a simple model (determining whether or not this is the only or the best model is beyond the scope of this work). This model is based on two observations:

1) The styrene blocks are short (Table 1 group 4 / group 2 = 5.8).

2) There is only one configuration at the end of each block.

Together these suggests that the styrene units in the block will be influenced by a 'steric handle', i.e. the fixed configuration of MSS at the end of the block. This lifts the degeneracy of the relative configurations S_mS_rS and S_rS_mS. It is convenient to dispense with **relative** configurations and to think in terms of **absolute** configuration, here arbitrarily denoted u and d.

There are now 8 possible S<u>S</u>S environments, as shown in the vector v

$$v = \begin{matrix} uuu \\ uud \\ udu \\ udd \\ duu \\ dud \\ ddu \\ ddd \end{matrix}$$

The relative populations P of these configurations in the block can be calculated using a transition matrix **a** approach[9],

$$P = (\Sigma\, i\, A_i\, \mathbf{a}^i)V \qquad .....\{VI\}$$

where V is the initial populations of configuration vector, v, corresponding to the block end. The block end is a MSS unit, provided only with tacticity statistics of second order or lower are dealt with this can be simulated using the matrix v with identical values for pairs of elements (uuu = duu, dud = uud etc). Each pair of elements refers to **one** configuration of the end block.

It is necessary to perform the sum as the block end has a fixed configuration which imposes constraints on the allowed configuration in the next triads in the block. This effect diminishes further along the block.

The above discussion of the ratios of peak intensities suggests that the transition matrix **a** should contain elements of value 0.556 and 0.444 (0.444 / 0.556 = 0.8 = s above, 0.444 + 0.556 = 1), if r = 0.556 then a trial form of **a** is,

r	0	0	0	r	0	0	0
1-r	0	0	0	1-r	0	0	0
0	r	0	0	0	r	0	0
0	1-r	0	0	0	1-r	0	0
0	0	r	0	0	0	r	0
0	0	1-r	0	0	0	1-r	0
0	0	0	r	0	0	0	r
0	0	0	1-r	0	0	0	1-r

a has the property that $\mathbf{a}^n$ for n>3 = $\mathbf{a}^3$. Therefore the sum in equation {VI} only requires terms in $\mathbf{a}^1$, $\mathbf{a}^2$ and $\mathbf{a}^3$. The challenge is to find coefficients A_1, A_2, A_3 and vector V such that P contains

populations consistent with the experimental data. It is found that for A_1= 7.53, A_2=3.33, A_3=16.77

and

V=

0
1
0
0
0
1
0
0

then P=

5.76
4.60
12.96
10.36
6.66
5.32
5.32
4.25

compared with experimental values of

5.76
4.61
12.96
10.37
6.64
5.31
5.34
4.28

The elements of V suggest non-zero values only for the configurations uud and dud.

To conclude, it is possible to show that the measured intensities of the eight peaks are consistant with a simple Bernoullian model of tacticity provided that we allow two constraints:

1) a single fixed orientation of the end group

2) the styrene blocks being short.

No input of the number of peaks expected from any one deconvolution is made to the Memsys5 algorithm. It has been demonstrated that the eight peaks in the ratio described above are more than an artefact of the deconvolution program. The fact that they can be fitted to such a simple model is probably an oversimplification. However, development of more appropriate models is beyond the scope of this current work.

5.2.3.4.2Carbonyl region of the spectrum

The MSM triplet region of the C_1 aromatic spectrum showed four peaks attributed to each maleic anhydride having two possible orientations with respect to the styrene. If this is the case then evidence for this should be seen in the carbonyl region of the spectrum. Figure 5.15 shows the carbonyl region of the 90 MHz ^{13}C spectrum with the Memsys5 deconvolution of this spectrum overlaid. The Memsys5 result shows five major peaks and one minor peak, as summarised in Table 5.16.

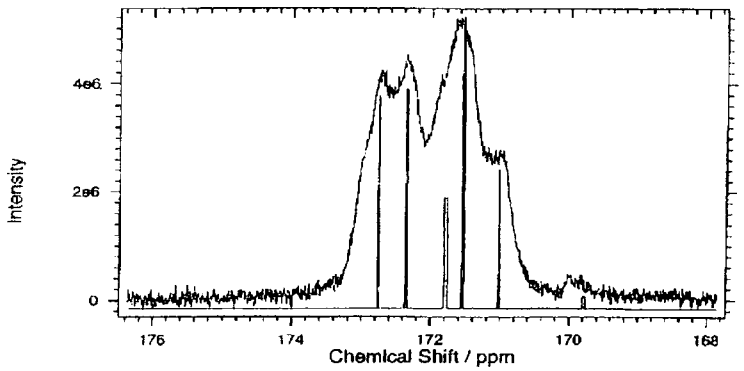


Figure 5.15 90 MHz ^{13}Cspectrum of carbonyl region with Memsys5 deconvolution overlayed.

Table 5.16 Peaks in the Memsys5 result for the Carbonyl Region of the ^{13}C Spectrum

Chemical shift /ppm	Intensity (% of overall cumulant)	Assignment
169.8	1.3 +/- 0.3	Unknown Minor impurity
171.0	13.9 +/- 0.4	P1
171.5	29.1 +/- 2.5	P2
171.8	11.1 +/- 2.7	Unknown possibly Succinic Anhydride
172.4	22.0 +/- 0.6	P3
172.8	21.3 +/- 0.6	P4

The peaks of interest to the discussion of stereochemistry are those at 171.0, 171.5, 172.4 and 172.8ppm. These four peaks, attributed to carbonyl groups in the SMS triads, can be divided into two groups corresponding to the two different carbonyl environments (different both in terms of substituent effects [2,3] and stereochemistry). Hence, peaks P1 and P2 are grouped together, likewise peaks P3 and P4. Note, the **total** intensity within each group is the same.

Peaks P1 and P2 can be attributed to the carbonyl group MA1 of Figure 5.17 on the basis of the greatest steric effect. The effect is seen in both chemical shift and stereochemistry, the latter being indicated by the greater effect on the relative intensities of the peaks. Therefore, P3 and P4 are assigned to carbonyl group MA2. The relative intensities of P1 and P2 indicate that when a styrene unit adds to a polymer with a maleic anhydride unit at its growing terminus it does so with a preferred relative configuration. The preferred configuration accounts for 68% of the total. For the addition of a maleic anhydride unit to a styrene terminus there seems to be little preference.

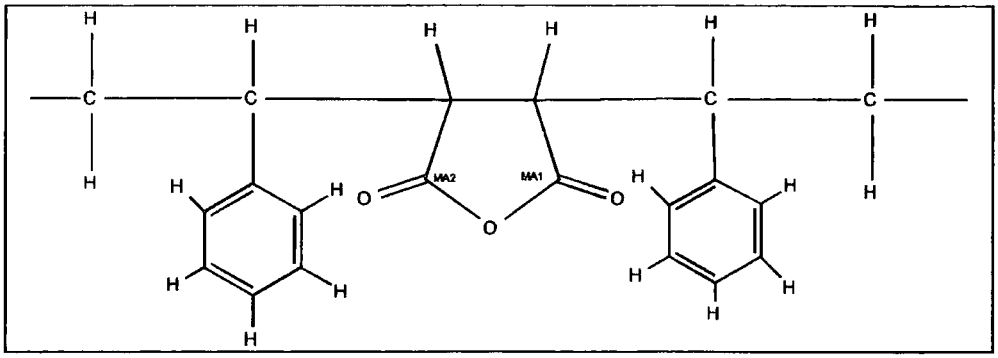


Figure 5.17 Styrene / maleic anhydride showing assignment of carbonyl groups

5.2.3.4.3 Aliphatic region of the spectrum

To derive microstructural information from this region of the spectrum requires the use of spectral editing techniques[1]. In the work presented here, this region is compromised by the large carbon signal from the solvent and the poor sensitivity. This illustrates the advantage of the approach presented here: the aromatic C1 region is not compromised by overlapping peaks or poor sensitivity. Furthermore, the microstructural information is obtained without the need for spectral editing.

5.2.3.5 125 MHz ^{13}C Spectrum

The failure of the Memsys5 results at 90 MHz to confirm the re-assignment of the polymer block ends was a consequence of the lack of resolution at this magnetic field strength. Acquiring the spectrum at the higher field strength may consolidate the block end re-assignment and also help to clarify the polymerisation model.

Initial attempts at acquiring a spectrum on the new DRX500 spectrometer at Unilever Research proved troublesome. A large spectral discontinuity in the C1 aromatic region of the ^{13}C spectrum made design of PSF parameters impossible. The source of the discontinuity was eventually traced to filter breakdown and the troublesome part replaced by the manufacturer. However, the spectrometer time available for this work was limited due to other demands on the

spectrometer and the signal : noise ratio observed in the final spectrum is not what was originally hoped for (see Figure 5.18). Nevertheless, the gross spectral features are consistent with those observed at 90 MHz.

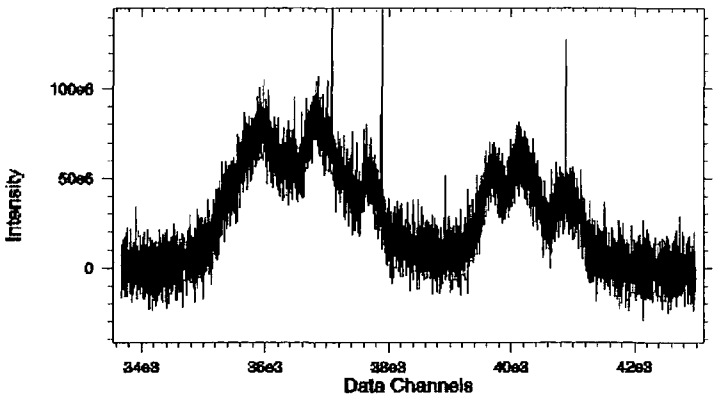


Figure 5.18 125 MHz C1 Aromatic Region

The similarity of this spectrum with that acquired at 90 MHz is initially surprising given the extra dispersion at the higher field strength. This lack of resolution is highlighted if the PSFs used at 90 MHz are compared with that at 125 MHz.

Table 5.19 Comparison Of PSF Widths At Different Fields

^{1}H Frequency / MHz	PSF Width / ppm
90	0.65
125	0.70

In order to establish if this difference in PSF width was significant the 125 MHz spectrum was reprocessed with the narrower PSF. The narrower PSF resulted in each of the end groups

starting to split into two peaks and a subsequent increase in the error bars associated with the Memsys5 intensities. Therefore, this difference in width is believed to be genuine and suggests that the higher field strength will not give the insight into the polymerisation process that was at first imagined.

It should also be noted that there are a number of very sharp peaks evident in the later 125 MHz spectrum which are probably due to monomer or lower molecular weight oligomers, suggesting that during the time evolved between acquiring the spectra at different fields the polymer had started to degrade. These sharp peaks, although only very small in terms of overall intensity, can lead to an underestimation of the applied PSF width - suggesting that the true PSF may be slightly larger than the 0.70 ppm quoted.

The difference in NMR line-width between the monomer and polymer peaks manifests due to the T_2 contribution to the line-width. T_2 is the time constant that describes the decay of transverse magnetization due to a loss of phase coherence between the nuclear spins. The T_2 value gives information about the distribution of resonant frequencies and about the local fields experienced by the magnetic moments of the nuclei. The local fields are related to the structure and nature of the local magnetic environment around the nucleus. As the local magnetic field in low molecular weight molecules fluctuates very rapidly and can average to zero, the internal local fields are weak and yield long T_2s or narrow resonant lines. This is the case for the monomers. The atoms in solids or highly viscous polymer solutions are in nearly fixed positions, and the internal magnetic fields are large resulting in a rapid loss of phase coherence. Therefore, the T_2s in polymers are very short and the resonance lines broad. It is worth noting that there is an additional contribution to T_2 which is not molecular in origin - the rate of decay of transverse magnetisation is influenced by the inhomogeneity of the external magnetic field. The experimentally observed T_2^* is the sum of the internal molecular T_2 and the contribution resulting from the non-uniformity of the applied field.

T_2 should be roughly independent of the applied magnetic field strength; if anything it should increase. The lack of extra dispersion at 125 MHz is probably due to the effect of the disorder in groups relative remote from the site of interest becoming more apparent in the spectrum.

For comparison the 125 MHz spectrum has been processed with Memsys5. The calculation of polymer composition and attempts at fitting to Markovian statistics are reported below. It is interesting to note that the 125 MHz spectrum was acquired with a much greater digital resolution in the hope of improving the quality of the deconvolution, i.e. more points per Hz. However, this had the drawback of making the spectrum un-useable with the MassInf algorithm; the time involved for any one iteration of the algorithm proved prohibitive.

5.2.3.5.1 Memsys5 result and calculation of polymer composition

The intensities of the Memsys5 derived peaks are given in Table 5.20, they are expressed as a percentage of the total cumulant.

Table 5.20 125 MHz Memsys5 Intensities

Group	Intensity	Lit. Assignment	Reversed Assignment
1	23.0280 +/- 0.6008	MSM	MSM
2	9.8819 +/- 0.49711	SSM	MSS
3	9.99365 +/- 0.56616	MSS	SSM
4	54.1325 +/- 2.5374	SSS	SSS

Substituting values into equations{i} and {ii} gives:

% Styrene = 74.6 +/- 6.6

This is again in good agreement with the Aldrich value of 75% . The error bars are much larger than for the 90 MHz spectrum. This is probably due to the poorer sensitivity of the 125 MHz spectrum.

5.2.3.5.2 First Order Markovian Model

Substituting the values from Table 5.20 into equations {iii} and {iv}, again for both block end assignments, indicates that the probabilities $P_{M/SS}$ and $P_{M/MS}$ are not equal. Therefore, the polymerisation model does not fit a first order Markovian model and is consistent with the results from the 90 MHz deconvolution.

5.2.3.5.3 Second Order Markovian Model

Again substituting the values from Table 5.21 into equation {v} and allowing two standard errors on the peak at 141.91ppm, the styrene content can be calculated if a second order Markovian model is assumed. The styrene content is presented in Table 5.21 for both end group assignments.

Table 5.21 Polymer Composition Based On Second Order Markovian Statistics

Number Of Standard Errors On Peak At 141.9ppm	% Styrene With Literature Assignment	% Styrene With Reversed Assignment
-2	73.7	73.6
-1	74.2	74.1
0	74.7	74.5
+1	75.2	75.0
+2	75.7	75.4

From Table 5.21, it is clear that either block end assignment could give the desired composition and the deconvolution thus fails to give conclusive support to the stereochemical argument for the reversed assignment. In this context the results obtained at higher field have proved somewhat disappointing. Furthermore, the higher field results indicate that the intensities of the block end peaks are similar, which is in agreement with Bhuyan and Dass.[2]

5.2.3.5.3 SSS Triad Region

The error bars associated with this deconvolution are much larger than for the 90 MHz spectrum and it is difficult to draw any conclusions as to whether the peaks are still in the same constant ratio.

5.3 Conclusions

^{13}C NMR, in conjunction with probabilistic data processing, has been successfully used to determine the composition of a styrene / maleic anhydride copolymer. Reproducible results have been derived, using both the Memsys5 and MassInf algorithms, for spectra acquired at two different

magnetic field strengths. The results are consistent with those reported by Aldrich and some in-house mass-spectrometry data.

Given the poor quality of the spectra, in terms of both sensitivity and resolution, the level of extra information that can be recovered this type of probabilistic data processing is encouraging. For the 90 MHz spectrum the internal consistency demonstrated by comparison with the deconvolution of the carbonyl region, and the agreement with the literature composition, supports the conclusion that the extra structure observed in the spectra is real.

The polymerisation method has been shown to be not consistent with a first order Markovian model. Furthermore, it has been demonstrated that a second order Markovian model cannot account for the intensity observed in the spectra. It has been postulated that the intensity of the eight peaks observed in the SSS region of the spectrum can be described by simple empirical expressions.

A literature assignment has been reversed based on a stereochemical argument. The intensity results, even at the higher field, have failed to confirm either assignment. It has been demonstrated that acquiring data at higher magnetic field strengths is not necessarily an advantage as far as assigning assignment peaks from a copolymer is concerned and strengthens the case for data processing / deconvolution techniques.

The MassInf algorithm has proved un-useable for large datasets or spectra which require many trials for design of the optimum PSF.

References:

1. D J T Hill, J H O'Donnell, P W O'Sullivan, *Macromolecules*, **19**, 9-17, 1985
2. K Bhuyan and NN Dass, *Indian Journal of Chemistry*, **29A**, 376-378, 1990.
3. B E Buckak and K C Ramey, *Polymer Lettera Edition*, **14**, 401-405, 1976.
4. Sigma-Aldrich Catalogue entry 20063-8.
5. Private communication between Author and Aldrich Chemical Co.
6. Microanalysis by Butterworth Laboratories Ltd., Ref.: BL1/008.
7. Unilever Research Port Sunlight Laboratory Notebook 8306, p60.
8. A E Tonelli, *Macromolecules*, **16**, 604-607, 1982.
9. Copolymerisation, vol. XV111, Ed. G.E.Ham, John Wiley & Sons, 1964.

CHAPTER 6:

The application of Linear Prediction and Maximum Entropy data processing to NMR spectra containing a range of peak widths.

The ^{27}Al NMR spectra of aluminium chlorohydrate systems

Using the MaxEnt Solutions Ltd. implementation, Maximum Entropy data processing is limited to spectra that contain peaks of similar width. This chapter describes the application of linear prediction techniques as a precursor to Maximum Entropy deconvolution. The aim is to use linear prediction methods on a spectrum containing peaks of very different widths with a view to producing sub-spectra containing only peaks of similar width. The sub-spectra are then suitable for Maximum Entropy data processing.

The ^{27}Al NMR spectra of aluminium chlorohydrate solutions are examples of a system that exhibits peaks of very different widths. Linear prediction is used to produce a sub-spectrum containing only the sharp peaks or the broad peaks. The peak areas are maintained. An auto-program has been written to enable many hundreds of these spectra to be processed more timely and more reproducibly than a manual method of integration. The results of the auto-program are compared with the previous method of manual integration.

The sub-spectrum containing the broad peaks is processed with a Maximum Entropy algorithm. Three peaks are identified in the tetrahedral aluminium region of the spectrum. An attempt is made to correlate this extra structure with sweat reduction and hence antiperspirant efficacy.

6.1 Introduction

As shown in the earlier chapters of this thesis, Maximum Entropy data processing has been successfully applied to spectral deconvolution. It has also been shown, in Chapter 4, that the method can be used to improve the signal to noise ratio observed in a spectrum. This has been achieved without the loss of resolution that is normally observed with more conventional methods, e.g. window functions.

Apart from a few standard instructions the main inputs to either the Memsys5 algorithm or the newer MassInf algorithm are the point spread function (PSF) and an estimate of the noise level present in the data (Sigma). It has been recognised, throughout the earlier chapters, that the main limitation of the Maximum Entropy techniques is the fact that only one PSF can be used for each Maximum Entropy deconvolution. Therefore, the successful application of this type of data processing has relied upon the fact that one PSF can adequately describe all the peaks of interest in the spectrum.

Normally, if a peak is present in the spectrum that is wider than the chosen PSF the Maximum Entropy result will be split into more than one peak. Conversely, if a peak is present that is slightly narrower than the applied PSF, the algorithm will still attempt to fit the peak but the corresponding Maximum Entropy result will carry significantly increased errors. Provided that the peak width variation across the spectrum is not too large an average PSF has been shown to suffice for most applications. Although, as described in Chapter 3, the average PSF may be optimized from the program's output diagnostics, it is clear that for many NMR spectra a single PSF is not appropriate.

Methods are available to try to overcome this limitation of using one PSF. For example, if the variation in peak width across a spectrum is known to be a systematic function of frequency and can be mathematically modeled, it is possible to regrid the spectrum onto a different x-axis. The peak-widths are then forced to be similar and the peak intensities are maintained. If the difference in peak-width across a spectrum is large this regridding process results in either peak compression or

expansion. In the extreme, this can result in one peak expanded such that the interval between successive points is large compared with the peak-width itself, i.e. step functions are apparent between the points, and at the other end of the spectrum a peak described by an unnecessarily large number of points. Any Maximum Entropy deconvolution will then be compromised as the expanded peaks are likely to fail the Deconvolution Criterion (See Chapter 3), i.e. there will be an insufficient number of points describing a composite peak to enable a successful deconvolution. Rather than NMR spectra, this type of regridding is better suited to chromatography data, where the peak widths can be modelled as a function of retention time. Many examples of chromatograms which have been regridded onto a different x-axis are available in the literature. [1]

This PSF limitation has been highlighted in earlier work [1] and in a poster presented by Ebbels et al., on work carried out primarily by Ebbels but in conjunction with L.P.Hughes. [2] For a range of simulated bands that were not overlapped, with the same peak shape but different widths, Ebbels attempted to fit the spectra with a variety of PSF widths. He compared the Memsys5, MassInf and GIFA [3] algorithms. Ebbels concluded that lines wider than the PSF were reconstructed as clusters of thin lines, and for peaks that are thinner than the applied PSF the reconstruction was very poor leading to highly structured residuals. Ebbels suggests that, for these peak-width tests, Memsys5 and MassInf produced residuals below the noise, whereas for GIFA, the residuals were comparable with the noise level or greater than the root-mean-square (rms) noise-level. This suggests that the two MaxEnt Solution's algorithms, Memsys5 and MassInf, were capable of producing a better fit to the data for the same input PSF. It is noted that for spectra with a low signal:noise ratio MassInf generally produces integrals with smaller error bars than direct integration methods. It is worth remembering that the work presented by Ebbels was performed on simulated NMR spectra. However, he concludes that sophisticated methods such as Maximum Entropy can perform better than traditional methods for analyzing NMR spectra.

In the work presented in this thesis, an attempt has been made to overcome the limitation of using only one PSF for genuine NMR spectra. This is achieved by separating the peaks of similar widths into sub-spectra whilst maintaining their correct intensities.

One method of achieving this peak separation based on width is to remove the more quickly decaying components from the free induction decay (FID) and then to backward linear predict the FID based on the more slowly decaying spectral components. This backward linear prediction will then produce a sub-spectrum containing only the sharpest peaks. If correctly scaled, this can then be subtracted from the original spectrum to produce a further sub-spectrum that contains only the broad peaks.

Backward linear prediction is a well-established NMR technique.[4] It has been used routinely for the removal of rolling baselines and for predicting that part of the FID that has been lost during the radiofrequency pulse and receiver dead-time. The application of linear prediction to produce sub-spectra containing peaks of similar width as a precursor to Maximum Entropy processing is novel.

Specifically, this chapter describes the application of linear prediction to the solution-state aluminium NMR spectra of aluminium chlorohydrate systems. The technique of regridding is not appropriate in this case because the variation in peak width is not uniform across the spectra.

In this application, the linear-prediction technique is used to produce a sub-spectrum containing only broad peaks and a second sub-spectrum containing only sharp peaks. The spectral intensities are maintained. The sub-spectrum containing the broad peaks is then processed with a Maximum Entropy algorithm. The linear prediction program has been automated and a comparison with a manual method for quantifying the species present in aluminium chlorohydrate systems is discussed.

6.2 Background to aluminium speciation problem

6.2.1 Line broadening mechanisms in NMR spectroscopy

The reader is directed to Chapter 2 for a description of the possible origins of line-broadening which are important for the aluminium chlorohydrate systems studied in this chapter.

6.2.2 Aluminium Speciation

The Unilever Research, Port Sunlight, method for determination of aluminium speciation in aluminium chlorohydrate and related species by solution-state NMR has been summarised by Lee[5]. The background and manual integration method is briefly described below for comparison with the linear prediction method.

Aluminium chlorohydrate (ACH) and activated aluminium chlorohydrate (AACH) are used as high-efficacy aluminium antiperspirant actives. Their synthesis and characterisation have been described by Nazar et al.[6,7] The antiperspirant performance of these actives is modified by manipulation of the inorganic species present. ACH contains three types of structural unit that can be identified by ^{27}Al solution-state NMR. Al_{13} (Keggin ion) is a stable intermediate which undergoes activation by loss of one Al^{3+}. This results in a defect site producing the Al_{12} species. Dimerisation of Al_{12} can then occur to form Al_{24}. Higher molecular weight 'polymers' are also possible and are observed in the NMR spectrum as broad lines. The full synthetic process produces the following species (presented in order of increasing relative molecular mass):

Al^{3+} ... $[Al_{12}]$ ... Al_{13}^{7+} ... AlP1 ... AlP2 ... AlP3 ... higher polymers

AlP_1 is Al_{13} with one triad rotated by 60°. The P designation refers to a polymeric type with the relative molecular mass of the polymer increasing in the order P1 > P2 > P3.

The synthesis involves reacting NaOH with $AlCl_3$ at 95 °C. This reaction produces mainly Al_{13}^{7+} and a few $[AlO_6]$ monomers. Further heating produces the higher molecular weight polymers which are needed for anti-perspirant activity and are found in a typical ACH solution (Figure 6.1).

The precise structure of all these polymers is not known. Recent work by Allouche et al.[10] suggests that thermal treatment of an Al_{13} solution produces a cluster of two Al_{13} Keggin units connected by a ring of four octahedral $[AlO_6]$ units and is called Al_{30} or AlP2. It is obvious that

there are a large number of species present in solution and a full description is beyond the scope of this work; of greater importance to the antiperspirant industry is to limit the polymer size.

The activated system refers to the depolymerisation of the very high molecular weight ACH species to produce smaller molecules that are thought to be more efficacious because they can diffuse more readily to the sweat duct and provide a more effective blockade upon interaction with sweat components.

Based on the size of the species present in ACH it is believed that the AlP2 has the optimum molecular weight for sweat reduction. The quantification of ALP2 and the lower molecular weight species is an important analytical measurement for the optimisation of antiperspirant actives. Many NMR measurements have been carried out at Unilever, Port Sunlight, in an attempt to quantify the level of AlP2, Al^{3+} and Al_{13}^{7+} in solution and to try to relate these measurements to sweat reduction figures for a particular antiperspirant formulation.

Figure 6.1 summarises some of the aluminium species found in a typical ACH solution.

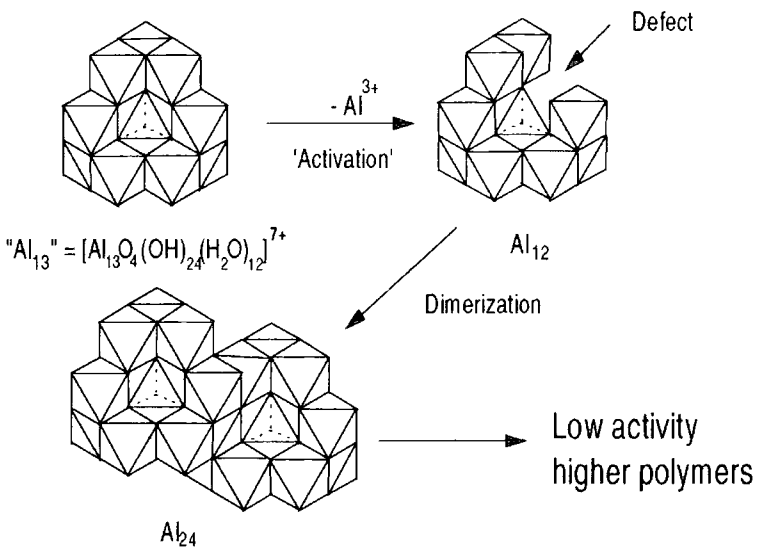


Figure 6.1. Some of the structural units found in ACH/AACH (AlP2 not shown).

(Figure courtesy of K. Gosling, Unilever)

The single Al nucleus of Al^{3+} gives a resonance that should allow quantification of all the Al in this form. Al_{13} and AlP2 contain Al nuclei in two environments, i.e. tetrahedral and octahedral. The tetrahedral signals are sharp enough to allow good quantification of the total tetrahedral intensity by manual integration methods. However, the octahedral signals are too broad to allow good quantification. In order to quantify all of the aluminium in the form of Al_{13} and AlP2, the corresponding tetrahedral signals are determined and a scale factor is applied to the octahedral region based on the known structure of each of these species. Al_{13} has one tetrahedral and twelve octahedral environments, AlP1 has one tetrahedral and twelve octahedral aluminium atoms and based on the work by Allouche et al. [10] AlP2 has two tetrahedral and twenty-eight octahedral aluminium environments. This accounts for the Al in the octahedral environments. (Section 6.2.3)

Table 6.1 summarises the NMR assignments for the low molecular weight species present in ACH/AACH. These peaks are used in the determination of aluminium speciation, together with the sharp signal at approximately 80 ppm from a sodium aluminate reference capillary.

Table 6.1 NMR activity of Al species in ACH and related materials

(# Relative to Al^{3+} which has been assigned to 0 ppm)

Species	Al sites * No.	Chemical Shift / ppm #
Al_{13}, Keggin ion	Tetrahedral * 1	~60, Sharp
Al_{13}, Keggin ion	Octahedral * 12	~10, Broad
AlP2	Tetrahedral*2	~70, Broad
AlP2	Octahedral*28	~10, Broad
Al^{3+}	Octahedral *1	~0, Sharp
Higher polymers	-	-

Figure 6.2 is the ^{27}Al spectrum of a typical AACH sample. The spectrum was acquired on a Bruker DRX500 NMR spectrometer equipped with an aluminium-reduced probe. The level of backgroud probe signal was measured and does not make a significant contribution to the following calculations for a typical antiperspirant active. The normal total aluminum level in these systems is 15 wt% by XRF.

10mm quartz NMR tubes (Wilmad 513-7PP QTZ) and 5mm quartz NMR tubes (Wilmad 528-PP QTZ) for the capillary were used to minimise the possibility of interfering aluminium resonances. A known weight of ACH was dissolved in about 5 cm^3 H_2O and made up to the mark in a volumetric flask. About 3 cm^3 of this solution was transferred to the 10 mm NMR tube. A known weight of sodium aluminate was added to 5 cm^3 of D_2O and this was transferred to the 5 mm NMR tube. The smaller tube was place inside the 10 mm tube, acting as a capillary. It was held in place using a 10 mm vortex plug. A direct polarization experiment with 2000 transients was used, spectral width 50 000 Hz, pulse-width 7 μs (i.e. ~50° pulse), and an interpulse-delay of 0.2 s. An exponential line broadening of 1 Hz was applied before Fourier transformation. In all cases the solvent used was H_2O with the D_2O, used in the internal sodium aluminate capillary, necessary for spectrometer locking.

It is clear from Figure 6.2 that there is a significant difference in the linewidths present in this spectrum. A representative PSF could not be found for direct Maximum Entropy processing.

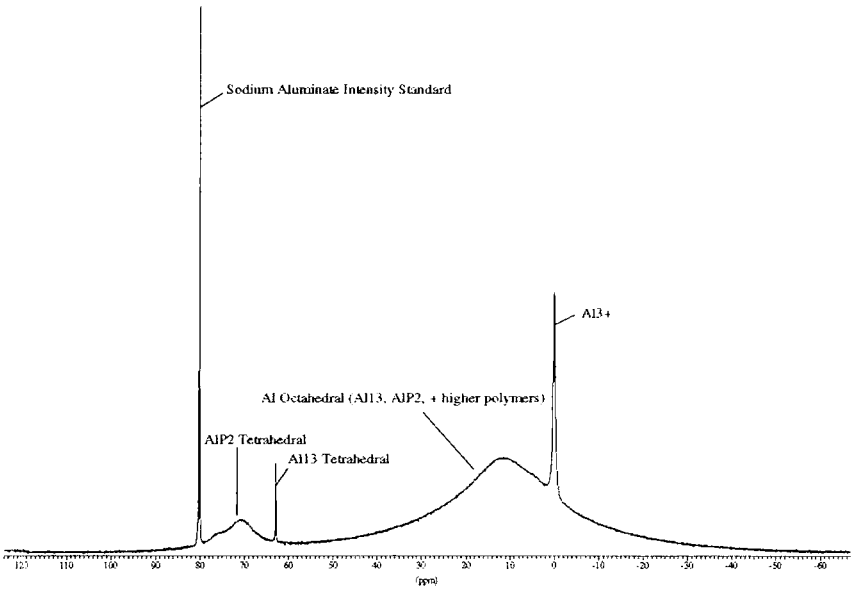


Fig.6.2 ^{27}Al spectrum of AACH

6.2.3 Quantification of aluminium speciation by manual integration.

Determination of aluminium speciation by a manual method of integration has been carried out for many years at Unilever Research, Port Sunlight. The existing method was described by Lee[7] and is summarised below.

The quantification of the aluminium species relies on the use of an internal standard of known concentration of sodium aluminate held in a 5 mm capillary within the ACH solution. The intensity standard is calibrated against a primary standard of aluminium nitrate (> 98% ex Aldrich Chemical Co.) as shown in Figure 6.3.

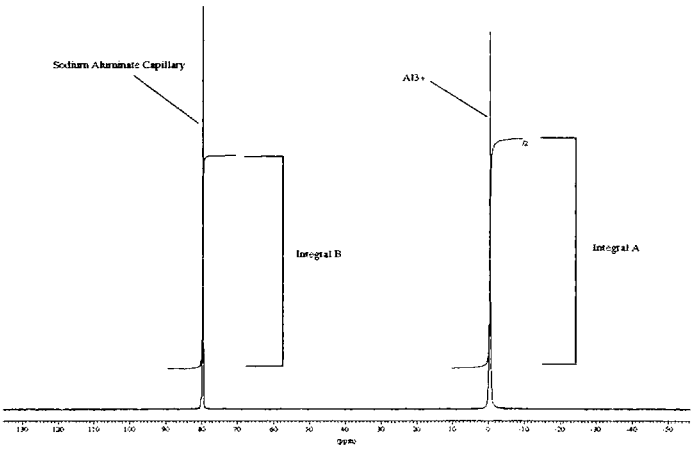


Figure 6.3. Capillary calibration

The number of moles of Al represented by the signal from the capillary (Al_{cap}) can be determined by equation {1}:

$$Al_{cap.} = B/A \times (M/375) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\{1\}$$

where A and B are the integrals corresponding to the aluminium nitrate (Al^{3+}) peak and the sodium aluminate peak respectively (Figure 6.3), M is the mass of the aluminium nitrate used and 375 is the molecular weight of the aluminium nitrate nonahydrate.

Figure 6.4 shows the integrals required to determine the number of moles of aluminium in the form of Al^{3+} and Al_{13} by means of equations {2} and {3}:

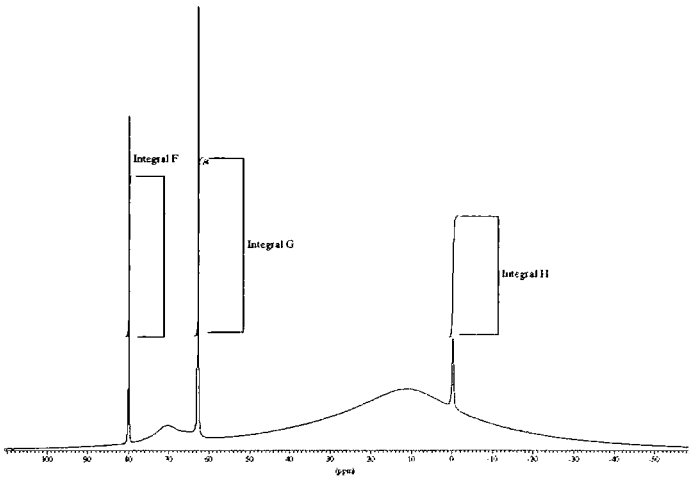


Figure 6.4. Integrals Required For Determination of Moles Al^{3+} and Al_{13}

Moles of aluminium in the form of Al^{3+} = (H/F) x $Al_{cap.}$ .....{2}

Moles of aluminium in the form of Al_{13} = (G/F) x $Al_{cap.}$ x 13 .....{3}

where F, G, H are the integrals of the sharp resonances shown in Figure 6.4 and $Al_{cap.}$ is calculated from equation {1}. The factor 13 in equation{3} accounts for the fact that for each tetrahedral aluminium contributing to integral G there are 12 octahedral aluminium atoms.

Determination of the number of moles of aluminium in the form of AlP2 requires the integrals shown in Figures 6.4 and 6.5. The total tetrahedral content is given by the intensity of (K-L). L is an arbitary integral offset introduced to make it easier to distinguish the integral from the spectrum. There is some uncertainty in this measurement due to overlap with the sloping baseline from the broad octahedral resonances. From this is subtracted the intensity due to the capillary (J) and the contribution due to Al_{13}. In some samples, the integral of the tetrahedral Al_{13} is very small compared with the integral of the tetrahedral AlP2, for example, and Lee comments that the Al_{13} integral can be difficult to measure directly from Figure 6.5. It is calculated as (J x G/F).

The number of moles of Al in the form of tetrahedral AlP2 is then given by:

(K-L-J- (J x G / F)).

Hence, the total number of moles of Al in the form of AlP2 is given by:

Moles of Al in the form of AlP2 = (K-L-J-(J x G/F)) x 15 x Al(capillary)/J .....{4}

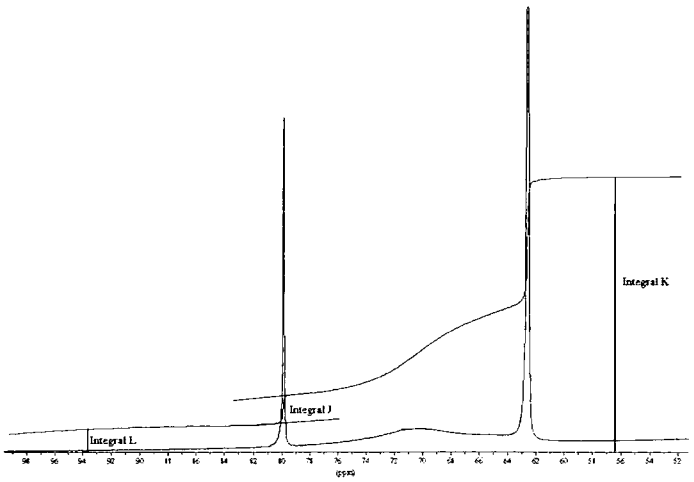


Figure 6.5. Determination of Total Tetrahedral Aluminium

The factor 15 arises from the fact that for each AlP2 there are two aluminium atoms in tetrahedral environments and 28 in octahedral environments.

Lee concludes that the total moles of aluminium based on speciation present in a AACH is calculated as:

$$\% \text{ Recovery} = \frac{\text{Moles } Al^{3+} + \text{Moles } Al_{13} + \text{Moles AlP2} \times 100 \times 27}{(\text{mass of sample} \times 25)} \quad .....\{5\}$$

Equation {5} is based on the assumption that there is nominally 25% aluminium in the sample.

Aluminium spectra have been recorded for many hundreds of these ACH systems and until the implementation of the automatic linear prediction and integration methods described in this report they have all relied upon the manual methods described above. The integrals are measured with a ruler and the results entered into a Lotus 123 spreadsheet for calculation of the levels of each aluminium species present.

6.2.4 Limitations of the manual method of integration

Lee points out that there may be additional structure on the peaks at ~70 ppm and ~10 ppm (described as AlP2). Indeed, evidence for this extra structure can be seen in Figure 6.2. There is a shoulder to higher frequency of the AlP2 resonance. He continues that one possibility is the presence of higher molecular weight polymers or polymers at least slightly larger than AlP2. The manual method of quantification cannot resolve this additional structure and in the absence of further information all of the intensity in the ~70 ppm peak is assigned to AlP2.

Other limitations lie in the measurement of the integrals themselves. The integral range for the sharp peaks is chosen to be as close as possible to the peak base (estimated by using 'normal' vertical expansions). The aim of this is to minimise the contributions from the underlying broad resonances. This is clearly a source of error. The selection of the integral window is subjective and may not be consistently selected when comparing one sample with another. Furthermore, for the broad peaks, intensity will be lost outside the region over which the integration is performed and there is again the problem of consistency from sample to sample. These are known shortcomings of the method used for comparison.

The manual method of plotting and integration is also very time consuming. Three different plot expansions are required in an attempt to try to minimise the errors described above. The automatic linear prediction program described below alleviates the need for manual plotting, and consistently selects the same integral regions for the peaks of interest. Whilst the method may not be completely quantitative it should be more reproducible than the manual method and hence more useful for following trends.

One limitation for all these techniques is the length of the receiver 'dead time' on the high-resolution spectrometer. For typical ACH/AACH samples only 65% of the aluminium is observed, based on equation {5}. This is compared to the total aluminium content as determined by XRF measuements. Approximately 35% of the aluminium intensity is lost and this is believed to be associated with the higher molecular weight aluminium species. Neither Maximum Entropy nor linear prediction methods can recover these lost signals; solid-state NMR or solid–echo techniques would be required if these species were of particular interest. Work has been carried out by R.K. Harris, University of Durham, on the solid-state NMR spectra of Al_{13} (Keggin ion).[11]

6.3. Linear Prediction

6.3.1 Background

Linear prediction is a widely used data-processing method and relies on the principle that each value in a time series can be represented by a fixed linear combination of the immediately preceding values. This principle has been found to be true for the free induction decays found in NMR spectroscopy. The mathematics underpinning the method is complex. A summary can be found in references [4,8], but the basic linear prediction equation is:

$$d_k = \sum_{j=1}^{m} a_j d_{k-j}, \qquad k = m, \ldots, M-1 \quad \ldots\ldots\ldots\{6\}$$

In equation {6}, the values 'a_j' are the linear prediction coefficients, sometimes referred to as the prediction filter, and the number 'm' is called the order of the prediction (NCOEF).

For the work presented in this report the linear prediction algorithm is the standard implementation found in the Bruker XWINNMR program. [8] The key program inputs are:

- NCOEF. This is the number of coefficients used for the linear prediction calculation. This parameter is empirically found. According to Hoch etc al.[8] the recommended approach to

determining the number of coefficients is to use a number much higher than the expected number of peaks given the S/N.

- LPBIN. The number of points contributing to the backward linear prediction. LPBIN is determined empirically and is set to a value between one and the time domain data size.
- TDoff. The number of points to be predicted. This is again found empirically and depends on the number of points removed from the start of the FID and hence on the T_2 of the faster decaying components.
- NSP. The number of left shifted points, i.e. the number of points deleted before applying the linear prediction filter. It should be numerically equal to Tdoff.

The classical use of backward linear prediction in NMR spectroscopy is to predict the first few points of an FID which may be corrupted due to ringing in the analogue circuitry of the spectrometer or lost completely in the receiver dead time. The consequence of these corrupted or lost points manifests itself as baseline curvature. As is discussed below, in this application the first hundred or so points of the FID, i.e. NSP, are deliberately discarded to remove the broad, fast decaying components and then these are predicted back using only the slowly decaying data points as the reference for the linear prediction. For the ACH systems, currently studied at Unilever Research, Port Sunlight, the number of points required to completely remove all the quickly decaying frequencies has been found empirically to be 206. The number of coefficients used for the linear prediction was also found empirically. The value of 4096 has been used in this work. This was selected by choosing that value which produced the spectrum that showed least discontinuities.

6.3.2 Implementation of Linear Prediction for AACH Spectra

6.3.2.1. Digital filtration

The ^{27}Al spectra presented in this report have all been acquired on a Bruker DRX500 spectrometer using digital filtration. Amongst the advantages of digital filtration are the removal of

spectral artifacts and the ability to select small spectral widths without signals being folded back into the region of interest.

Digital filtration is achieved by convoluting the time domain data with a sine x /x (sinc) function. This has the effect of applying a square wave to the frequency domain. The very sharp edges of the square wave function ensure signals cannot be folded back.

The convolution of the sinc function with the FID creates a problem for conventional backward linear prediction. The resultant FID starts from zero and ramps to a maximum value within the first few hundred points. The FID then decays as is normally observed. The sinc-function is applied as the FID is acquired. This means the original FID, i.e. without the ramp, is not available to the operator. This ramp is clearly seen in Figure 6.6. Backward linear prediction cannot be successfully applied to this FID.

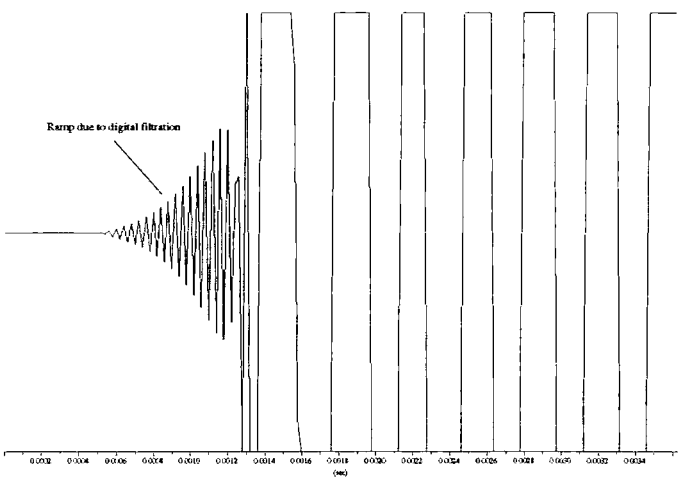


Figure 6.6 Effect of Digital Filtration on the FID.

However, this problem can be easily overcome by making use of the cyclic properties of the Fourier transform. The ramp can be conveniently moved to the end of the FID by making use of features associated with the Bruker data processing software. Fourier Transformation of a digitally acquired FID includes the application of a large phase shift. If the spectrum is then reverse Fourier

Transformed the data points of the resultant FID are cyclically shifted. The ramp is moved to the end of the FID.

The effect of this ramp can then be more easily removed either by applying a decaying exponential window to the FID (the time constant chosen such that the FID has effectively decayed to zero before the start of the ramp) or by reducing the time domain size. For this work, the decaying exponential was chosen such that a line broadening of 15 Hz was produced in the frequency domain. An article by Westler and Abildgaard[9] suggests that this approach may not be appropriate as it may produce 'frowns', a curling, at the edges of the spectra. There is some evidence that this may be the case for the work presented here, but this is not believed to be a problem for the quantification of the aluminium species and will be shown to have little effect when results are compared with the manual method of integration.

6.3.2.2 Autoprogram

The Bruker autoprogram function within XWINNMR provides a tool for adding your own functionality to the XWINNMR software. In this instance, the program performs a series of complex operations on the FID before automatically integrating, fixing the integral windows, and plotting the resultant spectrum. The autoprogram is written in 'C' language with an interface into XWINNMR. The program details will not be listed here but Figure 6.7 provides a flow chart that describes the basic program functionality. (See the Appendix for a program listing). The program relies on calls to standard parameter sets where constants such as NCOEF are stored. These parameter sets are an integral part of the program. Any subsequent fine-tuning of the program, e.g. if the aluminium system changes to include species which have different linewidths, can be accommodated by changing the constants in the parameter sets rather than the autoprogram itself. This makes the program more user-friendly in that the key program inputs will be recognisable to an experienced NMR operator.

The only subjective input at the time of running the program is the initial phasing of the spectrum. These phase parameters are carried through the program and are used to phase correct the

sub-spectra. Any phase errors will be easily recognised in the resultant sub-spectra, i.e. dispersion line-shapes are more easily recognised in the sub-spectrum of the sharp peaks, and a correction can then be applied to the original spectrum.

Figure 6.8 shows an example of the ^{27}Al spectrum of a typical AACH system and will be used to demonstrate the application of the linear prediction autoprogram.

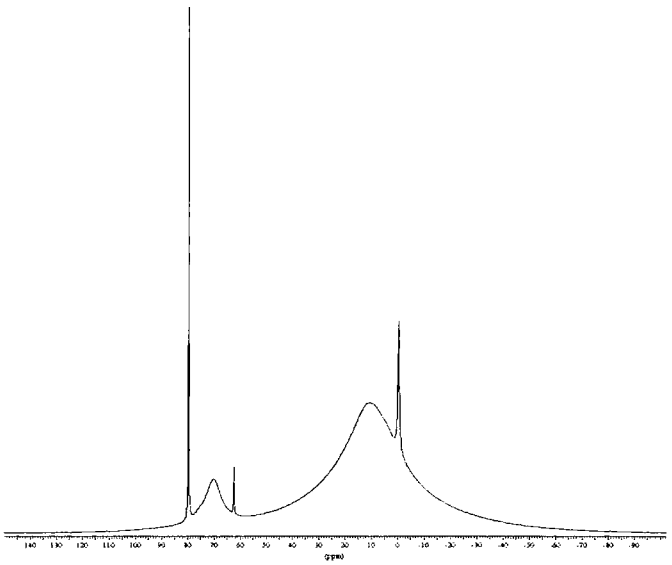


Figure 6.8. Typical ^{27}Al spectrum of AACH.

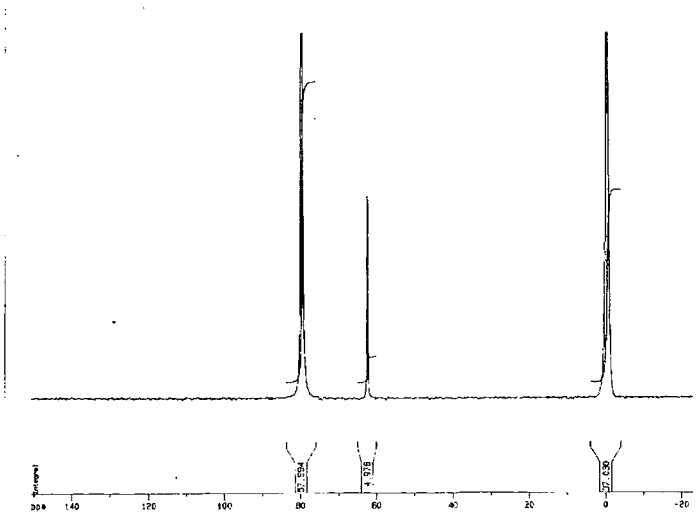


Figure 6.9. Sharp Sub-Spectrum following Linear Prediction showing Integral values. The vertical expansion has been increased to show the effective removal of all the broad peaks.

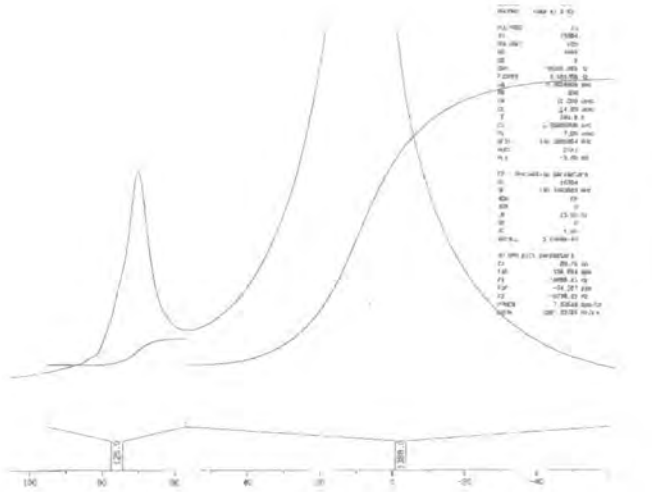
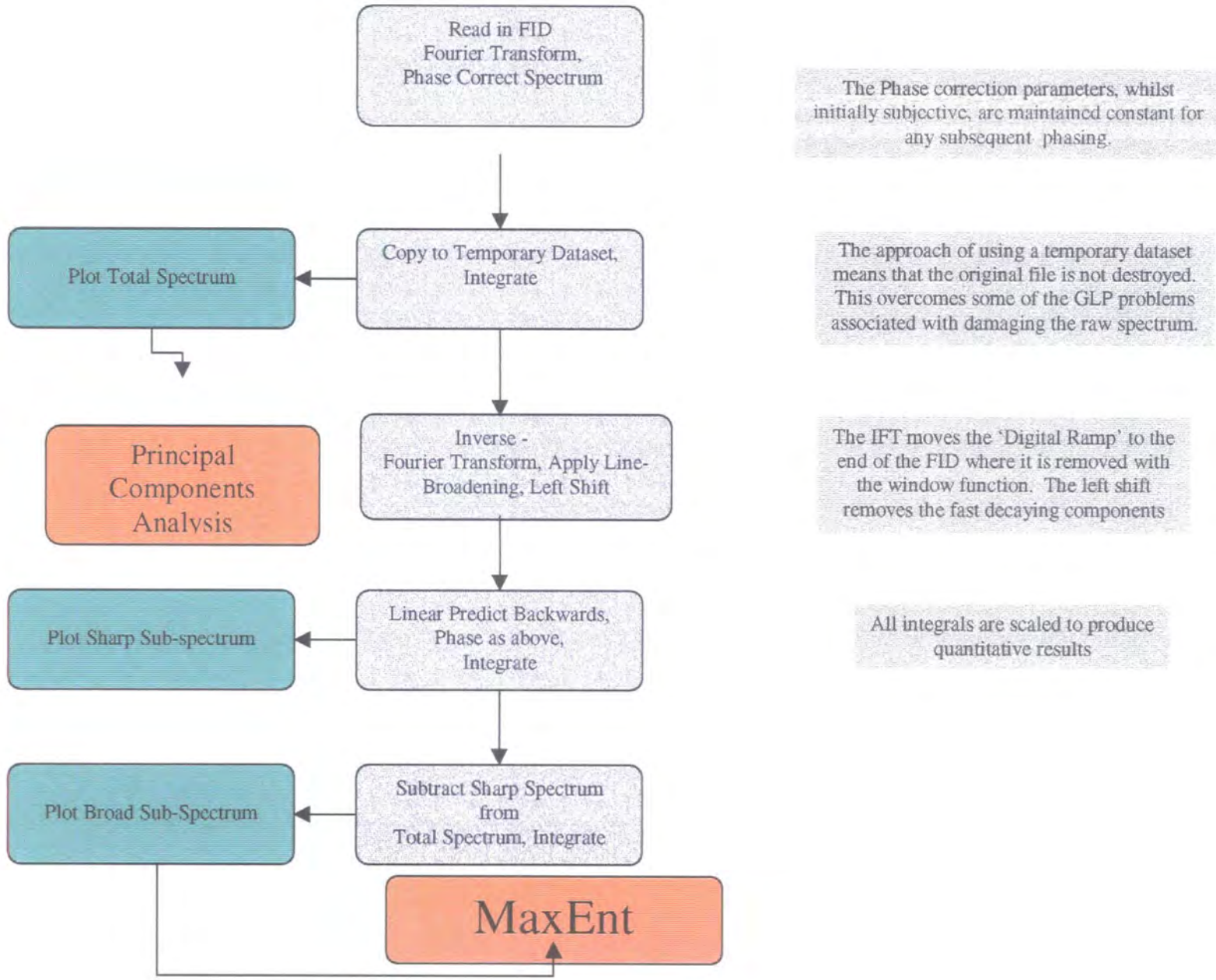


Figure 6.10. Broad Sub-Spectrum following Linear Prediction and Spectral Subtraction showing Integral Values. The vertical expansion has been increased to show that there is no evidence for the sharp peaks.

Figure 6.7. Flowchart Showing Autoprogram Functionality



Figures 6.9 and 6.10 show the sub-spectra derived from the linear prediction. The lack of discontinuities in the spectra indicates that a good separation has been achieved. The integral windows have been chosen such that the calculated levels of each aluminium species are similar to those already found using manual methods. This was considered to be the most appropriate method given the number of spectra that had already been processed and compared with sweat reduction figures. The size and position of the integral windows is consistent from one spectrum to another. The values are read directly from the numbers printed under the integral traces and typed into an Excel spreadsheet for calculation of, e.g. %AlP2.

The sub-spectra can now be more easily processed with a Maximum Entropy algorithm. (See section 6.4). Although the peak widths present in the broad spectrum are still somewhat different, (tetrahedral versus octahedral peakwidth) the sharper peaks have been completely removed and the tetrahedral peaks can be processed with a Maximum Entropy algorithm separately from the octahedral peaks.

Whilst processing many of these spectra it has become clear that there is a slight discontinuity in the Al^{3+} peak, at 0 ppm. The linear prediction parameters could not be optimised to accommodate all the sharp peaks in the spectrum because there is small difference in linewidth between the Al^{3+} peak and the other sharp peaks. As will be shown below this does not present too much of a problem when quantifying the aluminium species present.

6.3.2.3 Quantification of aluminum speciation using the autoprogram

Using the same arguments as described in section 6.2.3, the intensity standard is calibrated against a primary standard. The ^{27}Al spectrum, although it contains only sharp peaks, is still processed in exactly the same way with the linear prediction autoprogram. This ensures that the sodium aluminate integral is measured consistently from one spectrum to another. The calibration spectrum is shown in Figure 6.11 with the automatically scaled and printed integral values.

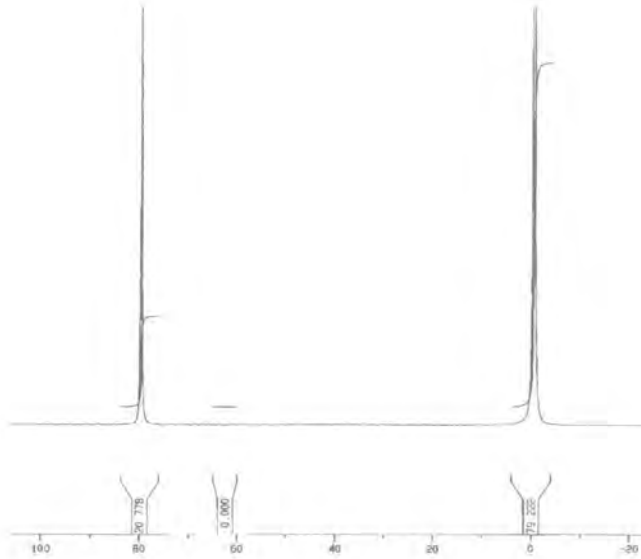


Figure 6.11. Capillary calibration spectrum processed with the autoprogram

Based on section 6.2.3,

$$Al_{cap.} = (f / g) \times (M / 375) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \{7\}$$

where f and g are the integral values shown in Figure 11 that correspond to the aluminium nitrate peak (Al^{3+}) and the sodium aluminate peak respectively. Then,

$$\text{Moles of Al in the form of } Al^{3+} = (c / a) \times Al_{cap.} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \{8\}$$

$$\text{Moles of Al in the form of } Al_{13} = (b / a) \times Al_{cap.} \times 13 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \{9\}$$

where a, b, and c are the integral values shown in the sub-spectrum of the sharp peaks, Figure 6.9. Finally,

$$\text{Moles of Al in the form AlP2} = \frac{d \times 12 \times Al_{cap.}}{a} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \{10\}$$

where d is the integral of the tetrahedral region shown in the sub-spectrum of the broad peaks, Figure 6.10.

6.3.3 Comparison with manual method of integration

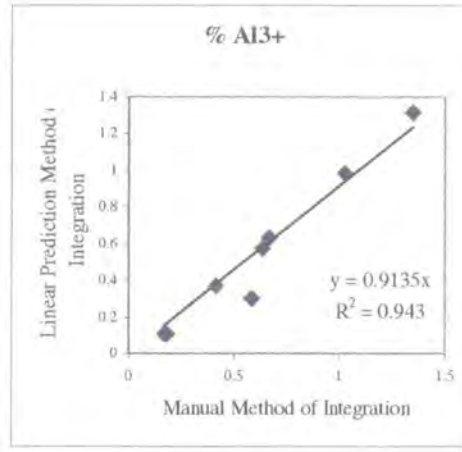
Table 6.2 shows the comparison between the results obtained using linear prediction and the manual method of integration for a typical AACH sample.

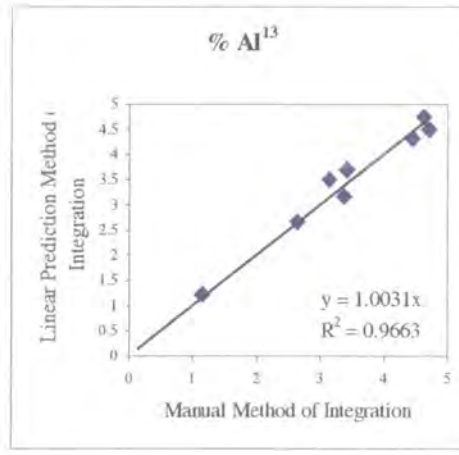
Table 6.2. Comparison of manual method and automatic integration

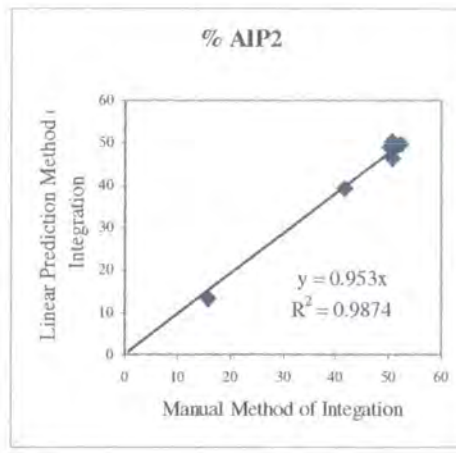
	% Al^{3+}	% Al_{13}	% AlP2
Linear Prediction Autoprogram	0.11	4.3	50.3
Manual integration	0.19	4.4	50.7

There is very good agreement between the two methods for this particular spectrum. The largest errors are associated with the Al^{3+} peak. This is to be expected given that the program is attempting to separate a very small sharp peak from a large broad peak. This failure to completely separate the peaks can be further explained by the slight variation in peak width seen for the sharp peaks. This gives rise to the spectral discontinuity described earlier. The relatively large difference in Al^{3+} intensity observed between the two methods is not a problem as far as antiperspirant activity is concerned, given the relatively small quantities present in these systems. It also has to be remembered that the linear prediction results may in fact be the true intensities. The manual method of integration is more likely to suffer from subjective errors.

In order to get a better idea of the differences between the two methods a variety of AACH spectra has been processed which covers the range of levels of each species found in the AACH systems. The results are compared in the following graphs.







The R^2 for each of the regression lines is better than 0.94, indicating that there is generally very good agreement between the two methods. Although the amount of each species may not be accurate, there is very good precision and hence the autoprogram can be used reliably for following trends in AACH processing.

The autoprogram is now used routinely at Unilever Research, Port Sunlight for the quantification of aluminium species in these systems.

6.3.4 ALP2 intensity versus sweat reduction

Aluminium spectra have been recorded for many hundreds of these ACH systems, but until the implementation of the automatic linear prediction and integration methods described in this report they have all relied upon the manual methods of integration. Nevertheless, a crude correlation has been found between the level of AlP2 in these systems and improved sweat reduction based on consumer testing. There are, however, discrepancies. For example, systems have been identified with a high AlP2 level, as measured by NMR spectroscopy, but which give a poor antiperspirant performance. The latest theories suggest that there may be more than one species contributing to the tetrahedral resonance which has previously been assigned to AlP2. This is supported by evidence from Harris's work on the solid ACH systems, which suggests there may be up to three different tetrahedral and octahedral sites.

P. Clarke, Statistics Unit, Unilever, has made an attempt to carry out a Principal Components Analysis of all the available ^{27}Al spectra in an attempt to correlate the intensity of each of the principal components with sweat reduction. This method reduces the many hundreds of spectra to a small number of pseudo sub-spectra that indicate the major differences that appear in the overall dataset. Preliminary results indicate that there are three different tetrahedral peaks present in these systems. The results of this analysis will be reported separately, but they add to the increasing evidence for extra structure in the tetrahedral region.

6.4 Maximum Entropy data processing

Maximum Entropy processing of spectra of this type has only been made possible by the reduction to sub-spectra that contain peaks of similar width. Whilst processing the ACH spectra with the linear prediction autoprogram it has become apparent that removal of the sharp peaks from the spectrum has made it easier to recognise, by eye, a number of unresolved peaks in the tetrahedral region of the spectrum. The intensity of these tetrahedral peaks appears to change as the ACH activation process is allowed to continue. The areas of these peaks may explain the discrepancies with the sweat

reduction figures. It may also enable the ACH processing conditions to be further refined so that the concentration of the efficacious AlP2 can be maximised. It is not necessary to process the sharp sub-spectra with a Maximum Entropy algorithm, all the peaks are well resolved.

The following figures are produced using the Maximum Entropy software and presented in data channels rather than ppm. This is because the conversion program does not recognise the chemical shift unit ppm. Nevertheless, it is clear from the spectra which peaks are due to tetrahedral aluminium and which are due to aluminium in an octahedral environment.

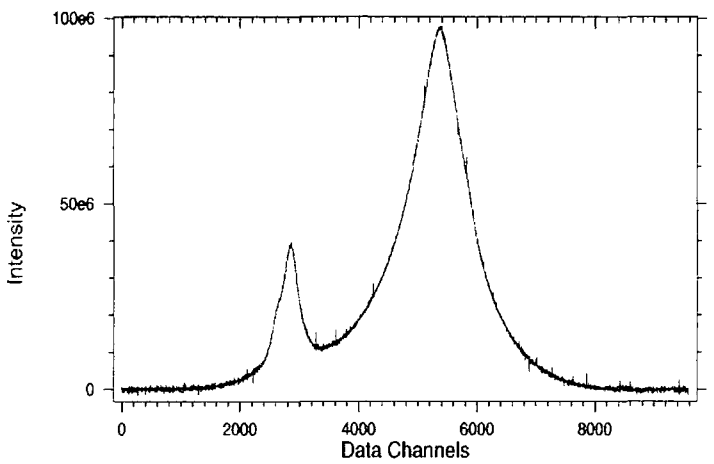


Figure 6.12. Broad Sub-spectrum showing structure on tetrahedral peak

Figures 6.12 and 6.13 show the shoulders that are apparent on the two of the sub-spectra. There is clear evidence for a peak to higher frequency of the main tetrahedral resonance, (See Figure 6.12), and also one to lower frequency. (See Figure 6.13).

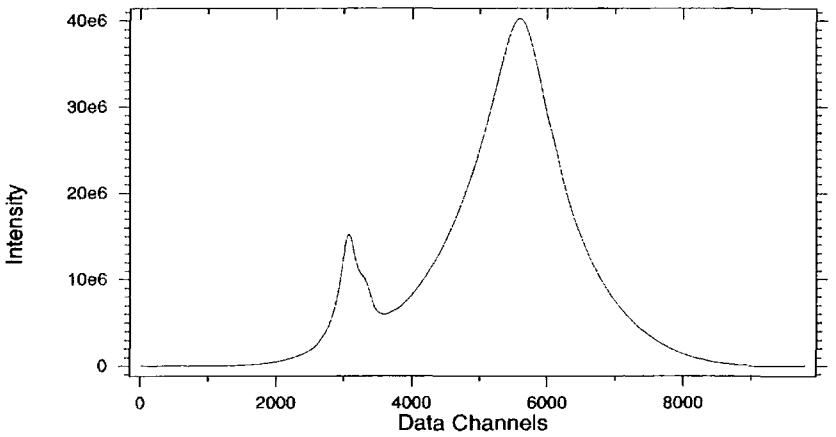


Figure 6.13. Broad sub-spectrum showing structure on tetrahedral peak

These spectra proved difficult to process with the Memsys5 algorithm. Firstly, optimisation of PSF parameters for the whole spectrum was difficult because of the large width difference between the tetrahedral and octahedral peaks. The wider octahedral peak biased the PSF width high. Secondly, the very high signal:noise ratio prohibited a MaxEnt deconvolution. For high signal:noise spectra the algorithm struggles to converge as it tries to fit a PSF to the spectrum within the noise. The problem is that with a range of linewidths, no single PSF gives residuals which are dominated by the noise.

The change in peak width was overcome by truncating the spectrum before the start of the octahedral peak, see Figure 6.14. The algorithm then 'pads' the data size to the next power of two so a fast Fourier Transform can be used. The PSF parameters could then be optimised, using the evidence values[1], based solely on the width and shape of the peaks due to the tetrahedral signals.

The optimum peak width was found to be 164 data channels, with a mixed peak shape of 85% Lorentzian and 15% Gaussian.

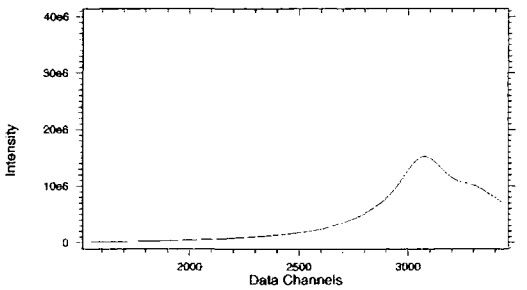


Figure 6.14. Truncated tetrahedral peak used for PSF design.

The failure of the algorithm to converge due to the high signal:noise ratio could not be overcome even by relaxing the convergence tests performed by the algorithm. (See Chapter 3). The high signal:noise ratio is due firstly to the large number of scans recorded and secondly to the application of a large line-broadening (15 Hz) during application of the linear prediction autoprogram. The large number of transients was acquired mainly for historical reasons to help with the manual integration.

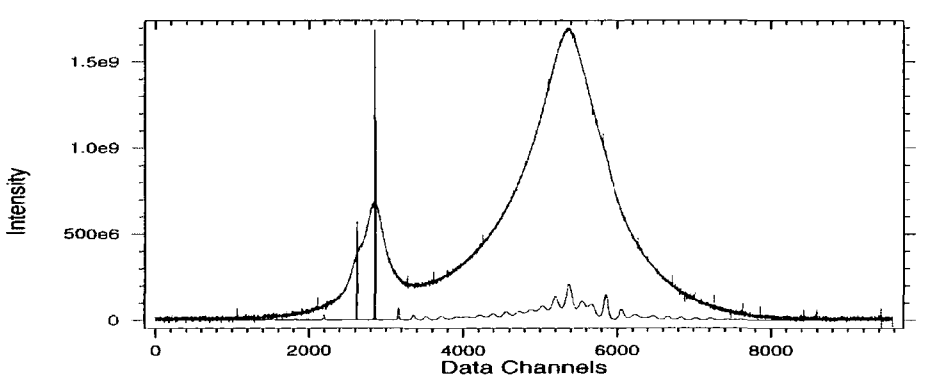


Figure 6.15. Spectrum of AACH overlaid with Maximum Entropy Result.

In order to continue with the deconvolution a small amount of random Gaussian noise was added to each spectrum. The level of noise added was found empirically to be insufficient to change the point of convergence of the algorithm but enabled it to converge. The Maximum Entropy results are shown Figures 6.15 and 6.16 overlaid with the noise-treated spectra.

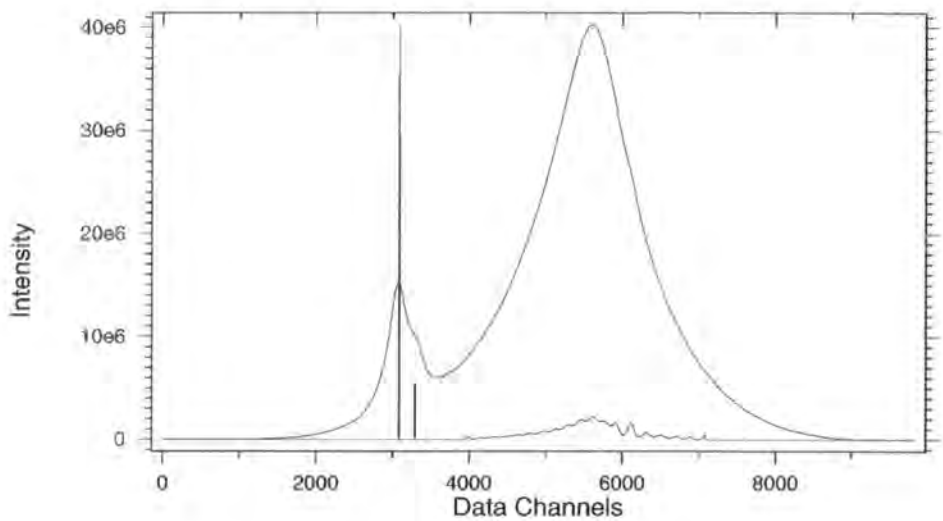


Figure 6.16. Spectrum of AACH overlaid with Maximum Entropy result

The Maximum Entropy results clearly show two tetrahedral peaks additional to the main tetrahedral resonance. Due to the rather narrow PSF the octahedral peak has been split into a large number of peaks which are not considered to be real. The areas of the tetrahedral peaks can be quantified and, given that baseline resolution has been achieved, the errors are likely to be much smaller than those associated with conventional integration methods.

Deconvolutions of this type have been performed for a number of different antiperspirant actives with known sweat reduction values. These measurements are taken from volunteers who undergo a hot-room test under controlled conditions. The tests are known to carry significant errors.

If the normalised tetrahedral aluminium intensity is plotted against mean sweat reduction a crude correlation can be observed in which the middle tetrahedral peak increases in intensity with sweat reduction as the other two peaks decrease in intensity (Figure 6.17).

This may suggest that the middle peak is from the tetrahedral aluminium in that species which is giving rise to the antiperspirancy. The difficulty in interpreting these data lies in the errors associated with the measurements of sweat reduction and also the errors associated with the Maximum Entropy deconvolution (the deconvolutions have not been repeated to establish the size of the errors).

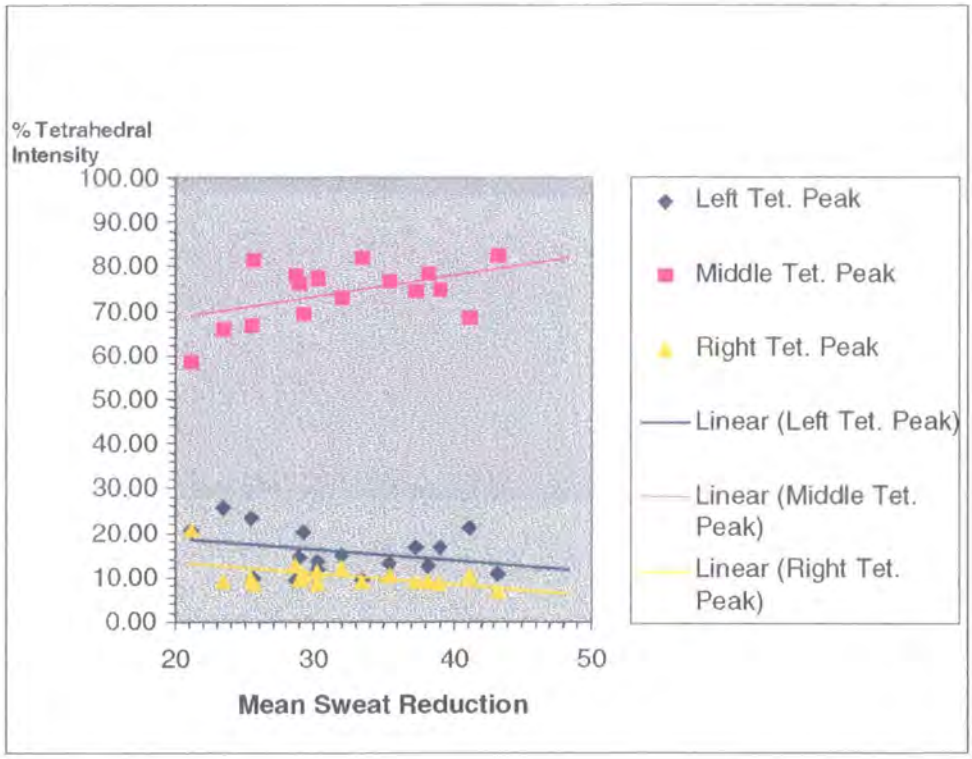


Fig. 6.17 Maximum Entropy intensities versus mean sweat reduction

6.5. Conclusions

Maximum Entropy data processing is limited to spectra that contain similar peak widths. The ^{27}Al spectra of aluminium chlorohydrate solutions contain a wide range of peak widths. This problem can be overcome with linear prediction techniques to produce sub-spectra that contain either only broad peaks or only sharp peaks. The sub-spectrum containing the broad peaks can be processed with a Maximum Entropy algorithm. The results indicate the presence of a least three tetrahedral peaks. This is consistent with some provisional work carried out by solid-state NMR spectroscopy and a Principal

Components Analysis of the spectra. A crude correlation has been found between the tetrahedral peak intensities and mean sweat reduction measurements from a number of antiperspirant actives.

References

1. *A User's manual for running 1D MaxEnt*, MaxEnt Solutions Ltd. 1994.

2. T. Ebbels, J. Lindon, J. Nicholson, *Quantitative Comparison Of Memsys5, Massive Inference and GIFA Spectral Processing using Simulated NMR Spectra.* Poster Presented at European NMR Conference, Edinburgh, 1999.

3. Delsuc M.A., in *Maximum Entropy and Bayesian Methods*, Cambridge, ed. J. Skilling, 285, 1989. GIFA is available at www.cbs.univmontpl.fr/GIFA.

4. J.C. Hoch, A.S. Stern, *NMR Data Processing*, Wiley-Liss, 1996.

5. K.S.Lee, *Measurement Sciences Standard Operating Procedure; Determination of Aluminium Speciation in Aluminium Chlorohydrate and Related Species by Solution State NMR.*, 1998.

6. G. Fu, L.F. Nazar, and A.D. Bain, *Chem. Mater.*, 1991, **3**, 602.

7. J. Rowsell, N. Taylor, C. Campana, and L.F. Nazar, submitted to *J. Am. Chem. Soc.I*

8. XWINNMR, version 2.1, Bruker Spectrospin, Karlshruhe, Germany.

9. W.M. Wester, F. Agilgaard,. *DMX Digital Filters and non-Bruker offline Processing III*, 1996. Web Page: http://garbanzo.scripps.edu/nmrgrp/wisdom/dig.nmrform.t

10. L. Allouche, C. Gerardin, T. Loiseau, G. Ferey, and F. Taulelle, *Angew. Chem. Int. Ed.*, 39, 3, 2000.

11. R.K. Harris, D. Apperley, Structure of Keggin Ion, Private communications.

CHAPTER 7:

Quantitative Analysis of Electrospray Mass Spectra

Electrospray ionisation is a widely used technique for introducing non-volatile compounds into a mass spectrometer. If these compounds contain more than one site capable of becoming charged multiply charged ions may be observed on a mass/charge axis.

Dyes are an example of analytes which have many charge sites available. This can lead to complex mass-spectra, which become increasingly difficult to interpret without the aid of reconstruction techniques. In this chapter, the Maximum Entropy approach to reconstruction is described qualitatively and then applied to two commercially available dye systems. Whilst the algorithm is widely available, it is most commonly used in positive-ion mode for reconstruction of data from bio-molecules, e.g. proteins. This work extends the current applications to data acquired in negative-ion mode and to systems containing more than one type of counter-ion.

It is shown that, for the complex dye systems, the zero-charged spectra can be easily reconstructed. These reconstructions are based on either the mass/charge ratio of sodium or that of the proton. This depends on whatever counter-ion is lost to produce the charged molecule. Minor spectral correlations due to sodium adducts are resolved. The techniques are extended to dye mixtures with successful separation of individual dye species. The form of the electrospray response as a function of analyte concentration is described in terms of the current models for electrospray ionisation.

7.1 Introduction

Electrospray ionisation is a technique suited to the introduction of polar, thermally labile compounds into the mass spectrometer. In this technique the solvent, containing the sample to be

analysed, is pumped through a fine needle which is maintained at a high voltage. This has the effect of forming a mist of highly charged droplets in the atmosphere of the spectrometer. As the ionic species in the sample solution emerge from the needle they move into the mass-analyser (in work presented here a quadrupole) in response to the imposed electric field.[1]

These ions become desolvated by some very low energy process that does not induce fragmentation. Two alternative mechanisms have been proposed for this process and are described by Constantopoulos[2]. The first is coulomb fission. In coulomb fission, the charge density increases as the solvent evaporates. The surface tension forces, which keep the droplet intact, are eventually overcome by the increasing charge density on the droplet causing it to divide. Coulomb fission causes droplets of various sizes. The second mechanism for desolvating the ions is ion-evaporation. Like coulomb fission, in the ion-evaporation model the charge density increases while the solvent evaporates. Instead of forming smaller and smaller droplets, coulombic repulsion overcomes the attraction of the charged ions to the droplet surface and ions are expelled directly from the droplet surface.

Constantopoulos continues, "although these proposed mechanisms do predict how ions in the electrospray droplet are transferred into the gas phase, neither mechanism predicts the preferential expression of particular species in the mass spectrum." Compounds, containing more than one charge site, are observed as singly or multiply charged ions. For many singly and doubly charged ions, the relative intensities of the ions in the spectrum do not reflect the relative concentration of the ions in solution, sometimes by several orders of magnitude.[3] Constantopoulos has described the effect of salt concentration on analyte response using electrospray mass spectrometry. [2]

In this work, maximum entropy methods are used to disentangle the mass-spectra of a mixture of analytes and the derived maximum entropy intensities are used to calibrate the electrospray response as a function of analyte concentration.

Despite the problems associated with the electrospray response and the complexity of the multiply charged spectra, the electrospray ionisation technique allows rapid, accurate and sensitive analysis of a wide range of analytes from low molecular weight (less than 200 Da) polar compounds to bio-polymers larger than 100 kDa. Generally, compounds of less than 1000 Da produce singly charged protonated molecular ions $(M+H)^+$ in positive ion mode. Likewise, these low molecular weight analytes yield $(M-H)^-$ ions in negative mode, although this is dependent upon compound structure.

For ions from a molecule of larger mass a series of peaks is observed in the spectrum of the multiply-charged ions. In positive-ion electrospray, each peak represents a given ion, with adjacent ions in the same series differing by, generally, plus or minus one proton. Generally, ions occur with mass-to-charge ratio:

$$\mathbf{(M + zH) / z} \qquad \text{.....}\{7.1\}$$

where:

M is the molecular mass of the analyte

H is the mass of the species responsible for the charge, normally the proton

z is the number of charges on a particular ion (assuming that **H** is singly charged).

In measuring the molecular weight for simple systems, the charge on any one of the ions is first calculated by solving a pair of simultaneous equations for any two consecutive ions in a series. Hence, the charge for all the ions in the series is calculated step-wise and the molecular weight deduced. For more complex systems, each component can give rise to a series of multiply charged ions and subsequent peak overlap can prohibit the calculation of accurate mass/charge ratios. Reconstruction is desirable.

Early methods for generating a zero-charged mass-spectrum produced a baseline that increased with mass and tended to introduce artifacts. [4] More recent methods require prior identification of the charge states in the multiply-charged ion series.[5] Widely used software, whilst allowing automatic assignment of charge states, is limited by directly transforming the multiply-charged spectrum onto a true molecular mass scale, i.e. components which are unresolved in the original data remain unresolved on the molecular mass axis[6]. A probabilistic approach to disentangling electrospray data involves the use of the Memsys5 algorithm and will form the basis of this work.[7] The Memsys5 electrospray interface has been developed for the analysis of biological samples in positive-ion mode and is beginning to be used routinely.[8,9]

It is worth noting that the true underlying spectrum of masses is sharper than the peaks in the original data. Signals in the multiply charged spectrum are inherently broadened by:

- the isotopic distribution of the elements in the analyte (if not resolved)
- the instrument will have a finite resolution

By incorporating this broadening into the program it should be possible to use Memsys5 to deconvolute it from the spectrum, thus enhancing the resolution achievable from electrospray data. The current implementation of Memsys5 is limited to a single estimate of the peak-shape present in the multiply charged spectrum. This is a Gaussian curve of constant width and is a poor model of both the above line-broadening effects. An estimate of the peak-width-at half-height is input to the algorithm.

In the work presented here, the negative-ion electrospray mass spectra of two dyes, Direct Red 80[10] and Direct Yellow 50[11], will be considered. An attempt is made to use the Memsys5 algorithm to reconstruct the mass spectrum of the zero-charged molecules for these systems. This feasibility study will then be used as the basis for the reconstruction of spectra from more complex dye mixtures. It is worth noting that the spectrum of the multiply charged ions from a mixture of two dyes is extremely complex with many overlapping bands. Manual estimation of molecular-masses

from data of this quality is very difficult even for the most experienced mass spectrometrist. The aim of this work is to identify the individual dye components in complex mixtures with as little user intervention as possible and attempt to quantify the concentration of each dye present.

7.2 Memsys5 Technique

A full mathematical treatment of how Memsys5 reconstructs a spectrum of the zero-charged molecule is considered to be beyond the scope of this work but is based on the principles explained in Chapter 2. However, a qualitative description is now presented.

The Memsys5 algorithm does not process the experimental spectrum directly but rather uses it for comparison. Initially, the algorithm generates a randomly selected spectrum of the zero-charged molecule, trial 1, and then applies the principles of electrospray ionisation, described in equation {1}, to estimate the corresponding spectrum of the multiply-charged ions. The difference between this trial spectrum of the multiply-charged ions and the experimental spectrum gives rise to residuals which are used to assess the probability of the initial estimate being a good match for the raw data. A number of trial spectra will fit the experimental spectrum within the noise. The Maximum Entropy approach is to select that spectrum with the minimum structure, i.e. the maximum entropy.

A few tens of trials will normally suffice for the algorithm to converge, and those trial spectra which agree well with the data, and which are also intrinsically plausible through having large entropy, make up the probability distribution of plausible results, i.e. the Memsys5 result. This probability distribution is then mathematically sampled to determine the optimum result; the width of the uncertainty distribution specifies the uncertainty associated with the optimum result and allows error bars to be calculated for any of its features.

Apart from the residuals, the calculation of appropriate trial spectra is also assisted by equation {1}. If a particular compound has a peak present at charge z, then it should also be present to some extent at charges z+1, and also at z+2, and so on, up to a maximum depending on how many

charges the molecule can accept. Depending on the particular analyte this may not always be true for all possible charge states; defect sites may appear depending on the relative reactivity of a particular site. This 'failure' can add to the difficulty in interpreting spectra of this type.

This expectation of peaks is incorporated into the Memsys5 analysis. However, a peak at a particular mass can only appear in the final Memsys5 spectrum if it appears, to some extent, in the spectrum of the multiply charged ions; this again holds with the principle of only selecting those spectra for which there is maximum evidence. This expectation effectively ties each series of spectral peaks together. As the connections are intrinsic, there is no need in a Memsys5 analysis for the user to identify charge states prior to processing.

Figure 7.1 is an example of the multiply-charged mass spectrum of an aqueous solution of Direct Red 80, with Figures 7.2 and 7.3 showing the Memsys5 result as the algorithm iterates towards convergence; the algorithm converged after 30 iterations, shown in Figure 7.4. Note the way in which the algorithm is able to 'lock onto' the main spectral features after only a few iterations; the final iterations are associated with adding fine structural details to the result.

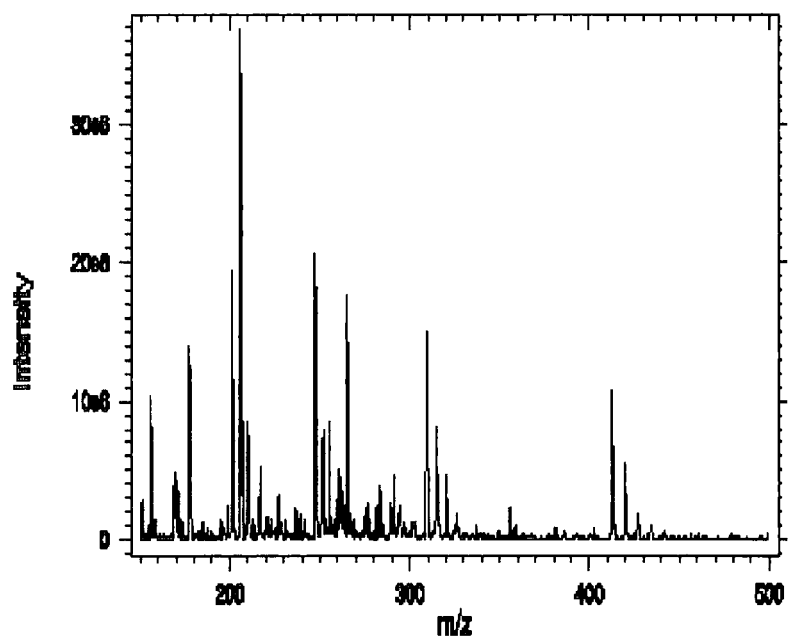


Fig. 7.1 Multiply charged spectrum of Direct Red 80

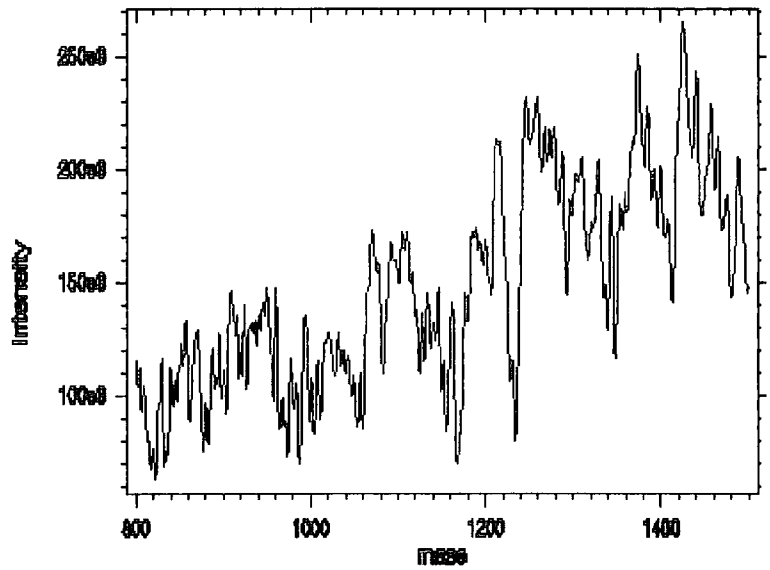


Fig. 7.2 Memsy5 result: iteration 1

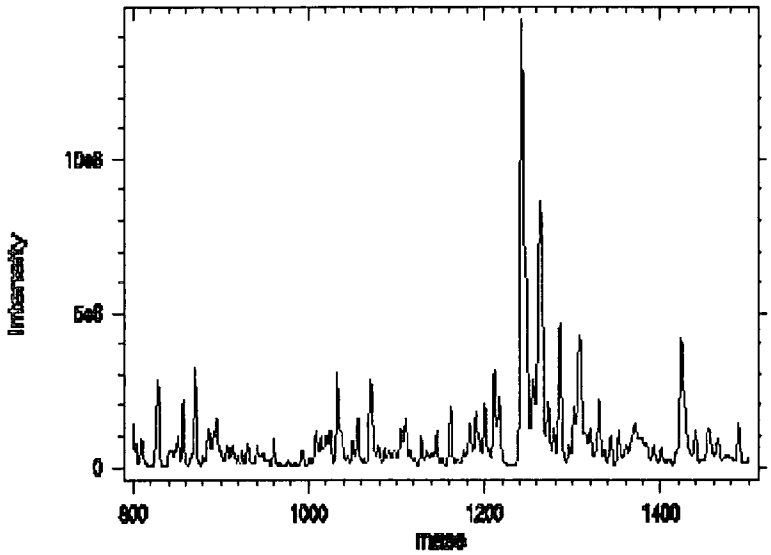


Fig. 7.3 Memsy5 result: iteration 5

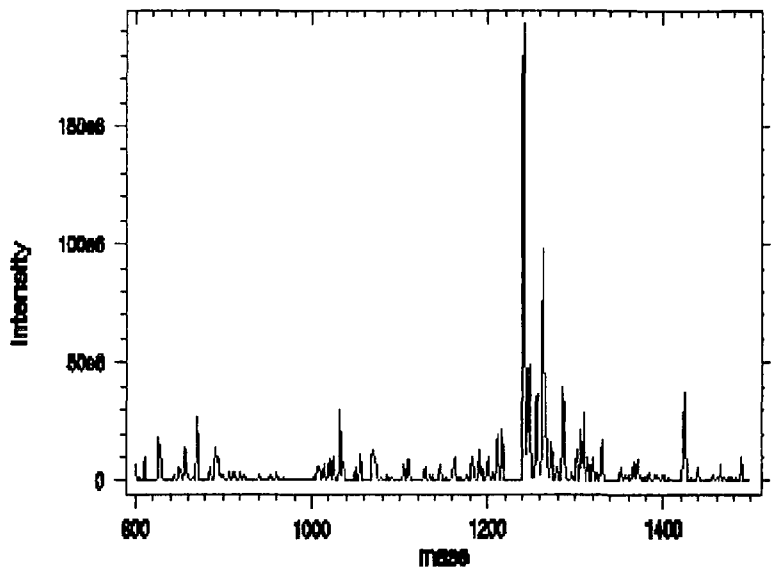


Fig. 7.4 Memsys5 result: iteration 30

One benefit of the multiple charging is that it effectively increases the mass range of the quadrupole spectrometer. The observed mass range in the multiply charged spectrum (Fig. 7.1) is 500 Daltons compared with a mass range up to 1500 m/z in Figures 7.2 - 7.4. The greater the degree of multiple charging the greater the apparent mass-range that can be achieved.

As described earlier, the probability cloud of plausible results is then sampled to determine the optimum result and the confidence limits. Fifteen samples are normally taken, five are displayed in Figure 7.5.

It is clear from Figure 7.5 that the main spectral features are present in each of the plausible spectra; the optimum result for this system has been derived with very little iteration. Minor fluctuations in the baselines give rise to the calculated estimates of uncertainty, which manifest themselves as standard error bars on peak intensity and position. For systems in which the spectra of plausible results are not consistent, the associated probability cloud would be wider and the error bars on any one spectral feature much larger. The determination of this probability cloud relies on the

signal:noise ratio in the raw data. High noise levels provide the algorithm with extra degrees of freedom, with many more plausible spectra, and much larger error bars.

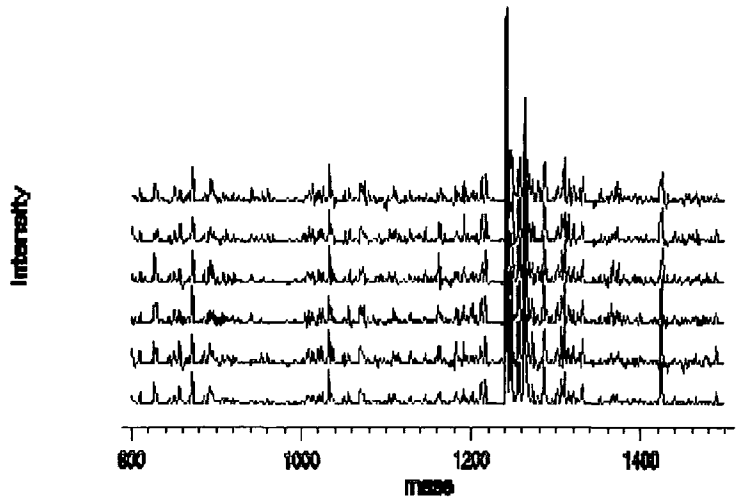


Fig. 7.5 Memsys5: plausible results

7.2.1 Input To Memsys5

Memsys5 requires a number of inputs in order to reconstruct an accurate estimate of the zero-charged spectrum, i.e.

- A measure of the peak-width-at-half-height (HWHM) is required. Given that, for the reasons described earlier, there is likely to be a range of peak widths present in the data the value of HWHM used in this work is an average width for all the peaks in the data. This can be determined by monitoring the algorithm's diagnostics following a series of trials (see section 7.3.1.1 for a description of how the HWHM is optimised). For this application, the peak shape is always assumed to be gaussian.

- The mass of the species responsible for the charge is required. For example, in equation {7.1}, this would be the mass of the proton. This must be a negative value if the data has been acquired in negative ion mode.
- A less critical input is an estimate of the fractional reduction in peak intensity towards higher and lower mass with respect to the envelope maximum for any particular charge state. This is assumed to be constant for all charge states and informs the program of the expected minimum intensity for adjacent charge states and is used to reduce any low probability correlation, primarily due to noise in the data. An accurate estimate of this input is only required for the best possible reconstruction, for most applications the default parameters are used.
- An estimate of Sigma is required. A description of Sigma can be found Chapter 3. In brief, it is a calculated value which estimates the noise in the raw data and any mismatch between the actual data and the applied gaussian bandshape. Apart from trials for determining HWHM, the default value is used and sigma is calculated internally by the algorithm. For this application, the algorithm assumes Poisson noise characteristics.
- An estimate of the expected output mass range and the number of Daltons per output point is also required, the former can be determined from the algorithm's diagnostics following an initial reconstruction over a large output mass range.

7.3 Experimental

Before spectra from dye mixtures are considered it is necessary to reconstruct the zero-charged spectra for the individual dyes to establish peak positions on the mass scale. These are then be used as the basis for assessing the quality of the reconstruction for the more complex dye mixtures.

All the electrospray spectra presented in this thesis were acquired on a VG Platform II mass spectrometer.

7.3.1 Direct Red 80

An electrospray mass-spectrum of Direct Red 80 was recorded in negative-ion mode by Dr. Mike Dale, Unilever Research, in 50/50 $MeOH/H_2O$ following reaction with triethylamine. The structure of Direct Red 80 is shown in Figure 7.6 and the multiply charged mass-spectrum is shown in Figure 7.7. (Only four charge states have been recorded).

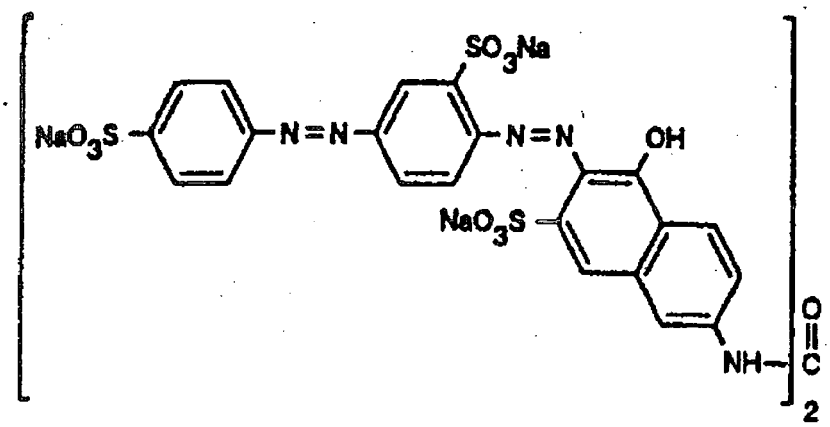


Fig.7.6 Direct Red 80. Relative Molecular Mass = 1373

The effect of the triethylamine is to produce the free-acid having a formula weight of 1241. The major peaks in the above spectrum can be easily rationalised without any Memsys5 processing. However, this system is a good starting point to establish the limitations of the Memsys5 technique. Direct Red 80 has six sulphonate groups which can each give rise to a different charge state, i.e. six multiply charged states may be observed in the mass spectrum. Of course, if a defect site is present not all of these will be observed.

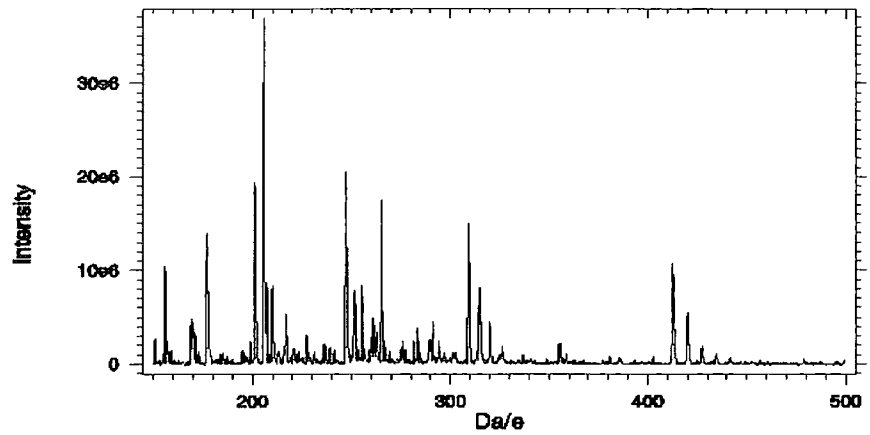


Fig.7.7 Direct Red 80: multiply charged data. Only four charge states acquired.

Is it necessary to use all six available charge states to achieve an acceptable reconstruction, or can satisfactory results be obtained by just processing the four more highly charged clusters of peaks, i.e. the peaks with the highest signal:noise ratio ?

Using the Memsy5 input parameter described in section 7.2.1, it is possible to use Memsys5 to reconstruct the zero-charged spectrum using either the mass of the proton or the mass of sodium as the basis. This will be demonstrated.

7.3.1.1 Four Charge States

Manually assigning the spectrum shown in Figure 7.7, i.e. with four charge states acquired, gives calculated masses given in Table 7.8.

Table 7.8: Manual assignment of spectrum of Direct Red 80

m/z	Assigned Ion	Calculated Mass
205.9	$[M - 6H]^{6-}$	1241.4
247.2	$[M - 5H]^{5-}$	1241.0
309.3	$[M - 4H]^{4-}$	1241.2
314.7	$[M - 5H + Na]^{4-}$	1240.8
320.3	$[M - 6H + 2Na]^{4-}$	1241.2#
326.4	$[M - 7H + 3Na]^{4-}$	1243.6
412.5	$[M - 3H]^{3-}$	1239.9
419.9	$[M-4H + Na]^{3-}$	1240.7
where M = free acid # (320.3 x 4) - (2 x 23) + 6 = 1241.2		

This is fine in as far as it goes, but there is some variability in the value of the calculated parent mass and it is not possible to identify any minor correlations which may be present.

In order to perform a Memsys5 reconstruction it was necessary to establish the peak HWHM from a series of trial reconstructions in which the input gaussian peak width was varied and the algorithm's output diagnostics, i.e. the evidence, plotted as a function of peak width. As described in Chapter 2, the evidence is the logarithm of the probability of finding that input peakwidth in the data; the more positive the value the better the estimate of the input HWHM.

The optimum HWHM was estimated at 4.7 data channels for a symmetrical gaussian peak (Figure 7.9). Given that the original data are likely to have a range of peak widths, due to isotope effects and increasing line-width with m/z, this value is only a compromise and is likely to resolve only bulk correlations.

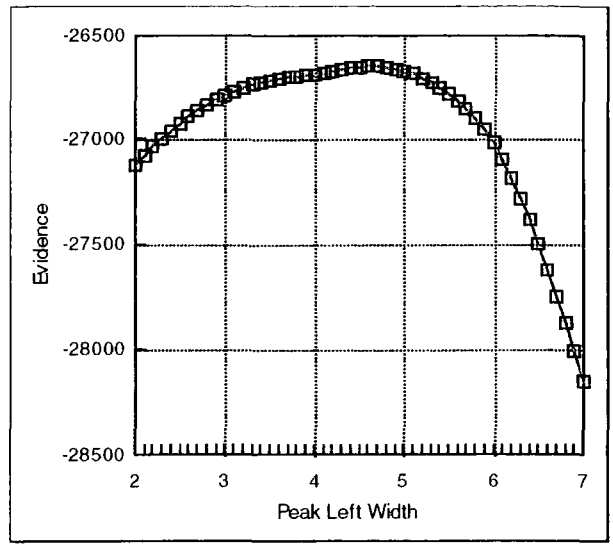


Fig.7.9 Evidence profile for a symmetrical gaussian peak

Whilst the maximum value in Figure 4 represents a compromise value for all the spectral peaks, the width of the evidence profile is indicative of a range of band widths being present in the raw data. The Memsys5 results based on this optimised HWHM value are shown in Figures 7.10 and 7.11.

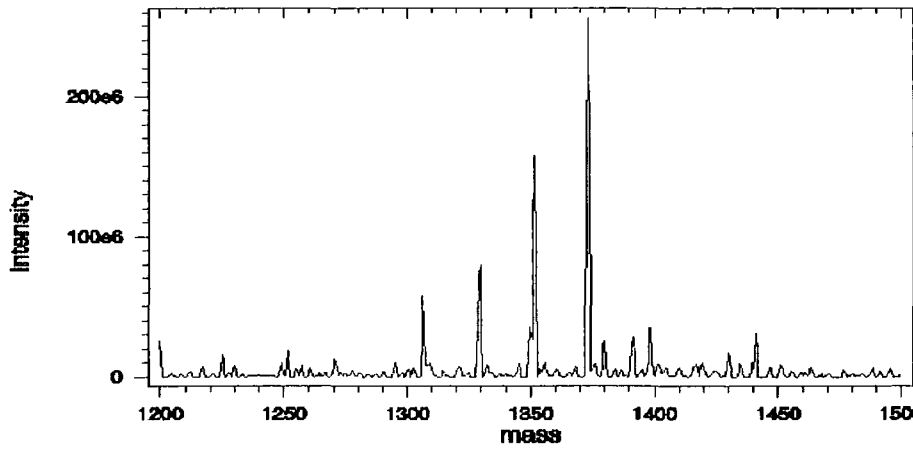


Fig. 7.10 Reconstruction based on mass of sodium

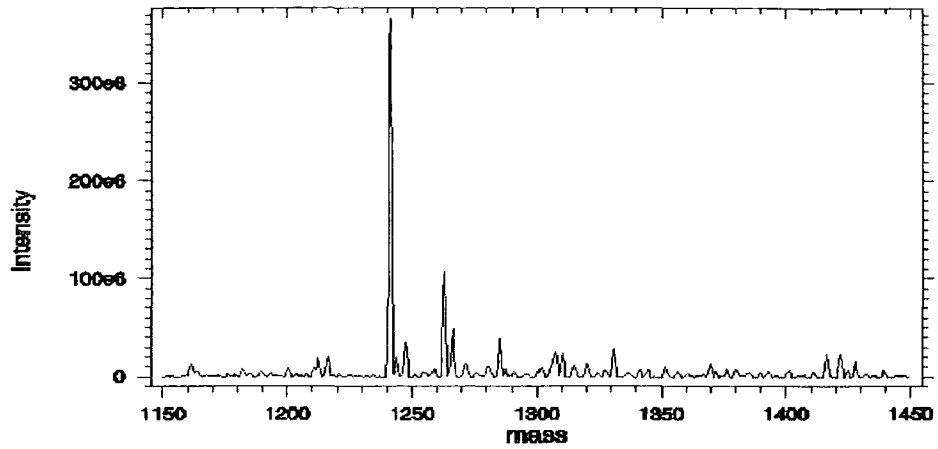


Fig. 7.11 Reconstruction based on mass of proton

Clearly, both reconstructions have identified a number of peaks. Their assignments are summarised in Table 7.12.

Table 7.12: Memsys5 Results

Species Responsible For Charge		Assignment
^{1}H	^{23}Na	
Mass	Mass	
1241.3		Free acid = [-6Na + 6H]
1262.9		[-5Na + 5H]
1285.5		[-4Na + 4H]
1307.3		[-3Na +3H]
1331.0		*[-2Na + 2H]*
	1306.6	[+3Na - 3H]
	1329.2	[+4Na - 4H]
	1351.4	[+5Na - 5H]
	1373.0	Salt = [+6Na]

Both reconstructions have identified correlations due to the sodium adducts present in this system. Additional peaks are probably due to minor correlations within the noise or the possibility of isotope peaks having been resolved if the applied PSF was in fact too narrow for some of the peaks in

the original data. Both reconstructions have identified three sodium adducts with some evidence for a fourth, although the intensity of this is such that it is very difficult to distinguish from the noise. Evidence for more adducts may be available if the raw data with six charge states is considered or if the true isotopic band shape is used as the PSF.

7.3.1.2 Six Charge States

The average HWHM value was again determined through an evidence profile. The Memsys5 result for the reconstruction based on the mass of the proton is shown in Figure 7.13.

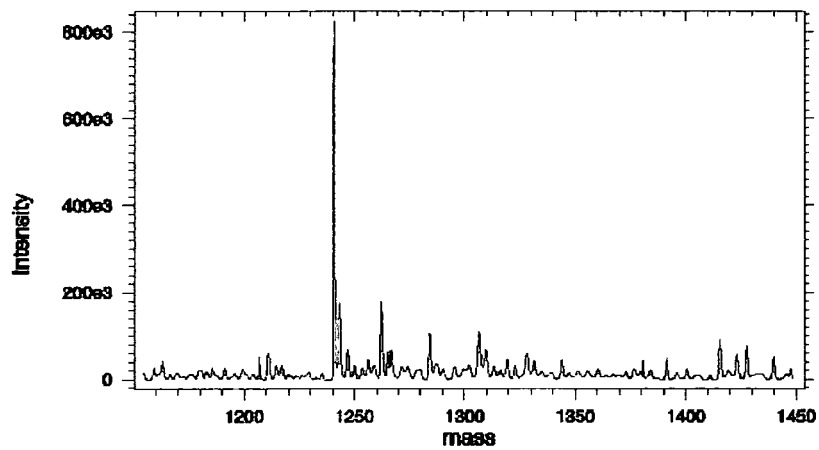


Fig.7.13 Reconstruction based on mass of proton: 6 charge states

The main features of the above spectrum are consistent with those obtained for the system containing only four charge states, i.e. parent mass at 1241 and at least three sodium adducts. However, the spectrum is complicated by the relatively poor signal:noise which is due to minor correlations in the noise of the multiply charged spectrum. In the multiply-charged spectrum the signal:noise ratio of the less highly charged clusters is low, and this lack of signal intensity is directly

responsible for the poor appearance, and relatively large errors, of the Memsys5 result. The same is true for the reconstruction based on the mass of sodium.

It would appear that recording all possible charge states is likely to deteriorate the Memsys5 result by introducing correlations due to a low signal:noise ratio. A further advantage of acquiring just four charge states is that the data will be recorded with better digital resolution and extra structure may be resolved. However, the reconstructed data cannot be used in any quantitative way because intensity will have been discarded by not recording all possible charge states. In conclusion, there is a trade-off between identifying those species present, where four charge states suffice, and determining the optimum result with full quantification, where all charge states must be processed.

7.3.2 Direct Yellow 50

An electrospray spectrum of Direct Yellow, was recorded by Michael Dale, Unilever Research, in negative-ion mode in 50:50 CH_3OH : H_2O. The structure of Direct Yellow 50 is shown in Figure 7.14 and the mass spectrum of the multiply charged ions is shown in Figure 7.15.

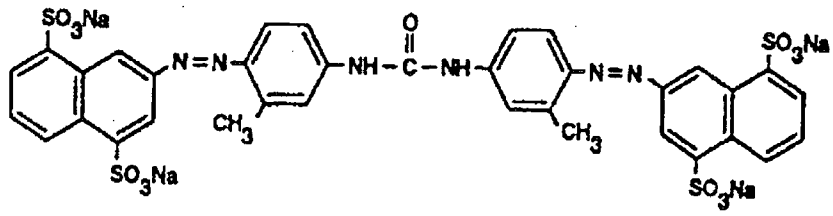


Fig. 7.14 Direct Yellow 50. Relative Molecular Mass = 956.

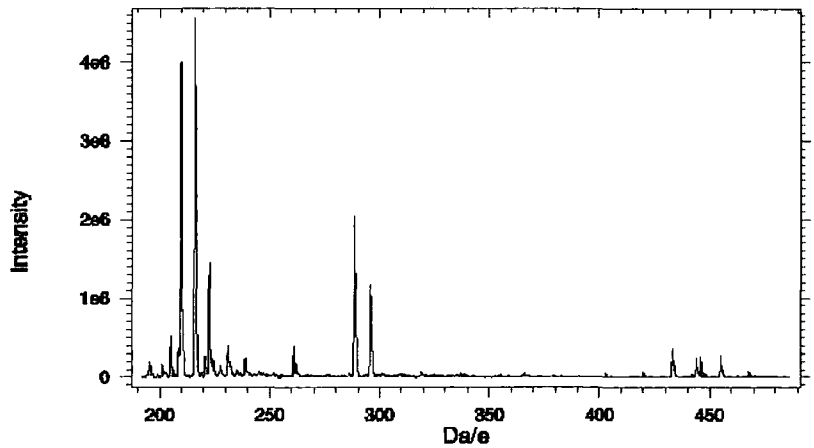


Fig.7.15 Multiply charged spectrum of Direct Yellow 50

The Memsys5 reconstructed masses are presented in Table 7.16 and Table 7.17 for the system with and without dialysis. As for the Direct Red 80 system, the non-dialysed system was studied as a function of the number of charge states.

Table 7.16. Direct Yellow 50 following dialysis.

3 charge states, ^{1}H reconstruction

mass	error	Assignment
868.563	0.079	[S - 4Na + 4H]

Table 7.17. Direct Yellow 50 without dialysis

Na Reconstruction				^{1}H Reconstruction				Assignment
3 Charge States		2 Charge States		3 Charge States		2 Charge States		
mass	error	mass	error	mass	error	mass	error	
				868.982	0.043	869.019	0.075	[S - 4Na + 4H]
				890.763	0.103	890.848	0.242	[S - 3Na + 3H]
913.287	0.317			912.890	0.272			[S - 2Na + 2H]
934.964	0.083	934.931	0.083					[S - Na + H]
956.901	0.045	956.977	0.095					[S]

where [S] = salt, i.e. free acid + 4Na

As expected, following dialysis the zero-charge spectrum can only be reconstructed on the basis of the ^{1}H. The algorithm correctly identified the lack of sodium in the system. The algorithm has found only one major correlation corresponding to a mass of 868.6 Da, i.e. loss of four sodium atoms followed by addition of four hydrogen atoms. For the system without dialysis, all four sodium adducts are observed if both the hydrogen and sodium reconstructed data are considered together. This is only true if three charge states are considered.

If the intensities of these sodium adducts are considered it may give some insight into the mechanism of protonation of these charge sites. From Figure 7.18, the relative intensity of the [S - 2Na + 2H] adduct is small, indicating that some sites may protonate preferentially.

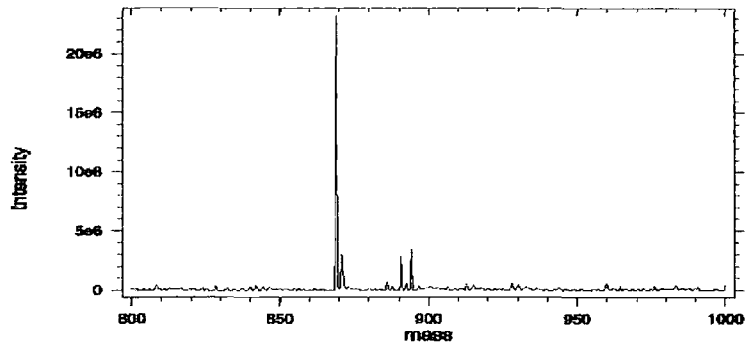


Fig.7.18 Maximum Entropy reconstruction showing sodium adducts

In conclusion, it would appear to be generally unnecessary to try and chemically simplify the multiply charged spectra of these systems; if a peak correlation is present above the noise level the algorithm should find it. Moreover, simplification of these spectra leads to useful information regarding sodium adducts being discarded.

7.3.3 Dye Mixtures: Electrospray Response Curves

The following dye mixtures were studied in 50:50 CH_3OH : H_2O

- 10:1, 5:1, 1:1, 1:5, 1:10 Molar ratio Direct Red 80 : Direct Yellow 50.

It was not possible to determine the absolute concentrations of the dyes present in solution because the dyes as received were known to be very impure. Aldrich quoted values of:

Direct Red 80 ~25% Pure

Direct Yellow 50 ~40% Pure.

The systems were not chemically treated with triethylamine or dialysed as in the previous examples. The aim was to assess the relative amounts of each dye present in solutiuon. Wherever possible a mass range was recorded to encompass all charge states. The HWHM of the Gaussian

PSF was estimated directly from the data by overlaying with a gaussian curve and adjusting its width until a good match was obtained by eye, i.e. no evidence profiles were generated. The most intense peaks in the spectra were used to determine the HWHM. The aim of this was to establish if a calibration of dye concentration could be obtained using the crudest optimisation available to the mass spectrometrist. Each spectrum was baseline corrected by fitting a polynomial curve through the noise present in the spectrum. No attempt was made to optimise this baseline correction.

Figure 7.19 shows the multiply charged spectrum for the 1:1 mixture. Figure 7.20 is the Memsys5 reconstructed zero-charged spectrum based on the mass of sodium.

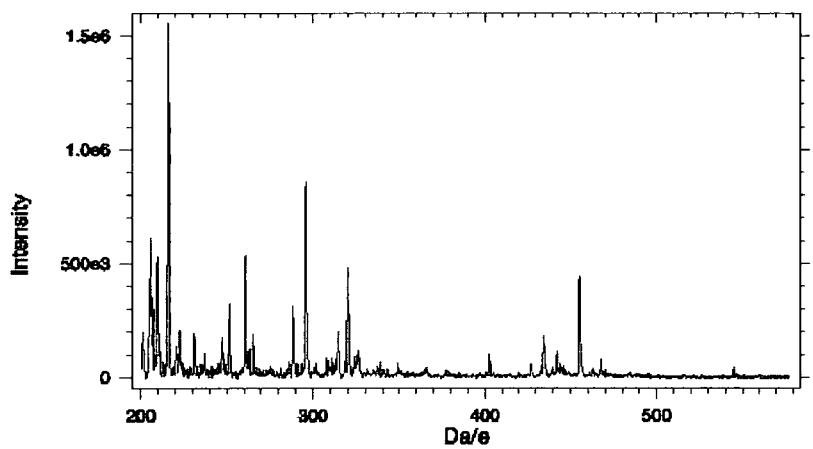


Fig. 7.19 Multiply charged spectrum: 1:1 dye mixture

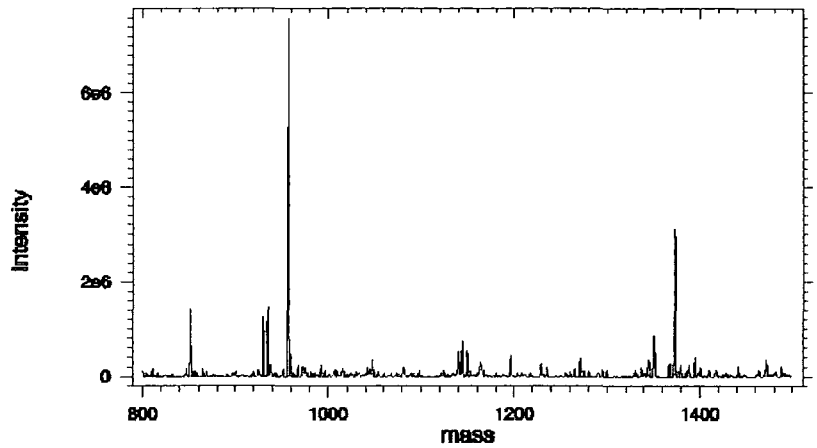


Fig. 7.20 Maximum Entropy reconstructed spectrum

Both dyes are easily recognisable in the reconstructed zero-charged spectrum with evidence for sodium adducts as before. Peaks are clearly evident at m/z = 956 and 1373, corresponding to the Direct Yellow 50 and Direct Red 80. Even at the molar ratio of 10:1 the less concentrated dye can easily be resolved. Appendix A7.1 is a summary of the reconstructed peak intensities and their associated errors for the reconstruction based on the mass of sodium. Only the most intense peaks are presented.

If the intensities from all these adducts is summed for each dye and a graph is plotted of electrospray response ratio (Direct Yellow 50 / Direct Red 80) against the ratio of the dye concentration present in solution, a response curve can be determined. For this system, an electrospray response curve is shown in Figure 7.21.

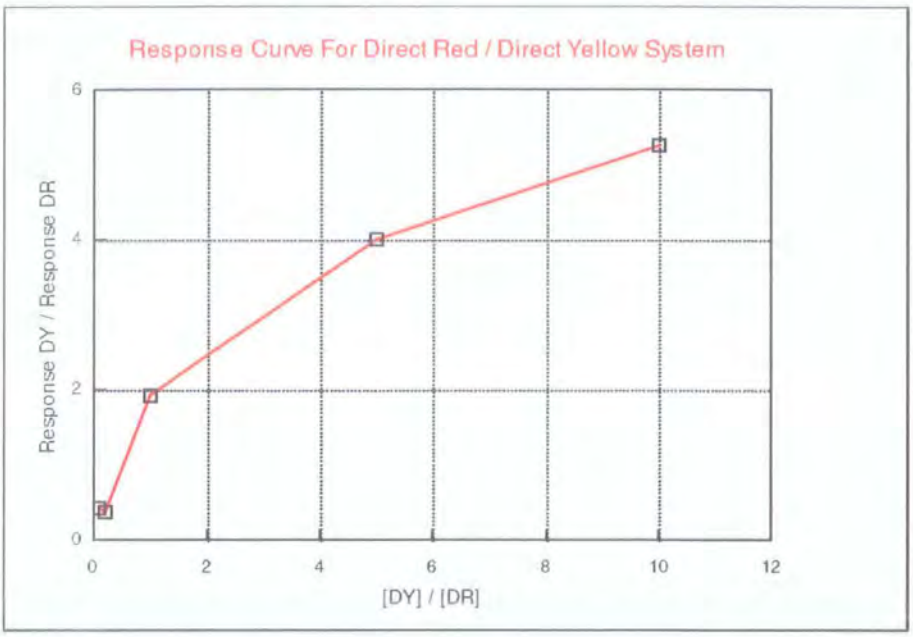


Fig. 7.21 Electrospray Response Curve

This response curve clearly deviates from linear behaviour. This is to be expected, and the literature proposes several equations to describe the form of such an electrospray response as a function of analyte-concentration[12]. These equations try to account for the fact that the response curve can be initially approximated to a linear model, followed by a reduction in gradient until a plateau is reached.

The simplest from of these equations is the ion evaporation model and is described for an analyte in the presence of an electrolyte. Tang and Kebarle[13] postulate that the production of gas phase ions is kinetically controlled and is dependent on the first order rate equation for the transfer of analyte and electrolyte ions into the gas phase, i.e.

$$I_{(A)} = P \cdot \frac{K_A [A^+]}{K_A [A^+] + K_E [E^+]} \quad \ldots\ldots\ldots\ldots\ldots\ldots \{7.2\}$$

where

$I_{(A)}$ is the detected analyte ion intensity

$[A^+]$ and $[E^+]$ are analyte and electrolyte concentrations. E will also include intensity from any background ions due to impurities.

K_A and K_E are the first order rate constants expressing the rate of transfer of analyte and electrolyte ions into the gas phase.

P is the proportionality constant expressing the "sampling efficiency" of the system.

Constantopoulos[2] notes that, like the Memsys5 derived reponse curve shown in Fig. 7.21, experimentally measured response curves reveal two distinct regions. In the first region, at low analyte concentrations, the response is linear. In the second region, at high analyte concentrations, the curve levels off and even decreases in intensity. Tang and Kebarle[13] were unable to fit the form of these response curves to the ion evaporation model over a wide range of concentrations. A good fit was obtained in the low concentration range of A^+ with the assumption that $K_A/K_E = 1$. However, the deviation of the predicted response curve from the experimental data at high concentrations of A^+ indicated that the actual value of K_A/K_E was greater than unity. Therefore, Tang and Kebarle changed their model to include terms due to surface activity and ion solvation effects in the hope of explaining the decrease in K_A/K_E at low concentrations. The extension of this model is summarised by Constantopoulos[2]. Only the main arguments are presented here.

The revised model is based on an ion depletion phenomenon. Its main points are:

1. Ion evaporation is assumed to occur at the droplet surface.
2. Those analyte ions with the higher surface activity have a higher concentration at the surface than in the bulk liquid.
3. A^+ ions evaporate at a higher rate than E^+ ions.

4. At high A^+ concentrations, A^+ ions are rapidly supplied from the bulk as A^+ ions evaporate from the surface. This maintains the constant ratio A^+ / E^+ ions at the droplet surface and results in a higher K_A/K_E ratio.
5. At low A^+ concentrations, all of the A^+ ions are at the droplet surface, where they evaporate at a high rate. There are no A^+ ions in the bulk to replenish the surface, so there is a depletion of A^+ ions at the surface relative to E^+ ions. This results in a decrease in the K_A/K_E ratio.

Whilst this revised model does explain the shape of the response curve observed for this two analyte system there is a second model proposed by Enke[14] which also explains much of what is observed with electrospray ionisation. Enke proposed an equilibrium partionioning model to explain the shape of these curves. The main features of this model are:

1. The excess charge on the droplet is fixed. This determines the total number of ions, which are un-neutralised by counter-ions, at the droplet surface.
2. All of this excess charge is at the droplet surface.
3. The droplet surface and the interior of the droplet are regarded as two independent phases, between which the ions can partition. This partitioning can be described by the equilibrium constants for each analyte.
4. The equilibrium constants account for effects due to: solvation energy, ion-paring energy, charge density, hydrophobicity, the nature of the counter-ion and the polarity of the solvent. Each of these can be found by experiment.
5. Ions evaporated from the surface are not replaced by other ions from the interior.
6. The model is independent of the mechanism of ion evaporation.

According to Constantopoulos[2], this model fits the experimental data over a wide range of concentrations with the same value of K_A / K_E , but fails to predict the electrospray bahaviour in terms of the suppression effects caused by the electrolyte. Constantopoulos[2] details the equilibrium partitioning model for the case of a single analyte and electrolyte. For completeness, the main stages

in the argument are presented here, with the extension to this system in which two analytes are present. The main stages in the partitioning model for a single analyte and electrolyte are:

1. Define partitioning constants, K_A and K_E in terms of the concentration of free ions, counter ions and ion pairs, i.e.

$$(A^+X^-)_i \rightleftharpoons (A^+)s + (X^-)_i;$$

$$(E^+X^-)_i \rightleftharpoons (E^+)s + (X^-)_i;$$

$$K_A = \frac{[A^+]_s [X^-]_i}{[A^+X^-]_i}$$

$$K_E = \frac{[E^+]_s [X^-]_i}{[E^+X^-]_i}$$

where X^- represents the counter-ions and s and i indicate surface and interior droplet phases.

2. Use K_A and K_E to establish a partition constant for the analyte and electrolyte competing for the counter-ion, this is shown in the diplacement reaction below. This gives the relationship between free analyte (A^+) and electrolyte (E^+), in terms of K_A, K_E, and the ion-pairs (A^+X^-) and (E^+X^-), i.e.

$$(A^+X^-)_i + (E^+)_s \rightleftharpoons (A^+)_s + (E^+X^-)_i$$

$$\frac{K_A}{K_E} = \frac{[E^+]_s [E^+X^-]_i}{[A^+X^-]_i [E^+]_s}$$

The terms $(A^+X^-)_i$, $(E^+X^-)_i$, $(E^+)_s$ are difficult to measure experimentally. Constantopoulos notes that the aim is to relate these terms to measurable quantities.

3. Eliminate $(A^+X^-)_i$, $(E^+X^-)_i$ by expressing these in terms of the total concentration of A, (C_A), the total concentration of E, (C_E) and the excess charge, (Q). Q can be determined from the spray current and the flow rate. (Q is also equal to the sum of the free ions at the droplet surface). There is an assumption made at this point. The model assumes that Q is independent of C_A and

C_E. This is unlikely to be a valid assumption; conductivity and hence Q will depend on on the concentration of A and E.

4. Constantopoulos continues with the derivation until he produces an equation which is quadratic in $[A^+]_S$, i.e. the concentration of the free analyte ions at the droplet surface.
5. Introducing the condition that $C_A << Q$, the quadratic reduces to a linear from, i.e.

$$[A^+]_S = C_A \cdot \frac{(K_A / K_E)}{(K_A / K_E - 1 + C_E/Q)}$$

6. Assuming that the response observed for analyte A, R_A, is proportional to $[A^+]_S$, then

$$R_A \propto C_A \left(\frac{1}{\text{Const} + C_E / Q}\right)$$

Constantopoulos[2] does not explicitly describe the case of more than one analyte but states that in such a case the values of C_A (for each analyte), C_E , K_A/K_E and Q "have a unique effect on the response of each analyte ion." It can also be argued that the same analyte ratio is unlikely to be observed in the gas phase as in the solution state. This suggests that a direct relationship can not be established for a solution containing more than one analyte; calibration experiments are needed for semi-quantitative analysis. Relative repsonse factors need to be determined over appropraite concentration ranges.

Constantopoulos[2] does not explicitly deal with the presence of multiply charged ions or adduct formation. He assumes that all the adducts will have the same contribution to the electrospray response. Multiple charging must have an effect on the magnitude of Q; it is unlikely that Q is independent of analyte concentration as described in note 3 above.

It is possible to extend the model presented above to more closely reflect the system studied here, i.e. two analytes. If the total ionic concentration is low then C_E could be replaced by the concentration one one of the dyes, say C_B. This gives:

$$\frac{R_A}{R_B} = \frac{C_A\ (\text{Const} + C_A/Q)}{C_B\ (\text{Const} + C_B/Q)}$$

If C_A and C_B are very much smaller than Q then the response is likely to be linear with concentration. If C_A and C_B are approximately equal to Q then the curve is quadratic. If CA is small and C_B large then the response will also be non-linear. This type of behaviour is consistent with the shape of the response curve derived from the Memsys5 intensities.

7.4 Conclusions

The Memsys5 electrospray interface has been successfully used to reconstruct the zero-charged mass-spectra of dyes acquired in negative-ion mode. The reconstructions provide evidence for the presence of sodium adducts; the reconstructions are made based on the mass of hydrogen and sodium.

The techniques have been extended to study a mixture of two dyes and a response curve has been determined based on the intensities of the peaks found in the Memsys5 reconstruction. The graph is initially linear before beginning to curve. The shape of this graph has been shown to be consistent with the current models of electrospray ionisation. It has been assumed that all the ions have the same response factor during the electrospray ionisation. This assumption appears to be valid. The Memsys5 results have demonstrated that if quantitative electrospray is needed, a bracketing experiment must be performed in which a control system is studied over a similar concentration range to the system of interest.

7.5 References

1. R. B. Cole, *J. Mass Spectrom.,* **35**, 763-772, (2000).
2. T. L. Constantopoulos, G. S. Jackson, C. G. Enke, *J. Am. Soc. Mass Spectrom.,* **10**, 625-634, (1999).
3. Z. L. Cheng, K. W. M. Siu, R. Guevrement, S. S. Berman, *J. Am. Mass Spectrom.,* **3**, 281, (1992).
4. M. Mann, C. K. Meng and J. B. Fenn, *Anal. Chem.*, **61**, 1702, (1989)
5. J. C. Cottrell, B. N. Green and S. A. Javis, *Proceedings of the 39th Conference on mass spectrometry and allied topics,* Nashville, TN, 234, (1991)
6. 'Transform', VG MassLynx User's Guide, Vol.1, Fisons Instruments, Ver. 2.00, 213, (1994)
7. MSL Solutions Ltd.
8. A. G. Ferrige, M. J. Seddon, S. Jarvis, *Rapid Commun. in Mass Spectrom.*, **5**, 374, (1991)
9. A. G. Ferrige, M. J. Seddon, B. N. Green, S. A. Jarvis, J. Skilling, *Rapid Commun. in Mass Spectrom.*, **6**, 707, (1992).
10. Direct Red 80, Aldrich Chemical Co. ref: 36,554-8
11. Direct Yellow 50, Aldrich Chemical Co. ref: 20,189-8
12. R. D. Smith, J. H. Wahl, D. R. Goodlett, S. A. Hofstadler, *Anal. Chem.*, **65**, 13, (1993).
13. L. Tang, P. Kebarle, *Anal. Chem.,* **63**, 2709-2715, (1993).
14. C. G. Enke, *Anal. Chem.,* **69**, 4885-4893, (1997).

CHAPTER 8: CONCLUDING REMARKS

The work presented in this thesis has attempted to demonstrate, that despite the scepticism that prevails within the NMR community, probabilistic data processing does have its place within the spectroscopist's armoury. The scepticism arises mainly from the apparent ability of the different algorithms to 'get something for nothing'. The approach taken in this research is to take a pragmatic approach to using the algorithms on a range of different real, i.e. not simulated, NMR and mass spectra. The systems were careful chosen such that any conclusions drawn from the algorithm's output could be verified by argument based on literature results and data supported from other techniques.

Following a brief description of the theory and practical operation of the software the first system chosen was the NMR spectra of sodium carboxymethyl cellulose. The NMR spectrum of SCMC is typical of many native polymers. The peaks are very broad and little information can be extracted without sample degradation to smaller oligomers or monomer. Maximum Entropy techniques were applied to the NMR spectrum of an intact SCMC sample. The Maxim Entropy derived peaks are shown to be consistent with those observed in the literature. The work was extended to demonstrate how these techniques could be used for spectral de-noising. The NMR spectrum of a typical fabric washing powder shows bands with very poor signal:noise ratio which are assigned to SCMC. The Memsys5 algorithm was used to de-noise the spectrum and show that the result is consistent with the main spectral features of SCMC. This type of application is of interest to the soaps industry for policing patent infringements by competitors. Very low levels of SCMC can have a significant effect on product performance.

The ability of the Memsys5 algorithm to achieve simultaneous resolution enhancement and signal:noise improvement was demonstrated by the processing of the NMR spectrum of a styrene / maleic anhydride polymer. The resultant Memsys5 intensities were used to calculate a polymer composition which was shown to be consistent with the known

monomer feed ratio and supporting data from mass spectrometry and elemental analysis. The consistency of this result meant that more confidence could be placed on the subsequent discussion of polymer microstructure. The polymerisation was shown to be a least a second order Markovian process and the conclusions were shown to be internally consistent by checking them against the results from a different region of the NMR spectrum. It is demonstrated that for this particular system there is no advantage to acquiring the NMR spectrum at a higher magnetic field strength. Probabilistic data processing is the only method for extracting the level of information needed to derive both polymer composition and microstructure. A literature chemical shift assignment is also reversed on the basis of a stereochemical argument. Unfortunately, the Memsys5 derived intensities failed to confirm this reassignment.

One of the major failings of Maximum Entropy based deconvolution techniques have been the inability to deal with spectra in which there is a wide range of peak widths. Only one point spread function could be used as the deconvolution model. Such a difficult system is the ^{27}Al NMR spectrum of aluminium chlorohydrate salts. These systems are used in anti-perspirant products and have been studied extensively for compositional information using a laborious manual method of peak integration. The problem of different peak widths was overcome by employing linear prediction techniques, prior to any Maximum Entropy processing, to reduce the NMR spectrum into a series of sub-spectra. Each sub-spectrum contained peaks of similar width and could be then be processed with the Memsys5 algorithm. This process has been automated and has now been implemented at Unilever Research. During 2000, the auto-program saved approximately one day effort per week.

Three peaks were identified in the tetrahedral aluminium region of the spectrum. An attempt was made to correlate this previously unreported structure with sweat reduction and hence antiperspirant efficacy.

Maximum Entropy techniques have been available for some time to disentangle mass spectra of multiply charged ions from an electrospray ionisation process. The previously reported application were all based in positive-ion mode. This work has been extended to

include the negative-ion mode mass spectra of dyes. The Memsys5 results are used to resolve the effects of sodium adducts on the mass spectrum. The technique has also been extended to include dye mixtures, with successful spectral separation of individual dye species. The form of the electrospray response as a function of analyte concentration has been described in terms of the current models for electrospray ionisation.

These results demonstrate that this type of data processing should be available to practicing spectroscopists. It will never replace simple fourier techniques but should be used as a last resort, when all other avenues have been exhausted. That said, the results can be remarkable and if a spectroscopist can afford the investment in time needed to successfully grasp the software the level of extra information that can be recovered from a spectrum will lead to further insights.

Appendix: Listing of Linear Prediction Autoprogram

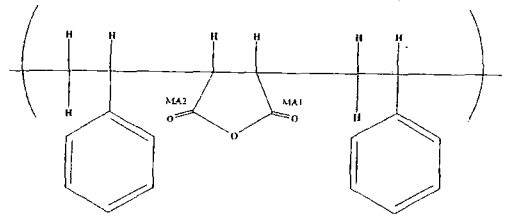
```
/*  Aluminium Data Processing    */
/*  L.P. Hughes; 1999      */
#include <lib/util.h>
/*Proc_err (ERROR_OPT, "Autoprogram for Linear Prediction and Plotting of
Aluminium Spectra,
L.P. Hughes, Dec 1998"); */
if (Proc_err(QUESTION_OPT, " Have you typed rpar les3 proc before running this
program ?" ) == ERR_CANCEL )
{STOP}
if (Proc_err(QUESTION_OPT, "Have you read in and correctly phased the first spectrum
?" ) == ERR_CANCEL )
{STOP}

/*  READ IN CURRENT DATASET AND COPY TO NEW FILE TEMP    */

GETCURDATA
WRPA("temp",1,1,"u","nightnmr")
WRPA("temp",2,1,"u","nightnmr")
WPAR("les3","proc")

/*  READ IN PHASE PARAMETERS FROM ORIGINAL DATASET  */

DATASET ("temp",2,1,"u","nightnmr")
GETCURDATA
VIEWDATA
RPAR("les1","proc")
em
ft
RPAR ("les3","all")
pk
abs
rmisc ("intrng","intrng.al3")
plot
DATASET ("temp",1,1,"u","nightnmr")
GETCURDATA
VIEWDATA
```

```
/* SET UP FILE TEMP READY FOR LINEAR PREDICTION        */

RPAR("les","proc")
ft
ift
RPAR("les1","proc")
em
ls

/* LINEAR PREDICT BACKWARDS, INTEGRATE + PLOT SHARP SPECTRUM   */

RPAR("les2","proc")
ft
RPAR("les3","all")
pk
GETCURDATA
VIEWDATA
WRPA("temp",3,1,"u","nightnmr")
apk
abs
rmisc ("intrng","intrng.al1")
plot

/* SUBTRACT SHARP SPECTRUM FROM BROAD SPECTRUM  */

DATASET2 ("temp",2,1,"u","nightnmr")
GETCURDATA2
DATASET3 ("temp",3,1,"u","nightnmr")
GETCURDATA
RPAR("les1","all")
ADD
DATASET ("temp",1,1,"u","nightnmr")
GETCURDATA
RPAR("les3","plot")
abs
rmisc ("intrng","intrng.al2")
plot
QUITMSG ("Program Finished")
```

Determination Of Polymer Microstructure By 13C Nuclear Magnetic Resonance Spectroscopy and Maximum Entropy Data Analysis

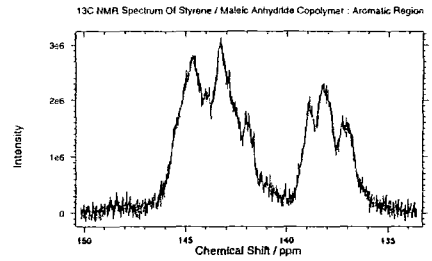
L.P. Hughes and K.S. Lee, Unilever Research, Bebington, Wirral, England.

The microstructure of a copolymer can have a profound effect on its physical and chemical properties and therefore determination of microstructure is an important analytical requirement. This poster describes the application of 13C Nuclear Magnetic Resonance Spectroscopy and Maximum Entropy (Memsys 5) data processing to a copolymer of styrene / maleic anhydride (75% styrene ex Aldrich Chemical Co.)

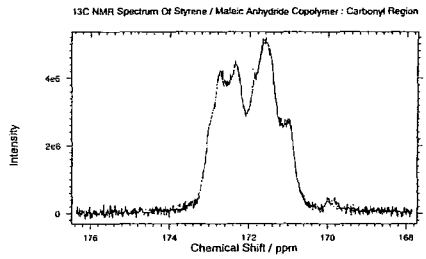


Both the aromatic and carbonyl regions clearly contain a number of unresolved peaks, the lack of any one discrete peak making optimisation of MaxEnt parameters difficult.

The 13C chemical shift is sensitive to both the presence of neighbouring groups and polymer tacticity resulting in severe overlap of spectral bands and subsequent uncertainty in assignment e.g. the quaternary aromatic region of the spectrum spans a range of ~ 10 ppm.



The difficulty in analysing the aromatic region of the spectrum has been previously described [1], however, spectral assignments are presented in the literature together with the comment that the monomer distribution calculated using these assignments could have up to a 20% error due to poor spectral resolution. [2,3]



Summary Of Findings

- The pragmatic use of Memsys 5 allows the determination of vinyl polymer microstructure and stereochemistry

1. Deconvolution

Memsys 5 gives a good deconvolution of this data with small error bars and residuals

2. Analytical Verification Of MaxEnt Result

The deconvolution is shown to be consistent with the monomer composition as determined by other analytical methods.

3. Exploration Of MaxEnt Result

The structure in the MaxEnt result suggests a 2nd order (or greater) Markovian model for polymerisation.
Literature assignments (based only on additivity of substituent effects) are discussed.

4. Rationalisation Of Model

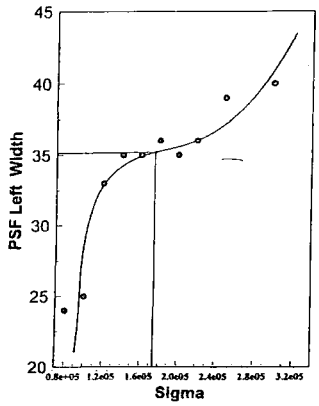
The reassignments are compatible with steric considerations i.e. the sensitivity of the C1 aromatic carbon of polystyrene to the through space effect of the γ substituent.

The interpretation of the C1 region is supported by the deconvolution of the carbonyl region of the spectrum.

[1] D J T Hill, J H O'Donnell, P W O'Sullivan, Macromolecules, 1985,19,9-17
[2] K Bhuyan and N N Dass, Indian Journal Of Chemistry,1990,29A,376-378
[3] B E Buckak and K C Ramey, Polymer Letters Edition Vol 14,401-405 (1976)

1. Deconvolution

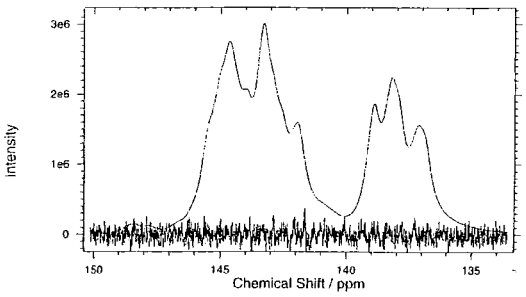
By designing a series of trials where the overall width of the applied bandshape was varied for each of a suitable range of values of sigma[a], a graph of optimum widths against input sigma could be plotted i.e. a sigma profile. The optimum total bandwidth and corresponding value of sigma was then determined by estimating the point of inflection.



Assessment of output diagnostics from subsequent trials based on this sigma profile, i.e. evidence profiles, gives rise to the optimum left and right width for the applied point spread function (PSF). Similarly, evidence profiles have been used to determine the shape parameters for the applied PSF.

Experience, with data of this quality, has shown that the width of the input PSF can be within ~ +/- 10% of the optimum without having a detrimental effect on the output deconvolution, especially when it is considered that the data is likely to contain more than one bandshape as a result of different mobilities within the polymer. It should be noted that each spectral region has been processed independently with sigma profiles being generated, where necessary, for each.

Aromatic Region: Mock Data and Residuals



[a] Sigma numerically defines the variation of the noise in the data combined with any mismatch between the applied point spread function (PSF) and the actual peakshapes in the data.

Group	^{13}C Chemical Shift / ppm	Cumulant (+/- one standard error)	Total cumulant for group (+/- one standard error)
1	136.87	5.27 +/- 0.63	24.21 +/- 0.19
	137.20	4.44 +/- 0.37	
	137.94	6.75 +/- 0.75	
	138.26	7.79 +/- 0.92	
2	138.91	9.56 +/- 0.22	9.56 +/- 0.22
3	141.91	5.53 +/- 1.79	5.53 +/- 1.79
4	142.54	4.44 +/- 0.26	55.61 +/- 0.26
	142.92	5.28 +/- 0.46	
	143.29	12.96 +/- 0.47	
	143.95	6.01 +/- 0.35	
	144.39	4.42 +/- 2.27	
	144.86	10.27 +/- 1.63	
	145.06	6.61 +/- 0.92	
	145.49	5.52 +/- 0.31	

Following references [2] and [3] the highlighted peaks are assigned to the polymer block ends.

These errors and an assessment of the residuals, i.e. experimental data minus MaxEnt calculated spectrum (free from noise), enable an assessment to be made as to the reliability of any feature in the MaxEnt result. Clearly, from the above figure the quality of the MaxEnt result is such that the intensity of the residuals is very close to that of the noise indicating that the MaxEnt calculated spectrum is in close agreement with the experimental data.

2. Verification Of MaxEnt Result

In order to verify the MaxEnt result the polymer composition is calculated from the MaxEnt intensities based on the literature assignments. It should be noted that reassignment of the block ends does not affect the calculation of composition.

75% Styrene ex Aldrich

Triad	Total Cumulant
MSM	24.21 +/- 0.19
SSM	9.56 +/- 0.22
MSS	5.53 +/- 1.79
SSS	55.61 +/- 0.26

Total Maleic = 0.5 x (2MSM +SSM + MSS) = 31.76
Total Styrene = SSS + MSS +SSM + MSM = 94.91

giving **74.9% Styrene**

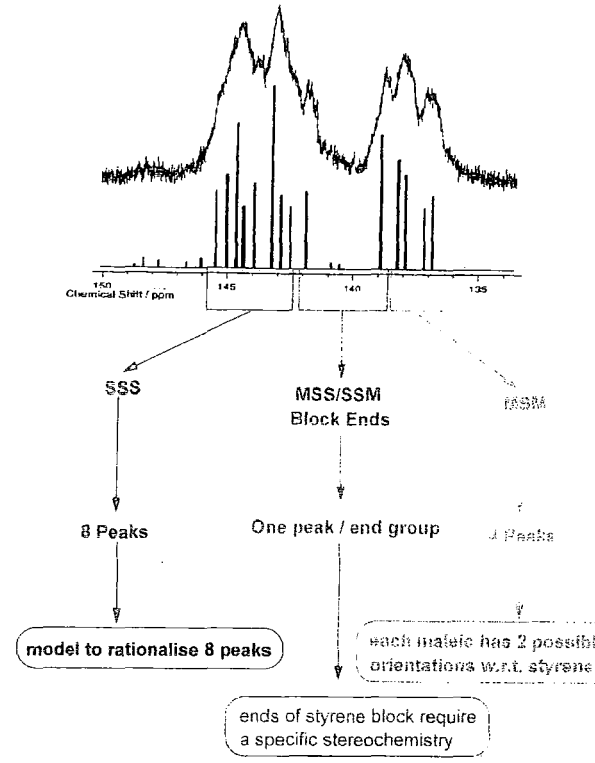
This is in good agreement with Aldrich, ^{1}H NMR and Mass-Spectrometry results and has therefore given us increased confidence as to the reliability of the MaxEnt deconvolution.

3. Exploration Of The MaxEnt Result

Stereochemistry

Assuming that the structure within each of the stereochemical groups, identified in the figure right, is largely the result of tacticity, leads to the conclusion that there is a strong correlation between microstructure and tacticity i.e. there is only one preferred conformation for the triads MSS and SSM, but four conformations for the triad MSM.

This in itself is evidence for a second order (or greater) Markovian model of polymer microstructure.



Microstructure

A first order Markovian model would require that the values of $P_{M/SS}$ and $P_{M/MS}$ were equal.

$$P_{M/SS} = A_{SSM} / (A_{SSM} + A_{SSS})$$

$$P_{M/MS} = A_{MSM} / (A_{MSM} + A_{MSS})$$

where A is the peak area.

However, using either of the assignments given below this is not the case. Therefore, a second order (or greater) Markovian model is required to describe the polymerisation process.

Calculation Of Composition Assuming Second Order Markovian Statistics.

From second order Markovian Statistics [4],

$$P_S / P_M = 1 + (A_{MSS} / A_{MSS} + A_{MSM}) (A_{SSM} + A_{SSS} / A_{SSM})$$

where P_M is the proportion of Maleic Anhydride in the polymer
P_S is the proportion of Styrene in the polymer

Number Of Standard Errors On Peak at 141.91ppm	% Styrene Based On Literature Assignment	% Styrene Based On Reversed Assignment
-2	60	90
-1	66	84
0	69	80
1	72	77
2	74	75

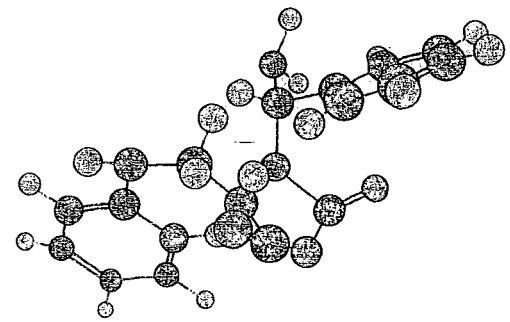
Chemical Shift / ppm	Literature Assignment	Reversed Assignment
138.91	SSM	MSS
141.91	MSS	SSM

Allowing two standard errors on the peak at 141.91ppm either assignment gives adequate agreement with a second order Markovian Model. This model will be checked using:
a) Massive Inference Algorithm
b) 125 MHz NMR

4. Rationalisation Of The Model

Aromatic Region Of The Spectrum

The literature assignments are claimed to be based on substituent additivity effects and statistical considerations. The problem with the assignments is that the authors assume that the β substituent will have the larger effect, ignoring the through space effect of the γ substituent. (They also ignore the direction of the through bond effect)



[4] J L Koenig, Chemical Microstructure Of Polymer Chains, Wiley-Interscience, ISBN 0471 07725-9

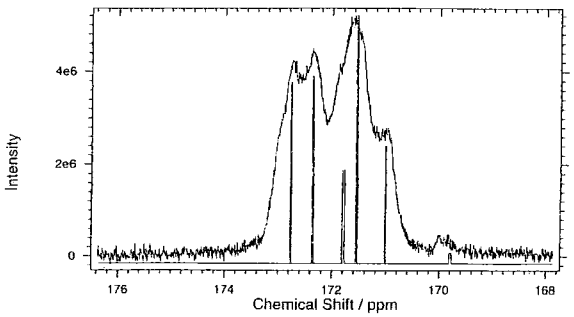
Demonstration of Consistency

Carbonyl Region Of The Spectrum

The MSM triad region of the C_1 aromatic spectrum showed four peaks attributed to each maleic anhydride having two possible orientations with respect to the styrene. If this is the case then evidence for this should be seen in the carbonyl region of the spectrum. The MaxEnt result shows five major peaks and one minor peak, summarised in the table.

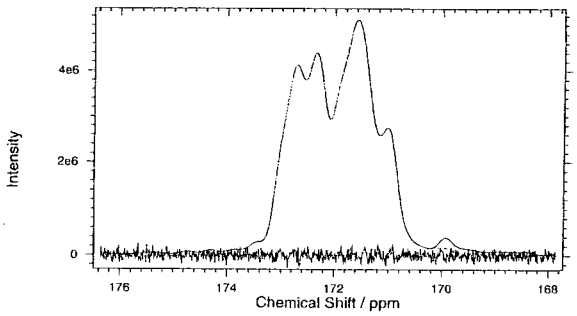
Chemical shift /ppm	Intensity (% of overall cumulant)	Assignment
169.8	1.3 +/- 0.3	Unknown Minor impurity
171.0	13.9 +/- 0.4	P1
171.5	29.1 +/- 2.5	P2
171.8	11.1 +/- 2.7	Impurity Probably Succinic Anhydride
172.4	22.0 +/- 0.6	P3
172.8	21.3 +/- 0.6	P4

^{13}C NMR Spectrum of Carbonyl Region With MaxEnt Deconvolution Overlayed.



The peaks of interest to the discussion of stereochemistry are those at 171.0, 171.5, 172.4 and 172.8ppm. These four peaks, attributed to carbonyl groups in the SMS triads, can be divided into two groups corresponding to the two different carbonyl environments. Hence, peaks P1 and P2 are grouped together, likewise peaks P3 and P4 (note the total intensity within each group is the same). Peaks P1 and P2 can be attributed to the carbonyl group MA1 (see structure) on the basis of the greatest steric effect (the effect of the substituent is seen in both chemical shift and stereochemistry, the latter being indicated by the greater effect on the relative intensities of the peaks). P3 and P4 are therefore assigned to carbonyl group MA2.

Carbonyl Region: Mock Data and Residuals



The quality of the deconvolution is indicated by the lack of structure found in the residuals and the small errors on the peak intensities.

Conclusions

1. The ^{13}C NMR spectrum of Styrene / Maleic Anhydride has been successfully deconvoluted using the Memsys 5 algorithm. The quantitative results of the deconvolution have been shown to be consistent with the known composition.

2. Based on the Memsys 5 output a microstructural model has been derived.

3. The assignments given in this current work have been shown to be physically consistent with the through bond effects of the γ substituent.

4. The deconvolution of the carbonyl region of the spectrum supports the reassignments presented in this work