Tina Memo No. 2001-014
Presented at a joint meeting of the Royal Statistical Society and the BMVA, 2001.
IVC Special Edition: The use of Probabilistic Models in Computer Vision., 21, 851-864, 2003.
Additional Appendix added to cover the problems of data fusion from Bayesian modules.

# Bayesian and Non-Bayesian Probabilistic Models for Image Analysis

P.A. Bromiley, N.A. Thacker, M.L.J. Scott,
M. Pokrić, A.J. Lacey and T.F. Cootes

Last updated
25 / 5 / 2007




Imaging Science and Biomedical Engineering,
School of Cancer and Imaging Sciences,
University of Manchester, Stopford Building,
Oxford Road, Manchester M13 9PT, U.K.

# Bayesian and Non-Bayesian Probabilistic Models for Image Analysis

P.A. Bromiley[1], N.A. Thacker[2] M.L.J. Scott,
M. Pokrić, A.J. Lacey and T.F. Cootes
Imaging Science and Biomedical Engineering,
School of Cancer and Imaging Sciences,
University of Manchester, Stopford Building,
Oxford Road, Manchester M13 9PT, U.K.

**Abstract**

Bayesian approaches to data analysis are popular in machine vision, and yet the main advantage of Bayes theory, the ability to incorporate prior knowledge in the form of the prior probabilities, may lead to problems in quantitative tasks. In this paper we demonstrate examples of Bayesian and non-Bayesian techniques with the use of selected examples from the area of magnetic resonance image (MRI) analysis. Issues raised by these examples are used to illustrate common difficulties in Bayesian methods and to motivate an approach based on frequentist methods. We believe this approach to be more suited to quantitative data analysis, and provide a general theory for the use of these methods in learning (Bayes risk) systems and for data fusion. Proofs are given for the more novel aspects of the theory. We conclude with a discussion of the strengths and weaknesses, and the fundamental suitability, of Bayesian and non-Bayesian approaches for MRI analysis in particular, and for machine vision systems in general. Overall we advise caution regarding the common assertion that the best approaches to all machine vision problems are necessarily Bayesian.

## 1 Introduction

This paper discusses the use of Bayes theorem in decision systems which make use of image data. We concern ourselves only with the use of the equation

$$P(H_i|data) = \frac{P(data|H_i)P(H_i)}{\sum_i P(data|H_i)P(H_i)} \tag{1}$$

for the interpretation of mutually exclusive hypotheses $H_i$, as the basis for algorithmic design. We compare such approaches with alternatives based on frequentist statistics. In the broadest possible terms, these two schools of statistical inference vary in their definition of a probability. Frequentist statistics demands that a probability should represent a genuine reflection of the frequency of occurrence of some event, whereas Bayesian statistics defines a probability as a degree of belief, allowing the incorporation of prior knowledge in the form of the prior probabilities.

Bayes theorem is a cornerstone of modern probabilistic data analysis. It is used as a way of constructing probabilistic decision systems so that prior knowledge can be incorporated into the data analysis in order to "bias" the interpretation of the data in the direction of expectation. The prior probabilities therefore have the greatest influence when the data under analysis is unable to adequately support any model hypothesis. Under these circumstances direct application of a purely data driven solution may not be directly possible. The use of Bayes theorem can appear to provide spectacular improvements in the interpretation of data. However, despite their popularity and widespread acceptance, there are often significant practical problems in the application of Bayesian techniques.

Firstly, Bayesian approaches use information regarding the distribution of a whole group of data to influence the interpretation of a single data set. Few would countenance modifying the estimate of a random variable by averaging with some weighted combination of the group mean, as it is well known that such a procedure would introduce bias, yet analogous procedures are accepted as reasonable in Bayesian approaches to data analysis. Another aspect of this problem is the suppression of infrequently occurring data or "novelty". Clearly, skepticism concerning the use of Bayesian approaches in areas such as medical data analysis, where pathological cases are often unique, would have some justification. We explain the effects of **bias and novelty** in Bayesian estimation in Section 2, using multi-dimensional MR volumetric measurement as an example.

[1]paul.bromiley@manchester.ac.uk
[2]neil.thacker@manchester.ac.uk

Many researchers have concentrated on the source of prior probabilities [16, 3]. Ideally this prior information could be established uniquely for a particular task, preferably emerging from a mathematical analysis. However, if Bayes theory is to make predictions regarding the likely ratio of real world events it must be accepted that the data samples used in Bayesian approaches must enforce the priors i.e. must represent a stratified random sample of the types of data under analysis[3]. Thus, in the absence of a deterministic physical mechanism giving rise to the data set there can be no theoretical justification for belief in the existence of a unique prior. If the circumstances in which the system is used change the expected prior distributions must also change. This situation is often encountered in data analysis problems, particularly those involving biological data sets where data are selected on the basis of rules that can vary with time.

In order for Bayes theory to be applied correctly the likelihood distributions ($P(data|H_i)$) of all possible interpretations $H_i$ of the data must be known. Unfortunately, in many practical circumstances, particularly in research, these distributions are not well known a-priori. Much of the computer vision community accepts the need to construct systems which learn in order to tackle difficult scene interpretation tasks. It could be regarded as a weakness if the computational framework used in a learning system demands that all possible interpretations of the data are available before a useful statistical inference can be drawn.

Beyond simple classification, any decisions based on Bayesian classification results should also be made on the basis of the Bayes risk, in order to minimise the cost of the decision. The aim in clinical support systems, for example, is to provide treatment which improves the prognosis of the patient, rather than just to provide the correct diagnosis. To illustrate the problems of **learning, priors and Bayes risk**, in Section 3 we give a specific example of a Bayesian system designed to evaluate the degree of atrophy in the brain arising from a variety of dementing diseases, and explain some possible solutions. We conclude that Bayesian decision systems cannot form useful components of learning systems without modifications that distance them from the original theory. We explain how both this and the previous example illustrate the difficulty of using Bayes theory for quantitative analysis, even on relatively simple problems.

If Bayes theory cannot be used directly to provide a useful diagnostic classification, a more appropriate method for presenting results for clinical interpretation must be found. We present one potential solution in the form of a single-model statistical decision system which has its origins in the so-called "frequentist" approaches to data analysis. The use of only one likelihood distribution avoids the need to specify prior probabilities and sidesteps issues of complexity. This can result in an approach to data analysis which is more in line with the need to construct systems which can learn incrementally, and yet still be capable of generating useful results at early stages of training. In Section 4 we illustrate a **single model statistical analysis** technique for the problem of change detection, under circumstances where the statistical model can be bootstrapped from the image data. This represents a significant step towards learning. We further describe a practical mechanism for **data fusion** within this framework and in particular provide a derivation for problems of arbitrary dimensionality. These systems generate data which is more quantitative than that generated by Bayesian methods, yet the data remains suitable for use in a Bayes risk analysis. Therefore, we believe that such systems have advantages over Bayesian algorithms and have great potential for use in computer vision.

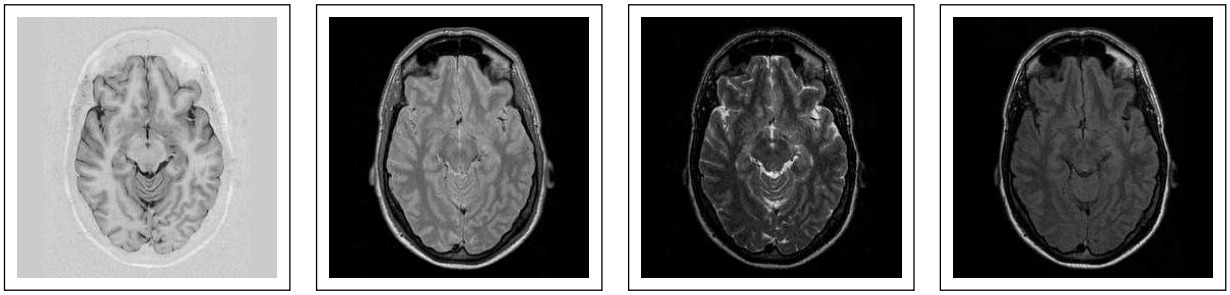
# 2 Bias and Novelty in Multi-dimensional MR Image Segmentation

Magnetic resonance imaging represents a very flexible way of generating images from biological tissues. From the point of view of conventional computer vision, analysis of these images is relatively simple as, for a given protocol, particular tissues generate a narrow range of fixed values in the data. Bayesian probability theory has been applied by several groups in order to devise a frequency-based representation of different tissues [13]. The conditional probabilities provide an estimate of the probability that a given grey level was generated by a particular part of the model. The conditional probability for a particular tissue class given the data can be derived using the knowledge of image intensity probability density distributions for each class and their associated priors.

The data density distributions are often assumed to be Gaussian, but for many clinical scans there is a high probability that individual values may be generated by fractional contributions from two distinct tissue components (partial voluming). In [19] we adopted a multi-dimensional model for the probability density functions which can take account of this effect. The conditional probability that a grey level $\mathbf{g}$ is due to a certain mechanism $k$ (either a pure or mixture tissue component) can be calculated using Bayes theory,

$$P(k|\mathbf{g}) = \frac{d_k(\mathbf{g}) f_k}{f_0 + \sum_i (d_i(\mathbf{g}) f_i) + \sum_i \sum_j d_{ij}(\mathbf{g}) f_{ij}} \tag{2}$$

[3]Strong Bayesians would argue that we can deal with degrees of belief, but then we have to accept that the computed probabilities do not necessarily correspond to any objective prediction of likely outcome.



(a) (b) (c) (d)

Figure 1: Image Sequences: IRTSE (a), VE(PD) (b), VE(T2) (c), and FLAIR (d).

where $d_k(\mathbf{g})$, $d_i(\mathbf{g})$, and $d_{ij}(\mathbf{g})$ are the multi-dimensional probability density functions for tissue component $k$, pure tissue $i$, and a mixture of tissues $i$ and $j$ respectively. The corresponding priors, $f_k$, $f_0$, $f_i$ and $f_{ij}$, are expressed as frequencies (i.e. the number of voxels which belong to a particular tissue type) whether pure tissues or a mixture of tissues. Note that in the first departure from a pure (fully justifiable) Bayesian model a fixed extra term $f_o$ (making an arbitrary assumption of uniform distribution) is included to model infrequently occurring outlier data [12].

The parameters of the model, such as covariance matrices, mean vectors and priors, can be iteratively adjusted by maximising the likelihood of the data distribution using Expectation Maximisation (EM) algorithm [27] (Appendix A). Once the data density models are obtained, the conditional probabilities can be calculated and probability maps derived for each tissue type, which estimate the most likely tissue volume fraction within each voxel.

The probabilistic segmentation algorithm has been implemented and tested on co-registered MRI brain images of different modalities chosen for their good tissue separation and availability in a clinical environment. The use of multi-spectral data enables decorrelation of statistical distributions and better estimation of partial volumes. The images used were variable echo proton density (VE(PD)), variable echo T2 (VE(T2)), inversion recovery turbo spin-echo (IRTSE), and fluid attenuated inversion recovery (FLAIR) (see Fig. 1). These images provide good separation between air and bone, fat, soft tissue (such as skin and muscle), cerebro-spinal fluid (CSF), grey matter (GM), and white matter (WM). Fig. 2 shows a scatter plot of the IRTSE and VE(PD) images, together with the model after 10 iterations of the EM algorithm. The final model agrees well with the original data. The partial volume distributions link the otherwise compact pure tissue distributions along the lines between them in accordance with the Bloch equations, which describe the signal generation process in MR. The final segmentation result is represented by probability maps for each tissue class and can be seen in Fig. 3. The probability maps range from 0 to 1 and can be used for boundary location extraction (e.g. a probability of 0.5 represents the boundary location between two tissues) or volume visualisation [15].

| | Air and bone | Soft tissue | Fatty tissue | Cranial Fluid | Grey Matter | White Matter |
|---|---|---|---|---|---|---|
| Air/bone | 16012.2 | 1032.0 | 21.9 | 126.7 | 371.5 | 0 |
| Soft Tissue | 1032.0 | 4076.9 | 1219.3 | 520.4 | 46.2 | 0 |
| Fatty tissue | 21.9 | 1219.3 | 1517.8 | 0 | 0 | 0 |
| Cranial Fluid | 126.7 | 520.4 | 0 | 445.9 | 759.2 | 84.1 |
| Grey Matter | 371.5 | 46.2 | 0 | 759.2 | 6465.8 | 4105.6 |
| White Matter | 0 | 0 | 0 | 84.1 | 4105.6 | 3548.7 |

Table 1: Typical priors assigned to each class (pure and mixed tissues). Zero values are fixed to eliminate biologically implausible combinations.

Table 1 gives the priors (i.e. number of voxels) assigned to each class (pure and mixture of tissues) to model different tissue types. As these values represent genuine frequencies of tissues they will change depending upon the region selected, and so too will any estimates of tissue proportion. For example, if the intention was to segment only the brain tissues (i.e. CSF, WM and GM) a region of interest could be chosen which contained only these tissues.

Fig. 4 illustrates how change in the region chosen to generate the priors may lead to significant change in the

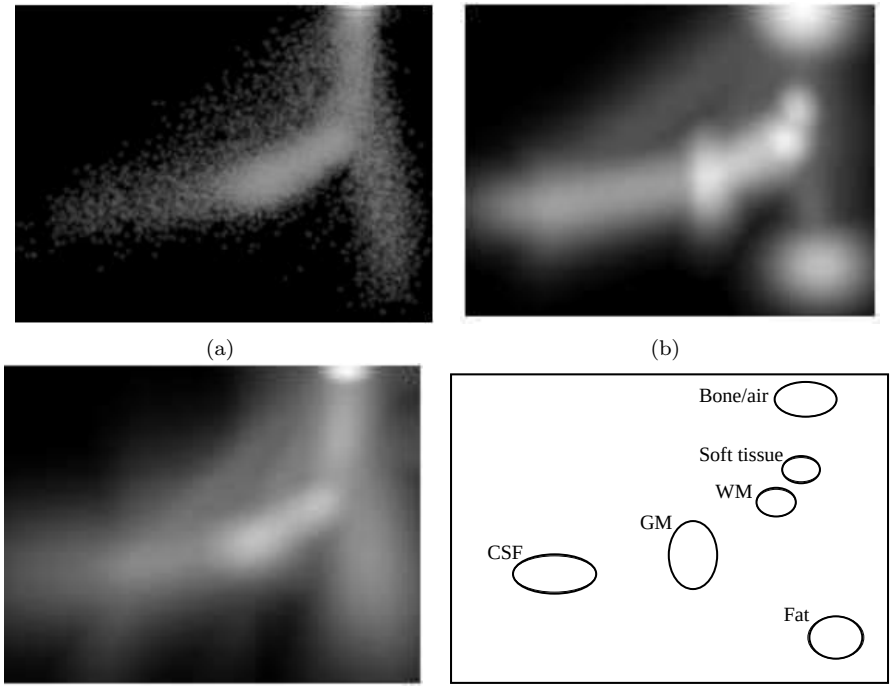



Figure 2: Scatter plots of IRTSE vs. VE(PD): original data (a); model using initial values of parameters (b); model after 10 iterations of EM algorithm (c); and schematic showing the origins of the data clusters (d).

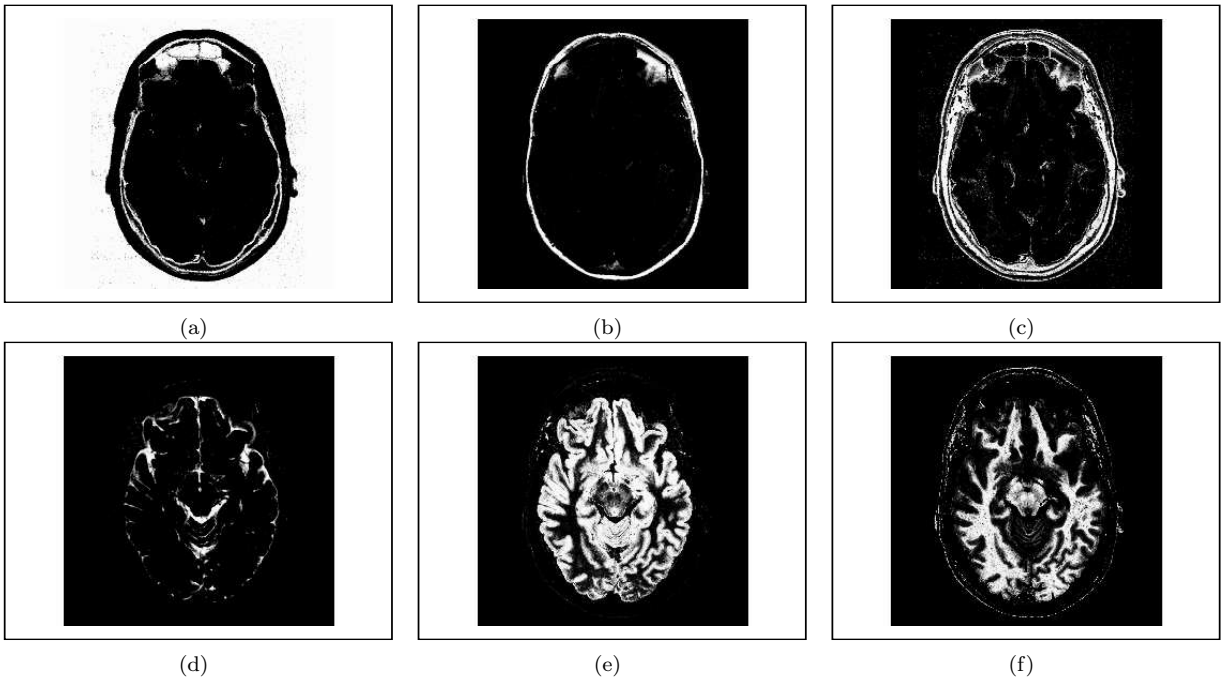
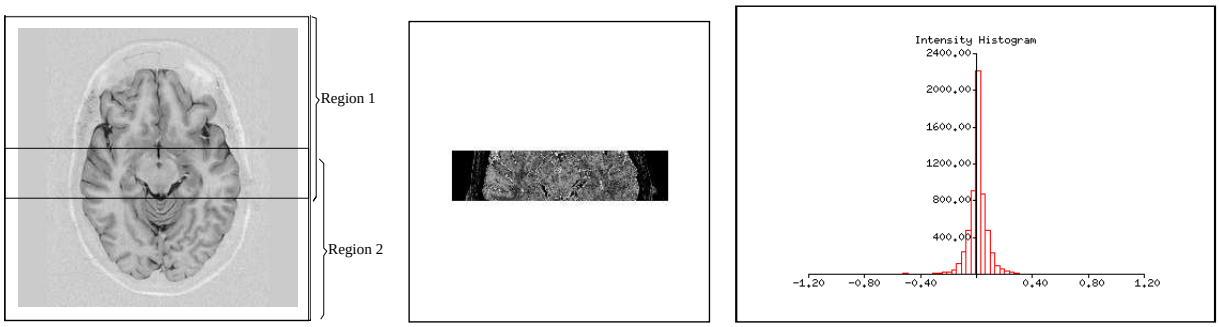



Figure 3: Probability maps for bone and air (a), fat (b), soft tissue (c), CSF (d), GM (e), and WM (f).

subsequent interpretation of the same data. Two overlapping regions of interest were defined, generating two sets of prior probabilities for the same region (their intersection). A subtraction of the two grey matter probability maps produced for this region shows significant differences between them: it can be seen from the histogram that the probabilities differ at approximately the 10% level.

Given the arbitrary nature of the process of region selection, it is evident that the prior information is not unique




(a) (b) (c)

Figure 4: Regions of interest (a) which provide different prior probabilities for the same region (their intersection), the result (b) of subtracting the grey matter probability maps for this region, and the histogram (c) of this subtraction result.

for a particular segmentation task. This exemplifies the prior selection problem. The common response to this, that the results will not change much over a large range of priors, is clearly over-simplistic (and unrealistically optimistic) as particularly important tissues can be eliminated altogether by careless selection of the bootstrap region. Unexpected data (novelty) will be suppressed in the analysis, and attempts to measure quantitative changes between two data sets will be affected by any bias introduced by inconsistent priors. Practical use of Bayes theory in such problems therefore requires careful control of prior estimates, particularly for quantification of small differences between data sets [25]. Solutions such as fixing the priors to be consistent between data sets represent a second arbitrary (and unsatisfactory) extension to the theory.

One possible approach to this problem, suggested by Laidlaw [17], involves estimating the priors locally to each pixel value. Unfortunately this does not extend to the use of partial volume models. Any region of grey-level values we observe can be attributed to the partial volume terms without the need for any pure tissue components. Such problems raise doubts about the validity, or even the need, to incorporate the priors at this stage of the analysis.

# 3 Learning, Estimation of Priors and Bayes Risk in Diagnostic Classification of Dementing Diseases.

In the previous example the suitability of Bayes theory for quantitative estimation was considered and the difficulties in obtaining quantitative measurements were discussed. However, a more conventional use of Bayes theory is in the classification of data. One of the important tasks in medical image analysis involves informing clinicians of the most likely interpretation of a large or complex data set for the process of decision support. In magnetic resonance imaging of the brain, for example, we may wish to take the data description (generated by the previous system) and perform an analysis of structure to identify abnormality or determine a categorisation. In previous work [22, 23] we have designed a system which is capable of diagnosing dementing diseases based on the pattern of atrophy in the brain. This is achieved through the analysis of cranial fluid volumes (such as may be achieved using the technique in the previous section) within a standardised co-ordinate system. After correction for head size and normal ageing[4] and care to represent the data in a way which takes correct account of the Poisson measurement process [24], twelve measurements of corrected volume were used in a simple Parzen classifier to estimate the probability of class assignment between one of four groups: normal; Alzheimers disease; Fronto-Temporal dementia; and Vascular dementia. This classification process makes direct use of Bayes theory and typical results are given in Table 2. Given the difficulty of clinically identifying subjects within these groups using psychometric tests, these results illustrate a separation between classes that appears sufficient to provide useful diagnostic information.

The use of Bayes theory once again requires the specification of a prior probability, which for this illustration has simply been assumed to be proportional to the sampled frequency within the data set. These prior terms establish the relative frequency of each model hypothesis and without them any classification result will be sub-optimal, in the sense that there will not be a minimum number of incorrect classifications across a sample group. In a clinical diagnostic task the prior probabilities are determined by the statistical make-up of the classified data sample. This process is often referred to as case adjustment.

The use of prior probabilities therefore requires a solution to the additional problem of ensuring that the prior

[4]The normal brain loses tissue which results in an increase of cranial fluid at a rate of typically 1% per year after the age of 40.

| Disease | Norm. | F.T.D. | Vas.D. | Alz. |
|---|---|---|---|---|
| $Normal$ | 7 | 2 | 8 | 1 |
| $Fronto-Temporal$ | 5 | 21 | 3 | 7 |
| $Vascular$ | 3 | 2 | 13 | 4 |
| $Alzheimers$ | 1 | 3 | 6 | 28 |

Table 2: Disease (rows) vs classification (columns) for a cross-validated Parzen classifier.

probabilities reflect the true frequency of occurrence of cases. Constructing a system with fixed priors based on a national average would be sub-optimal in any location that did not reflect this demographic, and even regional averages may vary over time. If the intention was to build an optimal, fully automatic classification system, then one which attempts to determine prior probabilities from the sample data as it arrives could be envisaged. However, if it is accepted that for moral reasons this data should be moderated by a medical expert, such a system might be considered inappropriate for the reasons outlined below.

The aim in any clinical environment must be to deliver treatment which improves the prognosis of the patient, rather than simply to obtain the correct diagnosis. Therefore, any decision based on the results from a diagnostic classification system should be made on the basis of the Bayes risk. For example, if the diagnosis is ambiguous between three possibilities, but two outcomes could have the same treatment, then this should influence the patient management. This assessment must be carried out by the expert, through a process of weighting decision support information with the experts own experience or other data. However, any system that can change its classification of the same data set over time (non-stationarity) due to the frequency of occurrence of other diseases will complicate experiential learning by the expert. Efficient learning therefore requires knowledge of the objective information content in the data, and expecting the expert to infer the effects of time-varying prior probabilities, rather then explicitly providing this information, is inviting difficulty. This exemplifies the difficulty of making quantitative use of data from a Bayesian module in a larger system.

Any system in which the priors are not fixed will not meet the requirement for stationarity which is an essential prerequisite for efficient learning and data fusion, whether involving a learning system or a clinician. One alternative is to construct a Bayesian classifier with equal prior probabilities, and to train the expert in how best to make use of this data. However, it is then difficult to believe that the outputs from the system contain any meaningful information beyond that present in the likelihoods. We must therefore conclude that the likelihoods provide the most appropriate way to present data for clinical interpretation. A clinician would then be in a position to make use of either their own experience, or a separate estimate of the current expected relative frequencies of diseases, in order to recommend treatment on the basis of risk. In fact, if experiential learning of the clinician is based upon likely outcome, specific calculation of probabilities of diagnostic classification may be considered an unnecessary diversion, making the specification of prior probabilities irrelevant.

# 4 Single Model Statistical Analysis for the Identification of Change in Magnetic Resonance Brain Images

The previous examples illustrate the difficulties encountered in determining all of the information needed to use Bayes theory correctly, and the consequences in terms of quantitative measurement. One approach to resolve these difficulties would be to attempt to construct statistical questions regarding expected data distributions that can be addressed without knowledge of the priors, or with fewer model components. Logically the minimum number of model components required would be one, and in that case the only quantity that can be obtained is the relative probability that a particular item of data was generated by the model. However, this is enough to identify data that is unlikely to be generated by the model (i.e. outlier detection).

There are several standard statistical techniques designed to operate using only one model, for example null hypothesis tests and the chi-squared probability. Although such techniques have already been applied widely in MR data analysis [10] these generally assume particular data density distributions which will not be applicable to arbitrary problems. Image analysis tasks frequently involve large amounts of data, and our recent work [4, 5, 6] illustrates how we might exploit this to bootstrap a statistical model of data behaviour from the data itself. Thus the technique does not require additional model components or prior probabilities, and avoids the need to explicitly build the single model. In addition, this technique produces output with a flat probability distribution, which provides routes to both self-test and data fusion, as will be discussed in the next section.

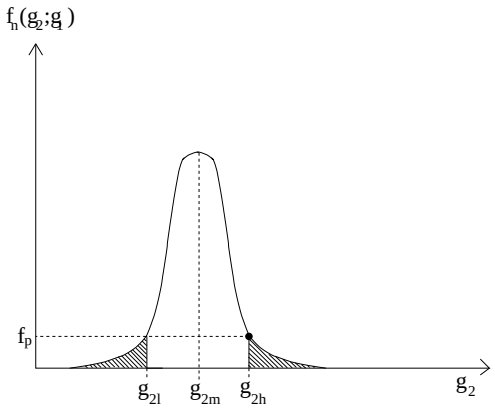



Figure 5: For any pair of corresponding pixels from the original images (the black point), the integration (the shaded region) is performed across all values smaller than $f_p$, the value at the point defined by the original image pixels.

The technique discussed here is designed to construct non-parametric models in order to estimate the probability that a particular item of data was generated by a particular process. It defines a probability that reflects how likely it was that the grey level values from corresponding pixels in an image pair were drawn from the same distribution as the rest of the data. A scattergram drawn from a sample of image data $S(g_1, g_2)$ is used as a statistical model of data behaviour. Taking a vertical cut through the scattergram identifies a set of pixels in the first image which all have the same grey-level value $g_1$. The distribution of data along this cut $f(g_2; g_1)$ gives the grey levels $g_2$ occurring at the corresponding pixels in the second image. If the scattergram is normalised along all vertical cuts, then these distributions become the probability distributions for the grey level value in the second image given the grey level value in the first,

$$\frac{f(g_2; g_1)}{\int_{-\infty}^{\infty} f(g_2; g_1) dg_2} = f_n(g_2; g_1). \tag{3}$$

Then, corresponding pairs of pixels from the original images are taken, and their grey levels used to find their coordinates in the normalised scattergram. An integration is then performed along the vertical cut passing through that point, summing all of the values smaller than the value at that point, $f_p$, as shown in Figure 5. The result is the probability $\varepsilon$ of finding a more uncommon pairing of grey levels, given the grey level in the first image $g_1$, than that seen at the original pixel pair,

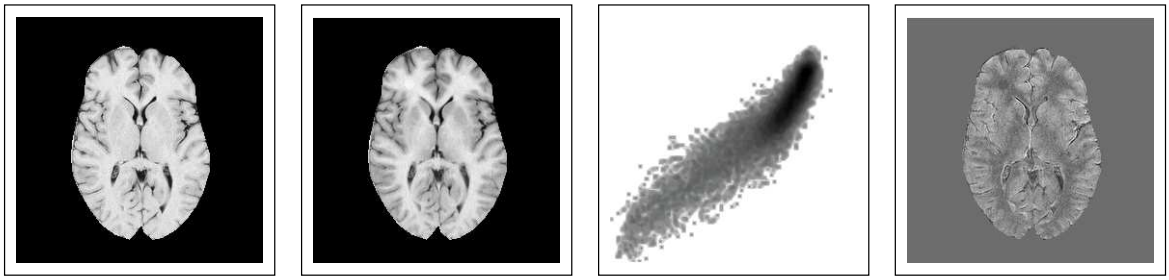
$$1 - \int_{g_{2l}}^{g_{2h}} f_n(g_2; g_1) dg_2 = 1 - P(g_{2l} < g_{2m} < g_{2h}; g_1) = \varepsilon, \tag{4}$$

where $g_{2m}$ is the mean grey level in the second image at pixels on this cut in the scattergram, and $g_{2l}$ and $g_{2h}$ are the limits of the integral. This follows directly from the original definition of a confidence interval, due to Neyman [18]. In addition, the ordering principle implicit in the technique results in the shortest possible confidence interval [9] [5].

The result of the integration is used as the grey level for the corresponding pixel in a difference image. Since it depends on the mean grey level for the pixels on this cut in the second image, any process which results in global differences between the images, such as a change in the level of illumination, will be ignored. The grey level values in the difference image relate directly to the frequency of occurrence of the pairing of grey level values seen at the corresponding pixels in the original images. This is exactly the type of measure needed to identify outlying combinations of grey-level values in a fully automatic manner. The technique therefore illustrates that useful and quantitative statistical analyses can be performed through comparison with a single statistical model bootstrapped from data itself.

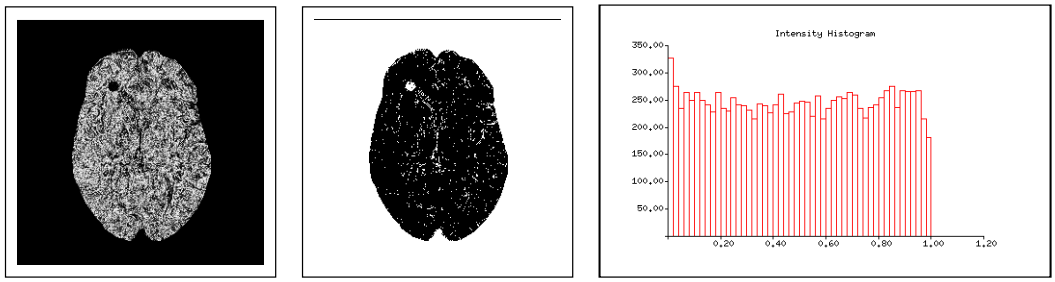
An important feature of this method is that, since the integration along vertical cuts in the scattergram performs a probability integral transform, the resulting difference image should by definition have a uniform probability distribution i.e. a histogram of the grey levels in the difference image will be flat. It follows that thresholding the difference image at some level $n$ will extract the $100n\%$ of the pixels that showed the most uncommon pairings of grey levels in the original images. Thus the estimated probabilities correspond to a genuine prediction of data frequency. Probabilities with these characteristics have previously been referred to in the literature as honest [8].

[5] Unfortunately a discussion of this and other related issues is beyond the scope of this paper



(a) (b) (c) (d)

Figure 6: The original MRI brain images (a,b), scattergram (c), and simple subtraction difference image (d), with a $2\sigma$ offset added to a small region of image (b).




(a) (b) (c)

Figure 7: The difference image produced by the new method (a) showing the altered region in the upper left. Thresholding at the 10% level (b) extracts the low probability pixels. The histogram of this difference image (c) is by definition uniform.

The importance of this feature in relation to the work presented here is that knowledge of the expected distribution for the output provides a mechanism for self-test [20]. Further uses of this property are discussed in the next section.

A potential application of this technique is the detection of MS lesions in MRI scans of the brain, an important issue in relation both to tracking the progression of the subclinical disease, and to therapeutic trials [26]. Such lesions can be difficult to detect, but can be highlighted using a contrast agent (GdDTPA), which concentrates at the lesion sites. Scans taken before and after the injection can be subtracted to help identify lesions, but the presence of the contrast agent also alters the global characteristics of the scan, so a simple pixel-by-pixel subtraction will not remove all of the underlying structure of the brain from the image.

Obtaining a gold standard for this work is difficult without extensive histological investigation. In order to simulate the imaging process, two T2 scans with slightly different echo train times (TE) of the same region of the brain were used. This simulates the effects of repeat scanning on different scanners after a significant time interval and the small quantitative changes which occur in the signal due to the presence of a contrast agent. The background was removed from the image so that the statistical model (scattergram) was estimated using only the tissues of interest. A grey-level offset at twice the level of noise $\sigma$ in the original images, too small to be detected visually, was then added to a small circular region of one of the brain images, simulating lesions in a testable manner. The synthetic data is shown in Fig. 6. The subtraction routine was applied to this data in an attempt to detect the change. The altered region is barely visible in the pixel-by-pixel difference image shown in Fig. 6. Fig. 7 shows the difference image generated using the new method, and the altered region shows up clearly. The altered region ceased to be detectable when the magnitude of the offset was reduced below around $1\sigma$. Fig. 7 also shows a histogram generated from the difference image and as expected this method produced the required uniform probability distribution, confirming the applicability of this statistical measure to this MR data. The possibility of such self-test is an important feature for statistical methods which are intended for application to arbitrary data analysis tasks.

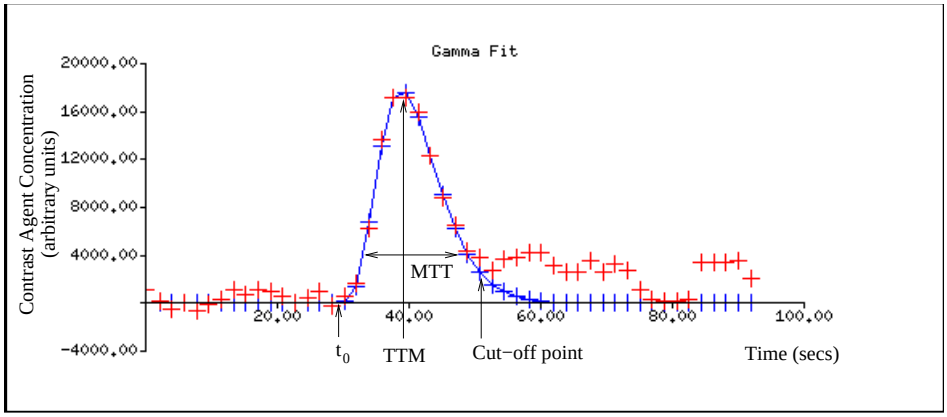



Figure 8: Gamma Variate fit of the concentration/time data for a bolus of contrast agent passing through the brain.

# 5 Data Fusion for Flow Abnormality Detection in Contrast Susceptibility Perfusion Data.

One of the major challenges in the construction of image interpretation systems is that of merging sources of information, or data fusion. Algorithms that produce output with a uniform probability distribution therefore confer an additional advantage, in that a standard technique exists to combine such distributions in a statistically rigorous manner [1], providing a route to data fusion. If j independent quantities $\omega$, each having a uniform distribution, are multiplied to produce a product $P$,

$$P = \prod_{i}^{j} \omega_i, \tag{5}$$

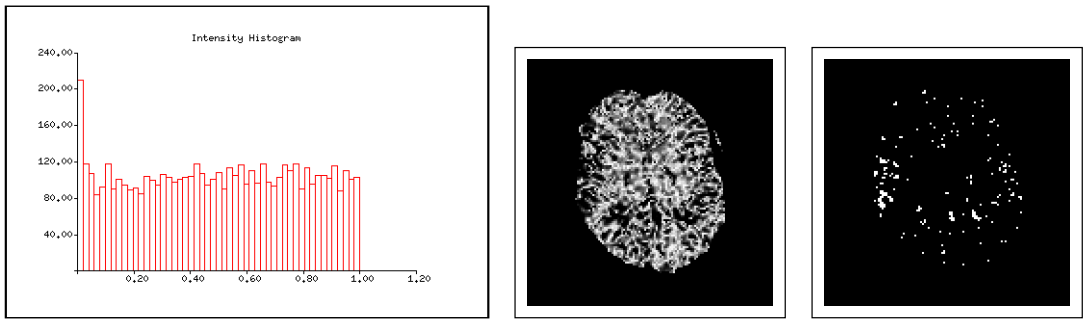
then this product can be normalised to produce a new quantity $P'$, which has a flat distribution, using

$$P' = P \sum_{i=0}^{j-1} \frac{(-\ln P)^i}{i!}. \tag{6}$$

This process is potentially nestable, providing a simple yet statistically principled method for data fusion. Although the above method can be derived easily for any fixed number of probabilities $j$, a general proof for any number of dimensions is needed before we can apply this to arbitrary problems. We have been unable to locate such a proof in the established literature so a general derivation for this equation for problems of arbitrary dimensionality is provided in Appendix C. This section demonstrates the use of this technique with an example involving the combination of independent maps extracted from dynamic MR images, to produce a single map showing all of the statistically significant information available.

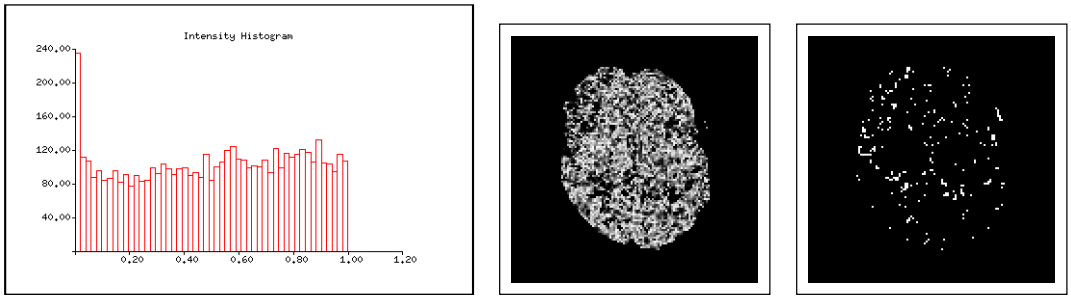
Dynamic susceptibility contrast-enhanced MR imaging can be used to image the passage of a bolus of contrast agent (Gd-DPTA) through the brain. The resulting temporal sequence of a particular slice through the brain can be processed to produce meaningful parameter maps of the rate and volume of blood passing through the voxels. This is done by fitting a Gamma Variate curve to the concentration of the contrast agent over time through a voxel/slice of the image [21, 2], as shown in Fig. 8. In order to exclude the recirculation of the contrast bolus from the fit, data after the point at which the agent concentration drops below a given fraction (in this case 20%) of the maximum is ignored. We hypothesise that the parameters of the Gamma curve can be approximated as if sampled from a Gaussian distribution, and if this is the case, then estimates of Relative Cerebral Blood Volume (RCBV) will be distributed like a Poisson random variable (Appendix B). Probability maps with flat probability distributions can be constructed, using conventional parametric techniques, for differences in both RCBV and Mean Time of Arrival (TTM) parameters. Figs. 9(b) and 10(b) show examples of such maps for a normal data set. Fig. 9(c) and Fig. 10(c) show the pixel maps of the 0-0.02 bin for the $\sqrt{CBV}$ and TTM error function distributions respectively. The near flat (uniform) distribution of probabilities (excluding fit failures) in these data, shown in Figs. 9(a) and 10(a), justify the initial model assumption.

To demonstrate that the method can detect genuine physiological change, we applied the method to scans of a patient before and after a surgical procedure to remove a carotid stenosis. In this case, genuine change between




(a) (b) (c)

Figure 9: Error function distribution on $\sqrt{CBV}$ difference map (a), probability map for the distribution (b) and pixel map of 0-0.02 bin for erf($\sqrt{CBV}$) difference map (c).




(a) (b) (c)

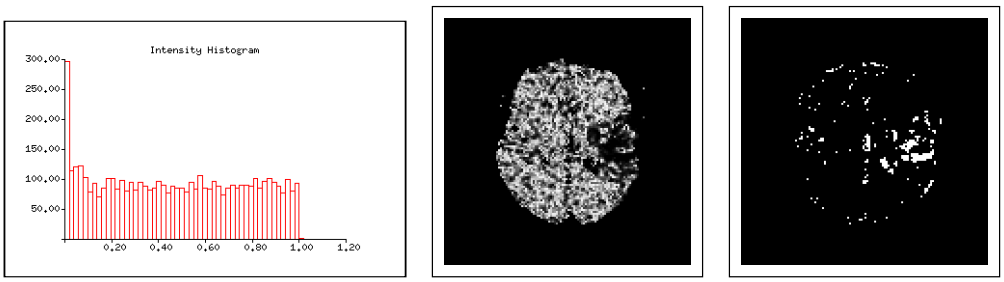
Figure 10: Error function distribution on TTM difference map (a), probability map for the distribution (b), and pixel map of 0-0.02 bin for erf(TTM) difference map (c).

the maps which is greater than that due to error would be expected. The error function distribution for the TTM difference map again has outliers in the 0-0.02 bin (Fig. 11(a)) but is otherwise flat. The pixel map of this bin (Fig. 11(c)) shows that most of the outliers are due to a change on the left (right for the observer) of the brain, most probably due to unblocking the affected carotid artery. The error function distribution of the $\sqrt{CBV}$ map (Fig. 12(a)) shows the same flat map with a peak, but the corresponding pixel map (Fig. 12(c)) does not show a gross change like the TTM map, but is more similar to the normal maps (Fig. 9(c), 10(c)). We believe that the changes seen on the $\sqrt{CBV}$ probability map (and corresponding pixel map) are predominantly due to perfusion changes in the grey and white matter, whereas those on the TTM map are due to changes in the time of arrival of the blood in the feeding arteries and draining veins.

The probability maps of the parameters $\sqrt{CBV}$ and TTM represent two physiological aspects of the same data, which we wish to combine in order to show the overall vascular differences pre- and post-operatively. Since the probability distributions of the maps are flat, the renormalisation technique described above can be applied to reflatten the product of the individual maps to produce a new map showing all of the statistically significant changes.

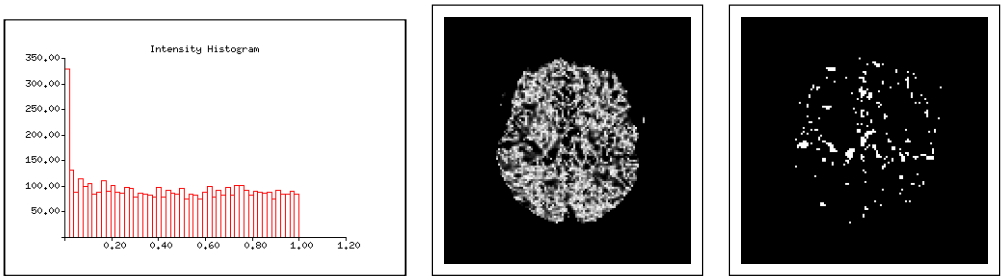
Fig. 13(a) shows that the combined re-flattened map for the carotid stenosis patient is flat (demonstrating that the $\sqrt{CBV}$ and TTM maps are independent) except for the peak in the 0-0.02 bin. Comparing the pixel map for this bin (Fig. 13(c)) with with those for the $\sqrt{CBV}$ and TTM shows that the information regarding perfusion differences has been preserved in the combined probability map (Fig. 13(b)) and that this map now contains all of the statistically significant information available.

In order to demonstrate the flexibility of this approach, we have also applied the method to the results from the non-parametric subtraction technique described in the previous section. A spatial correlation analysis can be performed by forming the product of the grey level of each pixel with the four nearest pixels. This is equivalent to forming the product of five images each having a uniform probability distribution, so the probability distribution of the product can be renormalised using the technique described above. However, the probability renormalisation technique assumes no spatial correlation, (see Appendix C). The probability values associated with pixels in the background, i.e. not in localised difference regions, will be randomly distributed and so will renormalise correctly. Localised differences will result in spatially correlated low probability pixels, and will produce low probability



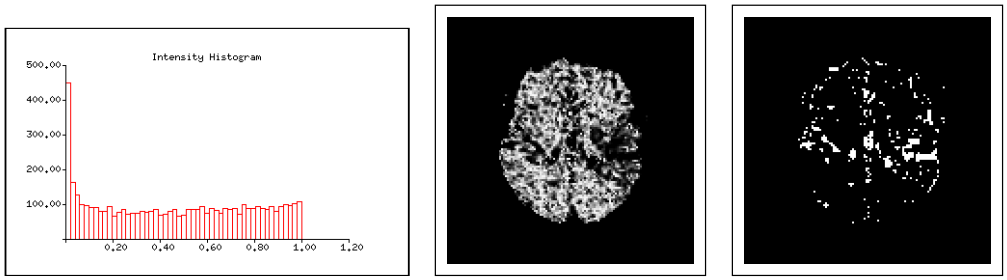

(a) (b) (c)

Figure 11: Error function distribution on patient TTM map (a), probability map for the distribution (b), and pixel map of 0-0.02 bin for patient erf(TTM) difference map (c).




(a) (b) (c)

Figure 12: Error function distribution on patient $\sqrt{CBV}$ map (a), probability map for the distribution (b) and pixel map of 0-0.02 bin for patient erf($\sqrt{CBV}$) difference map (c).




(a) (b) (c)

Figure 13: Distribution of the combined re-flattened probability map for the carotid stenosis patient (a), probability map for this distribution (b), and pixel map of the 0-0.02 bin (c).

products which will not renormalise correctly. Therefore, the probability distribution of the image showing the renormalised five-pixel product will feature a flat distribution for non-difference pixels, together with a spike close to zero containing the pixels in localised difference regions. This provides a simpler route to identification of the local difference regions when compared to the original non-parametric image subtraction result, in which the probability distribution is flat by definition for all pixels. Fig. 14 shows the result of applying this technique to the non-parametric image subtraction result for the synthetic MS data. The spatial correlation present in the data results in a reduction in the number of effective degrees of freedom [11], and so the histogram for the renormalised image shows the result of some degree of over-flattening. The localised difference region can immediately be extracted by thresholding at lower probabilities than would be used with the initial difference image, leading to less contamination by background pixels. In addition, since the distribution for background pixels is flat, the number of background pixels extracted in the threshold is known. Therefore, this number can be subtracted from the total number of pixels extracted in the threshold to leave the number of pixels in the localised difference regions, and so volumetric analysis of the difference regions can be performed. This implies that quantitative extraction can be performed using such methods without the need for prior probabilities.

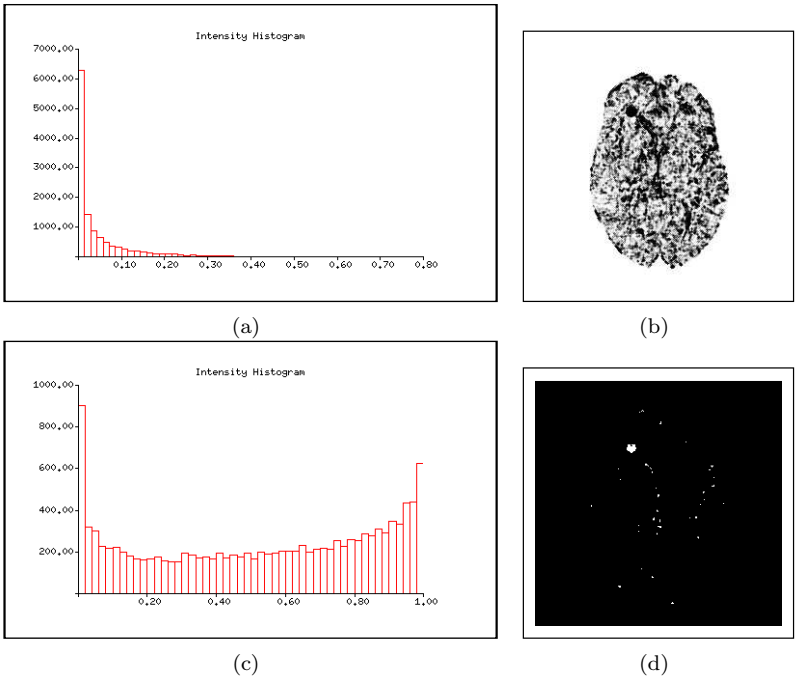



Figure 14: The results of spatial correlation analysis using probability renormalisation on the non-parametric image subtraction result for the synthetic MS data The histogram prior to reflattening (a) shows that the probability distribution of the product is not honest as it does not have the required distribution. The reflattened product (b) produces an approximately uniform histogram (c). The result of thresholding at the 0.5% level is shown in (d).

# 6 Conclusions

We have presented two distinct probabilistic methods for data analysis. The first, based on Bayes theory, has been used for both tissue quantification and categorisation of the spatial distribution of data in magnetic resonance images. We have shown how the issue of objective definition of prior probabilities is a problem in both cases. The use of non-objective priors raises problems concerning direct quantitative interpretation of data, associated with bias and suppression of novelty. We have also explained how Bayesian outputs do not form useful inputs for learning systems, which demand stationarity. Practical solutions to these problems reduce the role of the prior probabilities, and are therefore distanced from the basic theory.

Having identified the main problem as the need to work with multiple models, we have suggested an alternative form of statistical data analysis based upon frequentist methods. These can form useful statistical decisions using the distribution of data from single models. For many tasks these techniques eliminate the problems of unknown priors in a framework which is both flexible and supports data fusion. We have provided a general derivation for the combination of arbitrary quantities of independent data. The issue of self-test is also important, as it facilitates the creation of data analysis systems which are capable of testing the adequacy of their own assumptions. Although we have not demonstrated the application of both frequentist and Bayesian techniques to the same problem it is clear, at least for the atrophic disease classification, that theoretically it should be possible to replace the output probabilities from a Bayesian analysis with probabilities from single model techniques. The task of mapping these values onto a decision process which accounts for Bayes risk is equivalent to mapping the original Bayesian probabilities obtained from fixed priors. In this respect the work we have presented could be considered as possible components of a general approach to analysis of multiple hypothesis in complex data. The stages of this analysis would comprise:

- generation of flat probabilities for individual data sources (by bootstrapped or other methods);
- fusion of data using the probability re-flattening process;
- input of multiple hypothesis probabilities into a Bayes risk analysis system for decision selection.

At each stage in this process the outputs should be honest probability distributions, allowing the model assumptions and data independence to be validated.

It is our opinion that the difficulties in utilising Bayesian results in further analysis procedures should be regarded as a fundamental issue. Any system of modules generally requires that the data passed between them have known statistical characteristics (i.e. errors) [7]. Approaches based upon maximum likelihood have the techniques of covariance estimation and error propagation [14] to support the control of this process. Unfortunately, Bayesian statistics currently has no such tool-kit. Data fusion can only be achieved with knowledge of the assumed prior probabilities, in order to work back to the objective information content (Appendix D). In fact, this information is completely described by the likelihood distribution but not the Bayes probability. Data fusion using the outputs from Bayesian modules is therefore difficult as the data is fundamentally poorly represented for use in a larger system.

The area of computer vision has thankfully largely managed to avoid the Bayesian/frequentist debate which has dogged the statistical literature for many decades. However, given that statistical methodology is becoming the cornerstone of image analysis, it will inevitably become necessary for those in the area to be at least familiar with each side of the argument. Taking all of the factors presented in this paper into account, we conclude that the suggestion that Bayes theory should be the preferred vehicle for the solution of computer vision problems must be treated with scepticism.

## Acknowledgments

The authors would like to thank Professor Alan Jackson for his clinical guidance in the development of the techniques presented. Patrick Courtney was involved with development of the non-parametric subtraction technique in the early stages and also contributed valuable insights regarding the use of statistical modules in larger systems. We would also like to acknowledge the support of: EPSRC and MRC (IRC: From Medical Images and Signals to Clinical Information); DTI Medilink Scheme (Smart Inactivity Monitor using Array Based Detectors (SIMBAD)); Wellcome (Relating Cross-Sectional and Longitudinal changes in Brain Function to Cognitive Function in Normal Old Age); and the European Commission (An Integrated Environment for Rehearsal and Planning of Surgical Interventions).

## References

[1] ALEPH Collaboration, A Precise Measurement of $\Gamma_{Z\to b\bar{b}}/\Gamma_{Z\to hadrons}$. Phys. Lett. B313, pp. 535-548, 1993.

[2] M.M. Bahn, A Single Step Method for Estimation of Local Cerebral Blood Volume from Susceptibility Contrast MRI. MRM 33, pp. 309-317, 1995.

[3] C.M.Bishop, Neural Networks for Pattern Recognition, pp. 66 ff. Clarendon Press, Oxford, 1995.

[4] P.A. Bromiley, N.A. Thacker, and P. Courtney, Non-parametric Image Subtraction using Grey Level Scattergrams. Proc. BMVC 2000, pp. 795-804. BMVA, 2000.

[5] P.A. Bromiley, N.A. Thacker and P. Courtney, Non-parametric Image Subtraction using Grey Level Scattergrams. Image and Computer Vision, BMVC 2000 Special Edition, In press.

[6] P.A. Bromiley, N.A. Thacker, and P. Courtney, Non-parametric Image Subtraction for MRI. Proc. MIUA 2001, pp. 105-108, BMVA, 2001.

[7] P. Courtney and N.A. Thacker, Performance Characterisation in Computer Vision: The Role of Statistics in Testing and Design, Imaging and Vision Systems: Theory, Assessment and Applications, Jacques Blanc-Talon and Dan Popescu (Eds.), NOVA Science Books, 2001.

[8] A.P. Dawid, Probability Forecasting. Encyclopedia of Statistical Science 7, pp 210-218. Wiley, 1986.

[9] G.J. Feldman and R.D. Cousins, A Unified Approach to the Classical Statistical Analysis of Small Signals. Phys. Rev. D57, pp. 3873, 1998.

[10] Friston K J, Holmes A, Poline J-B, Price C J, Frith C D, Detecting Activations in PET and fMRI: Levels of Inference and Power. Neuroimage, 40, pp. 223-235, 1996.

[11] K.J. Friston, P. Jezzard and R. Turner, Analysis of Functional MRI Time-Series Human Brain Mapping vol. 1, pp. 153-171, 1994.

[12] K.Fukenaga, Introduction to Statistical Pattern Recognition, 2ed. Academic Press, San Diego, 1990.

[13] R. Guillemaud and J.M. Brady, Estimating the bias Field of MR Images. IEEE Trans. Medical Imaging, 16(3), pp. 238-251, 1997.

[14] R.M. Haralick, Performance Characterization in Computer Vision, CVGIP-IE 60, pp.245-249, 1994.

[15] A.Jackson, N.W.John, N.A.Thacker, R.T.Ramsden, J.E.Gillespie, E.Gobbetti, G.Zanetti, R.Stone, A.D.Linney, G.H.Alushi, S.S.Franceschini, A.Schwerdtner, A.Emmen. Developing a Virtual Reality Environment for Petrous Bone Surgery: A "State-of-the-Art" Review, Submitted to Journal of Otology and Neurotology, 2001.

[16] H. Jeffreys, Theory of Probability. Oxford Univ. Press, 1939.

[17] D.H. Laidlaw, K.W. Fleischer, A.H. Barr, Partial-volume Bayesian Classification of Material Mixtures in MR Volume Data using Voxel Histograms. IEEE Trans. Med. Imag., vol. 17 no. 1, pp. 74-86, 1998.

[18] J. Neyman, X-Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Phil. Trans. Royal Soc. London, A236, pp. 333-380, 1937.

[19] M. Pokrić, N. A. Thacker, M. L. J. Scott, A. Jackson, "Multi-dimensional Medical Image Segmentation with Partial Voluming", Proc. MIUA 2001, pp. 77-81, 2001.

[20] I. Poole, Optimal Probabilistic Relaxation Labelling. Proc. BMVC 1990, BMVA, 1990.

[21] K.A. Remp, G. Brix, F. Wenz, C.R. Becker, F. Guckel, W.J. Lorenz, Quantification of Regional Cerebral Blood Flow and Volume with Dynamic Susceptibility Contrast-enhanced MR Imaging. Radiology, 193, pp. 637-641, 1994.

[22] N.A. Thacker et.al. Quantification of the Severity and Distribution of Cerebral Atrophy Provides Diagnostic Information in Dementing Diseases. To appear Radiology, 2001.

[23] N.A. Thacker, A.R. Varma, D. Bathgate, J.S. Snowden, D. Neary, A. Jackson, Quantification of the distribution of cerebral atrophy in dementing diseases. Proc. MIUA 2000, pp 61-64, London. 10th-11th July, 2000.

[24] N.A. Thacker, F. Ahearne and P.I. Rockett, The Bhattacharryya Metric as an Absolute Similarity Measure for Frequency Coded Data. Kybernetika, 34(4), pp. 363-368, 1997.

[25] E.A. Vokurka., A. Herwadkar, N.A. Thacker, R.T. Ramsden and A. Jackson, Using Bayesian Tissue Classification to Improve the Accuracy of Vestibular Schwannoma Volume and Growth Measurement. To appear in AJNR, 2001.

[26] L.J. Wolansky, J.A. Bardini, S.D. Cook et al. Triple-Dose Versus Single Dose Gadoteridol in Multiple Sclerosis Patients. Journal of Neuroimaging 4(3), pp. 141-145, 1994.

[27] L. Xu and M.I. Jordan, On Convergence Properties of the E.M. Algorithm for Gaussian Mixtures, A.I. Memo No. 1520 C.B.C.L. Paper no. 111, 1995.

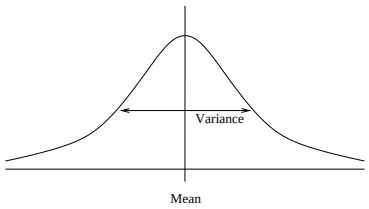



Figure 15: The Normal Distribution

# A Model Parameter Update using Expectation Maximisation

The EM algorithm is implemented in such a way that the **expectation** step recalculates multi-dimensional probability densities, $d_k(\mathbf{g})$, for pure and mixtures of tissues using the current parameters values. Once the probability density functions have been calculated the conditional probabilities, $P(k|\mathbf{g})$, can be derived and used for re-estimation of model parameters in the **maximisation** step of the algorithm in a maximum likelihood (i.e. least squares) manner.

The model parameters for pure tissues $i$ and for mixtures of tissues $i$ and $j$ which are iteratively updated are: the priors $f'_i, f'_{ij}$; the mean vector $\mathbf{M'_i}$; and the covariance matrix $\mathbf{C'_i}$. They take the forms

$$f'_i = \sum_v^V P(i|\mathbf{g}_v) \tag{7}$$

$$f'_{ij} = f'_{ji} = \frac{1}{2}\sum_v^V (P(ij|\mathbf{g}_v) + P(ji|\mathbf{g}_v)) \tag{8}$$

$$\mathbf{M'_i} = \frac{1}{V}\sum_v^V P(i|\mathbf{g}_v)\mathbf{g}_v \tag{9}$$

and

$$\mathbf{C'_i} = \frac{1}{V}\sum_v^V P(i|\mathbf{g}_v)(\mathbf{g}_v - \mathbf{M_i}) \otimes (\mathbf{g}_v - \mathbf{M_i})^T \tag{10}$$

where $g_v$ is the observed intensity value in voxel $v$, and $V$ is the total volume of all data analysed.

Using this representation it is possible to obtain the most probable volumetric measurement $V_i$ for each tissue $i$ given the observed data $\mathbf{g_v}$ in voxel $V$,

$$V_i(\mathbf{g_v}) = P(i|\mathbf{g_v}) + \sum_i P(ij|\mathbf{g_v})$$

# B The Sampled Normal/Gaussian Distribution

The normal distribution (see Fig. 15) is described by the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{(2\sigma^2)}} \tag{11}$$

where $\mu$ is the mean and $\sigma^2$ the variance. Given a sample of data $X_1 ... X_N$, approximate values for $\mu$ and $\sigma$ are given by $\bar{X} = \sum \frac{X_i}{N}$ and $\frac{\sum (X_i - \bar{X})^2}{(N-1)}$ respectively. The variance on the estimator of $\mu$ is $\frac{\sigma^2}{N}$, but the true value of $\sigma^2$ is unknown. Replacing it with the estimator of $\sigma^2$ gives the estimator of the variance of the calculated mean, $\frac{\sum (X_i - \bar{X})^2}{(N-1)\times N}$. The estimator of the area under the graph is the sum of the data points, and since this obeys Poisson statistics its error is $\pm\sqrt{N}$.

The errors on the perfusion parameters can be treated as analogous to the errors on the parameters of the Gaussian distribution. The CBV is equivalent to the area under the graph (N), so has an error proportional to $\pm\sqrt{CBV}$. TTM is equivalent to $\mu$, and MTT to $\sigma$, so the error on the TTM is proportional to $\pm\frac{MTT}{\sqrt{CBV}}$.

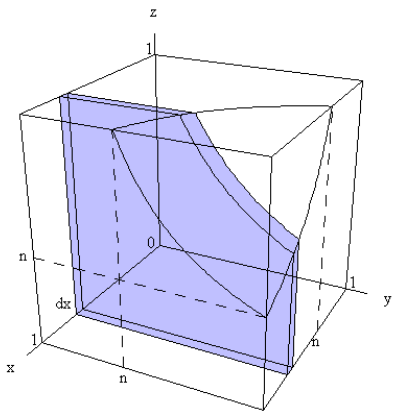



Figure 16: The sample space for the probability renormalisation in 3D, showing the element of integration (the shaded region) used to relate this to the 2D problem. The contour of constant probability is shown by the curved surface in the upper corner of the unit cube.

# C Probability Renormalisation

Given n quantities each having a uniform probability distribution $p_{i=1,n}$, the product $p = \prod_{i=1}^{n} p_i$ can be renormalised to have a uniform probability distribution $F_n(p)$ using

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!} = p + p \sum_{i=1}^{n-1} \frac{(-\ln p)^i}{i!} \tag{12}$$

The quantities $p_i$ can be plotted on the axes of an n dimensional sample space, bounded by the unit hypercube. Since they are uniform, and assuming no spatial correlation, the sample space will be uniformly populated. Therefore, the transformation to $F_n(p)$ such that this quantity has a uniform probability distribution can be achieved using the probability integral transform, replacing any point in the sample space p with the integral of the volume under the contour of constant p passing through this point, which obeys $\prod_{i=1}^{n} p_i = p = constant$. This can be expressed in terms of the volume of a hyper-region of one lower dimension by integrating over one dimension (let this be called x)

$$F_n(p) = p + \int_p^1 F_{n-1}(\frac{p}{x})dx \tag{13}$$

This is equivalent to dividing the integration into two regions using a plane perpendicular to the x axis which intersects the axis at $x = p$. Fig. 16 shows the element of integration that would be used in the 3D case, to relate the volume of the unit cube under the contour of constant probability to the 2D case.

Now, in the simplest case of $n = 1$, clearly $F_n(p) = p$, as no renormalisation is required. The solution for higher dimensions can then be derived by iterative application of Equation 13. This involves integration of terms in $(p/x)[-\ln(p/x)]^n$ which enter in the n=3 and higher cases. This integration can be performed using a simple substitution $x = pu$, $dx = pdu$

$$\int_p^1 (\frac{p}{x})[-\ln(\frac{p}{x})]^n dx = p \int_1^{1/p} (\frac{1}{u})[\ln u]^n du = p[\frac{1}{n+1}[\ln u]^{n+1}]^{n+1} = \frac{p}{n+1}[-\ln p]^{n+1} \tag{14}$$

Iterative application of Equation 13 therefore produces the series

$$F_n(p) = p - p\ln p + p\frac{(\ln p)^2}{2} - p\frac{(\ln p)^3}{6} + p\frac{(\ln p)^4}{24}.... \tag{15}$$

which can be written as

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!}. \tag{16}$$

# D Difficulties with Data Fusion using Bayesian Modules.

Let us imagine that a researcher wishes to build a modular vision system where a module delivers some evidence regarding a particular hypothesis $C$, based upon data from independent sources ($X$ and $Y$) which are then to be combined in a fusion process, such as that described in the paper above. Aside from the above mechanism of combination of hypothesis probabilities we can also combine the likelihoods by taking the product

$$P(X,Y|C) = P(X|C)P(Y|C)$$

However the researcher chooses instead to build modules which delivers MAP estimates ($P(C|X) \propto P(X|C)P(C)$ and $P(C|Y) \propto P(Y|C)P(C)$). This is a common tactic in computer vision in order to justify the use of prior knowledge and so improve the apparent performance of a module. Staying within a MAP framework, the corresponding fused output is $P(C|X,Y)$. However, we cannot compute this quantity using only the outputs from the MAP modules as

$$P(C|X,Y) \propto P(X|C)P(Y|C)P(C) = P(C|X)P(C|Y)/P(C)$$

The process of data fusion therefore requires an understanding of the role that the prior probabilities had in the construction of the output from a module. Unfortunately, if the prior knowledge was implicitly included in a way which did not specify the probabilities then this is going to be difficult to address.

This analysis has assumed that both data sets were interpreted using the same fixed prior $P(C)$. If the priors are inconsistent (as illustrated in the tissue segmentation example in the paper) then we have no right to perform any simple combination. The situation is equivalent to putting contradictory information in a database of facts or using sets of equations which are based upon conflicting assumptions in an analytic analysis. In both cases there is an immediate failure of logical consistency and any attempted inference must be expected to lead to a multitude of invalid and contradictory conclusions. When performing data analysis it is common to invoke the requirement of independence in order to avoid such problems, and we might suggest the use of independent priors. Initially this might look like a possible solution, but the concept of independence follows from characteristics of data. A true prior must, by definition, be the same for any set of data. By introducing a data sample into the specification of our prior knowledge we just transform the prior into another likelihood term, we are therefore fixing the Bayesian approach by abandoning it. This example goes to the very heart of the issue. The standard interpretation of prior probabilities appears to provide a way of incorporating prior knowledge which avoids the constraints which we would otherwise demand for any other source of information. Somewhere along the line, we must expect to have to identify and deal with the consequences.

One of these consequences is that fusion of data from modules which deliver MAP estimates requires us to know precisely what effect any prior knowledge has on the output, and in general we must undo this effect and work back to the likelihood before we can legitimately fuse the data [6]dditional problems of bias introduced by the use of MAP estimators is discussed in other documents on our web pages.. If you think it through, the idea that we can generally throw arbitrary prior probabilities into vision modules in order to improve performance appears has no theoretical credibility in the context of larger systems. However, the mathematical notation used to describe Bayesian systems does not tell us that there should be a problem, and with many publications describing Bayesian algorithms continuing to appear in journals and at conferences this conclusion appears to fly in the face of current received wisdom. This analysis demands that vision researchers develop a new level of critical thinking when formulating approaches for algorithm construction. We should reject the tendency to accept without question the introduction of arbitrary assumptions into algorithms just because someone presents them as priors.

[6] A