Tina Memo No. 2013-007
Presented at the BMVA one day meeting, London 2013.

Quantitative Pattern Recognition: Warts and All.

N.A.Thacker.

Last updated
2 / 20 / 2014

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Quantitative Use of Pattern Recognition: "Warts and All"

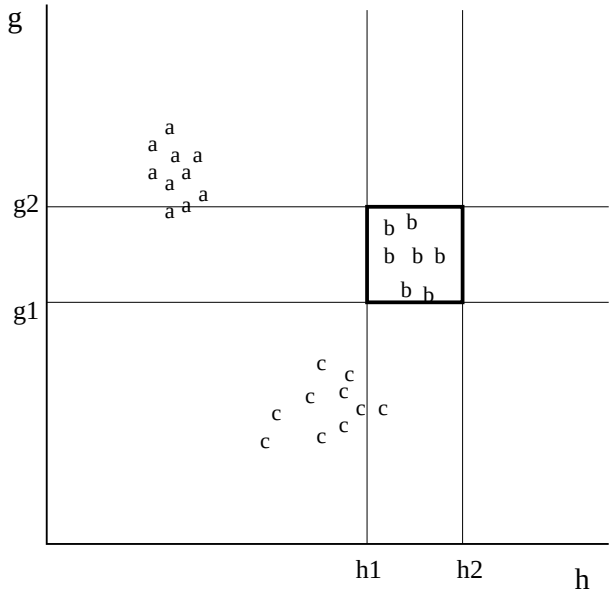
N. A. Thacker, ISBE, University of Manchester

Abstract

Pattern recognition is not like conventional statistical analysis methods. Much of the published work is non-quantitative. The aim is to develop methods which "seem to work", and might be called engineering. Some application domains (science, medicine) do not sit well with this approach. This talk is intended to explain why.

Ideas are illustrated with examples of MRI segmentation and crater counting on Mars.

Hypercube Classifier



A simple approach

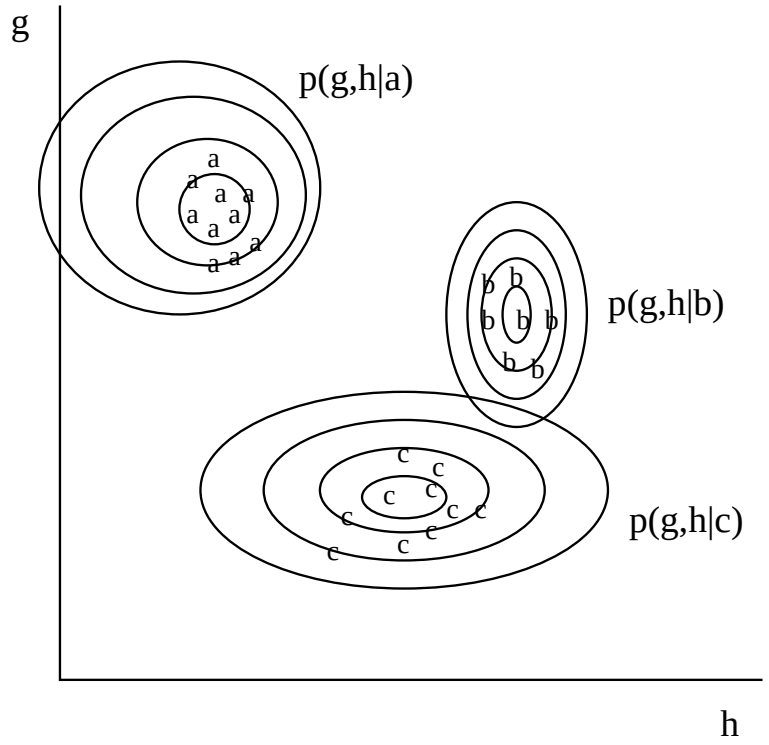
eg: a 'cuts' based analysis , binary image arithmetic;

$(G(i,j) > g_1) * (G(i,j) < g_2) * (H(i,j) > h_1) * (H(i,j) < h_2)$

Simple but does not compactly describe the true distribution of data.

Is there a theoretical optimal approach?

Bayes Classifier



A decision based upon the classification probability eg:

$$P(a|g,h) \;=\; Q(g,h|a)/(Q(g,h|a)+Q(g,h|b)+Q(g,h|c))$$

$$\approx p(g,h|a)Q(a)/(p(g,h|a)Q(a)+p(g,h|b)Q(b)+p(g,h|c)Q(c))$$

where

$$\int p(g,h|\omega)\ dg\ dh\ =\ 1$$

will minimise the classification error.

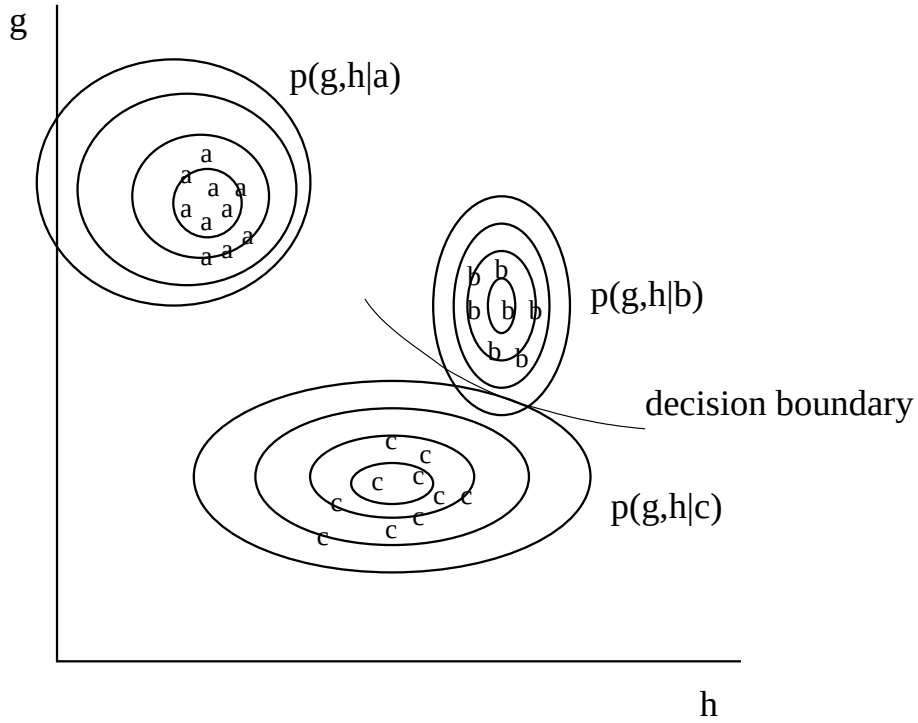
In general

$$P(\omega_m|\mathbf{X})\ =\ \frac{p(\mathbf{X}|\omega_m)Q(\omega_m)}{\sum_n p(\mathbf{X}|\omega_n)Q(\omega_n)}$$

Note: In practice when constructing classifiers from finite samples, difference vectors should be constructed using 'measure theory'.

Decision Boundaries

For a two class problem we can imagine a boundary curve between two probability distributions which passes through all points where $P(\omega_m|\mathbf{x}) = 0.5$



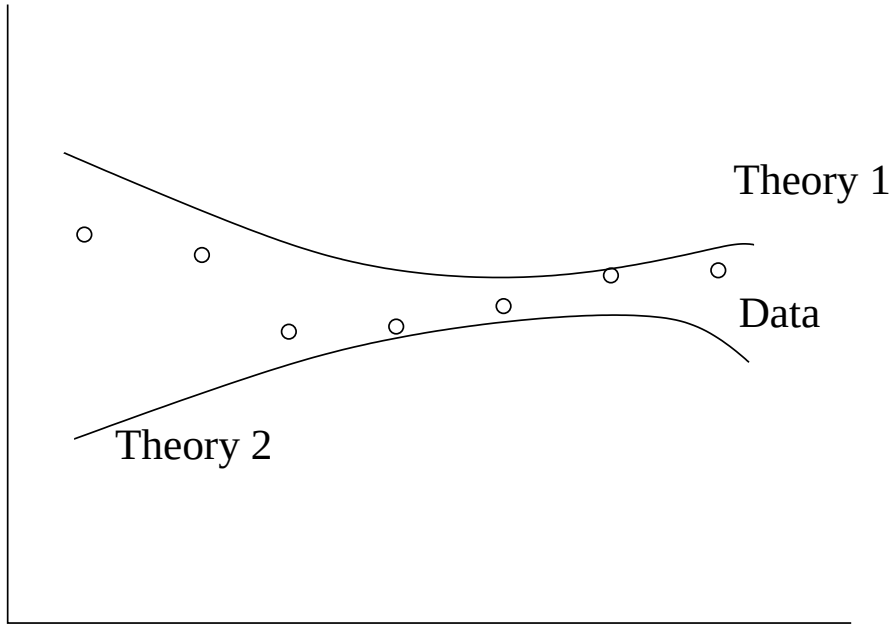
Rapid learning of decision boundaries in high dimensions is the basis for LDA, SVM, Decision Trees, Random Forests, (with or without Boosting).

Evaluation Methodology (Warts I)

Can we use the vast quantities of published results to undertake scientific analyses?

- Papers often focus on irrelevant aspects of algorithms and results can't be reliably replicated (experimenter effects, publication bias [Chatfield]).
- Algorithms are often adjusted using 'test' data which introduces bias.
- Test datasets are often poorly constructed and permit unexpected solutions to the problem. [Shamir].
- Gold standard reference datasets generally have their own biases and error.
- The "representative data" problem, precludes use for quantitative counting tasks.
- Algorithms compared using ROC curves. ROC curves are not really what we need! (See below)

Scientific Method

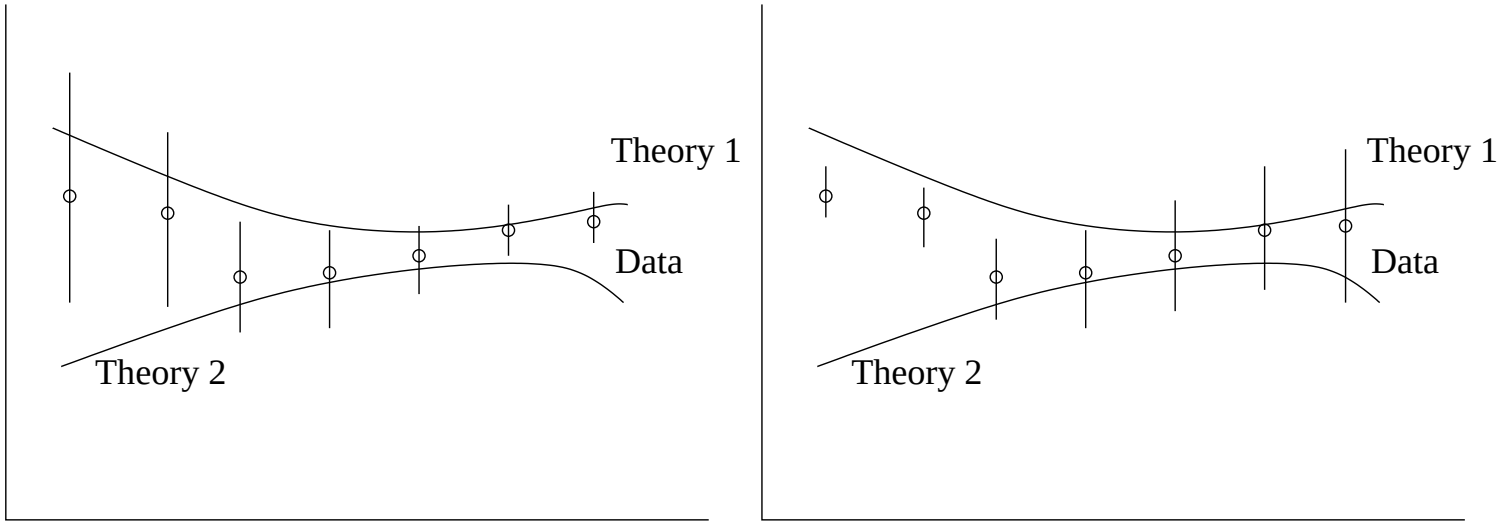


The scientific method involves quantitative comparison of theory and data.

Progress is made most reliably via "falsification" (Popper).

The specific values of the measurement errors are important.

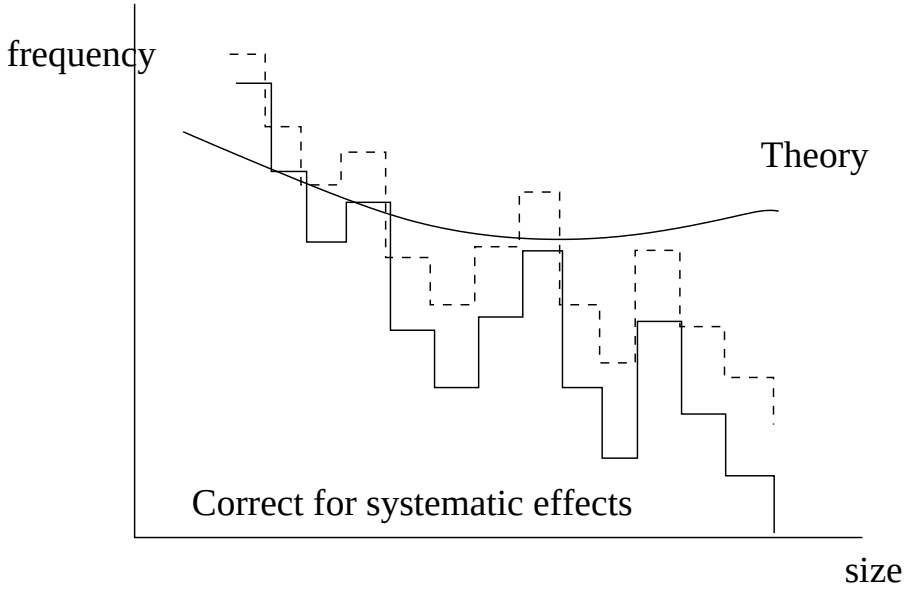
Scientific Method



The specific values of the errors are important.

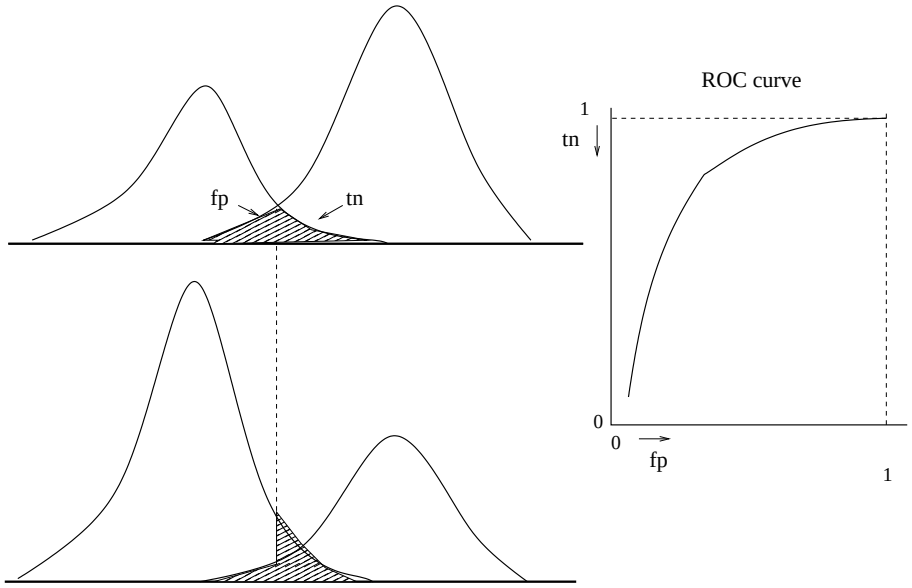
Systematic and Statistical Effects

e.g. size / frequency distribution



We cannot directly compare quantities of classified data to theory, due to mis-identification.

Can we use ROC's to Correct for Systematic Error?



The decision process is only 'Bayes Optimal' if the relative proportions of data classes during training and practical application match [Saerens].

Similarly the misidentification rate will change.

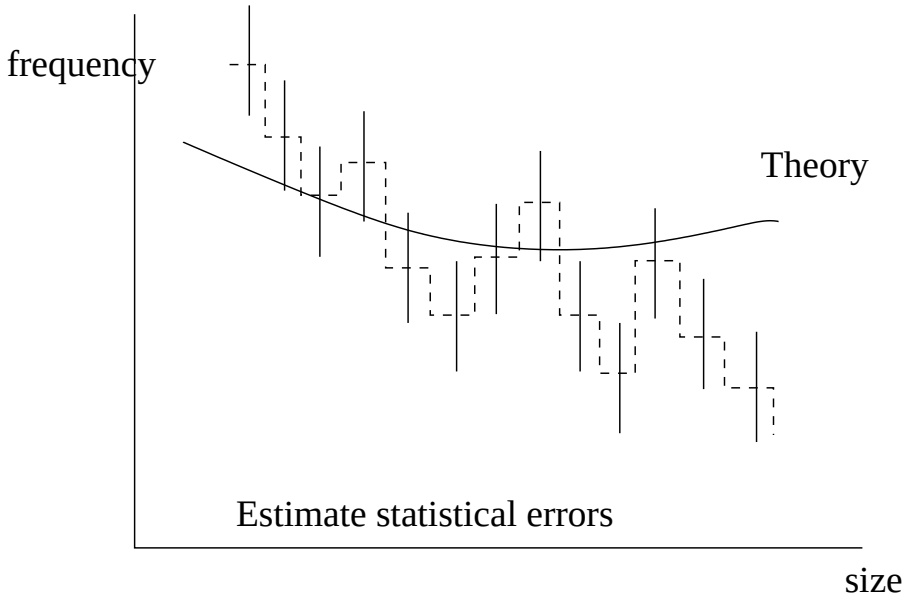
$$Q \;=\; Q' \;-\; Q'\,tn \;+\; (N \;-\; Q')fp$$

Systematic and Statistical Effects

$$Q' = \frac{Q \ - \ N \ fp}{1 \ - \ tn \ - \ fp}$$

eg: $Q = 42$, $N = 100$, $tn = 0.066$, $fp = 0.20 \rightarrow Q' = 30$

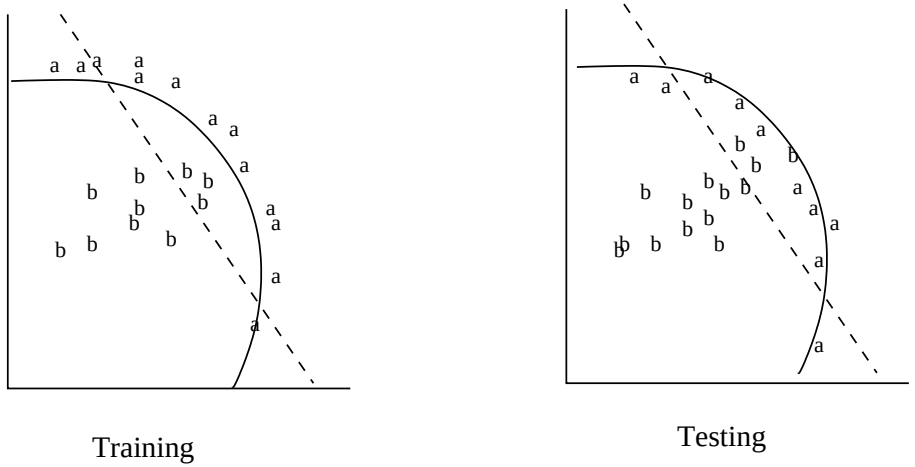
Even having corrected for systematic errors we need to know the statistical error. Is it Poisson/Binomial? (See Below)



In practice the shapes and positions of sample distributions can also change!

Algorithm Construction and Theory (Warts II)

- Demographics change the optimal decision boundaries.
- Probability estimates only correct if the density models are correct [Niculescu-Mizil].
- Data generation processes are poorly understood for arbitrary data.
- Sensitivity to parameter tuning (sometimes crude techniques seem to work better on new data).



Training Testing

It's not just 'where is the boundary?', but also 'how can the boundary move?'.

Summary of the Main Issues

We need to determine the following;

- Which $p(\mathbf{X}|\omega_n)$ do we use for incoming data?
  It needs to be based upon knowledge of typical datasets.
- How do we correct for misclassification?
- What are the errors on observed quantities?

We could use Monte-Carlo, but this requires us to construct a simulation.

Monte-Carlo's require detailed knowledge of the data, including expected sample quantities $Q(\omega_w)$.

If we have a good simulation even a 'cuts based' analysis is practical.

Without a good simulation we are left without any idea of what a pattern recognition algorithm is doing on a new (arbitrary) data set.

Solution : White Box Analysis!

White Box Correction for Misclassification

This best case performance is called the **Bayes Error Rate**, which occurs when $Q(\omega_m)$ is set to match the data sample.

In a sense $Q(\omega_m)$ must be considered an output of the analysis.
The $p(\mathbf{X}|\omega_n)$'s constitute our prior knowledge, not $Q(\omega_n)$!

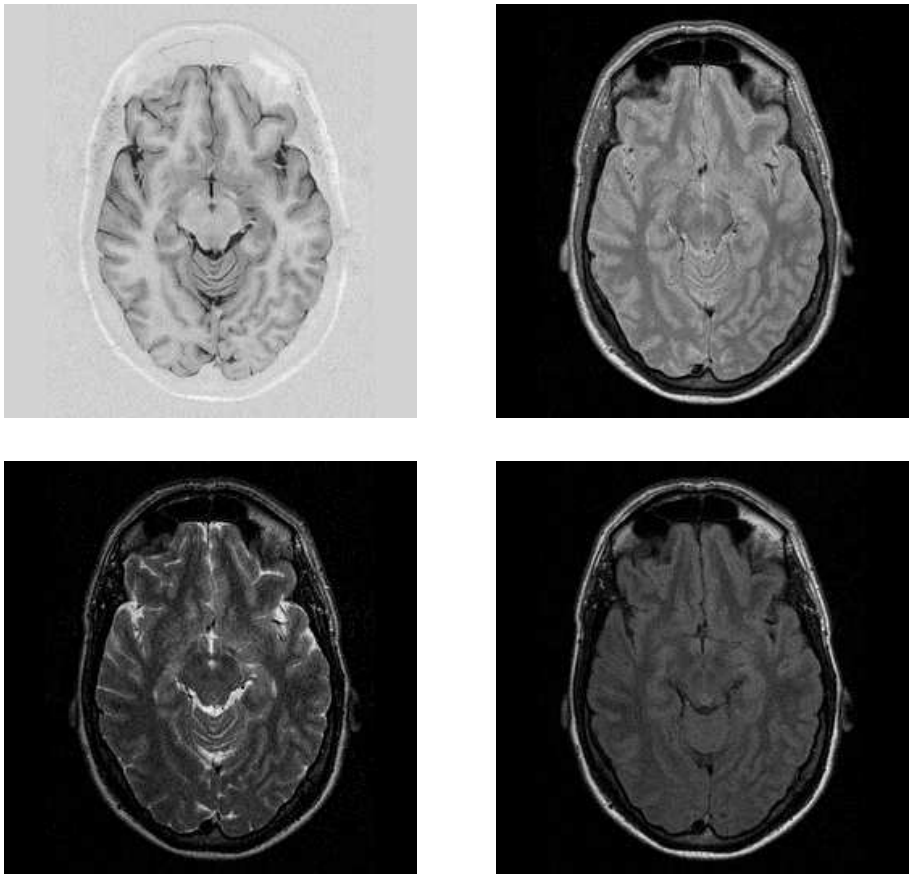
Some algorithms assume that the data sample is fixed.
eg: K Nearest Neighbours, discriminant analysis, Support Vector Machines (SVM).

Some algorithms allow us to estimate $Q(\omega_m)$ eg: Expectation Maximisation.

$$Q(\omega_m) \;=\; \sum_{i \in R} P(\omega_m|\mathbf{X}_i)$$

This is the Likelihood estimate, it is therefore the corrected estimate of quantity!

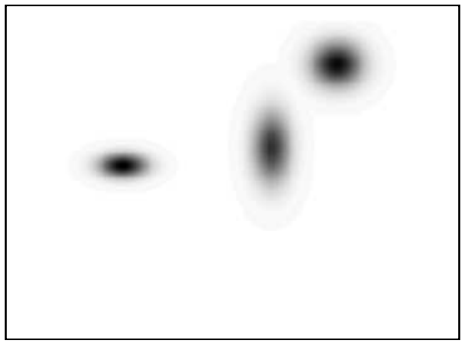
Example: Multi-Spectral Segmentation

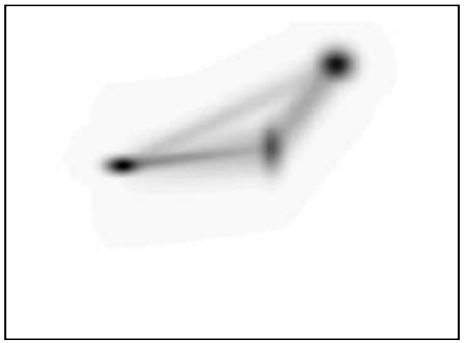


Each grey-level can be used as one component of the measured data vector $\mathbf{X} = (x_1, x_2, x_3, x_4)$.

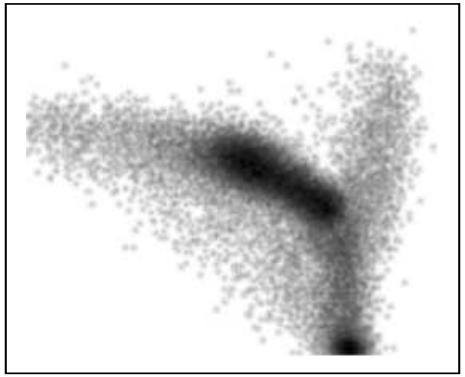
$P(\mathbf{X}|\omega)$ is based upon physics of the MR imaging process.

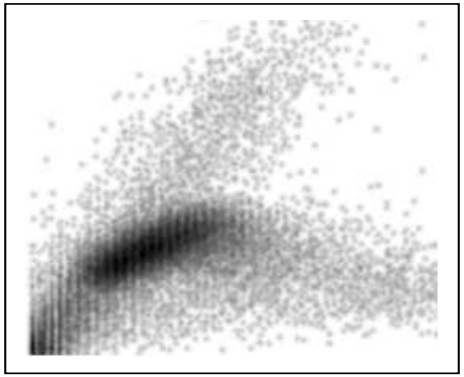
Pure and partial volume models (simulation).





Real Data x_1 *vs* x_2 and x_3 *vs* x_4





Multi-Spectral Segmentation

Parameters for the density model are estimated using **Expectation Optimisation**(EM), which involves iteration of the following steps;

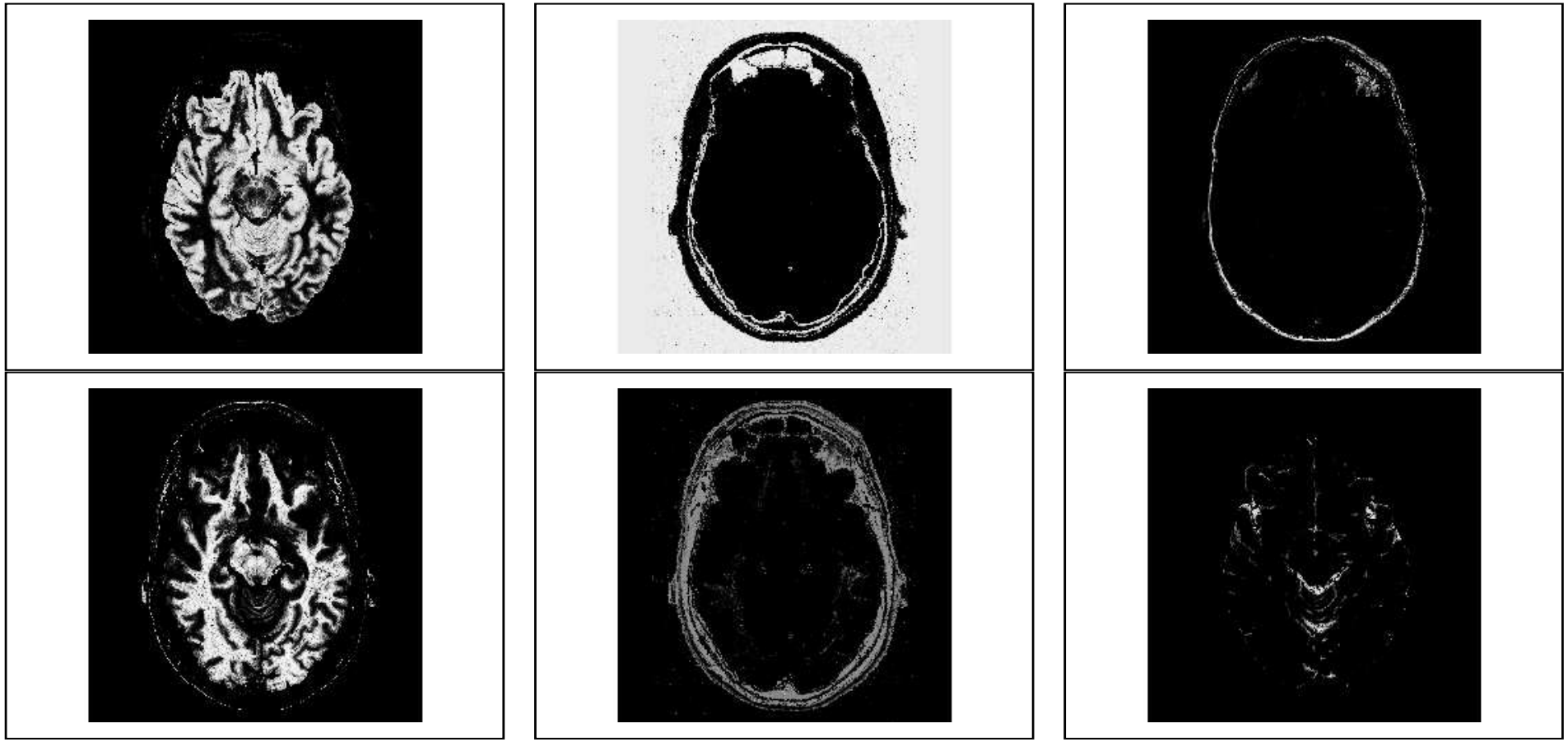
- use an initial estimate of probability classifications $P(\omega_m|\mathbf{X})$ to generate a new estimate of the parameters.
- use the new parameters to re-estimate $P(\omega_m|\mathbf{X})$.

There is a proof that this process will converge on a local optima of the Likelihood of the N data vectors $\mathbf{X}_n$.

$$L \;=\; \Pi_{i\in R} \sum_m Q(\omega_m) p(\mathbf{X}_i|\omega_m)$$

We can assess data conformity (representativeness) using 'goodness of fit'.

Multi-Spectral Segmentation



But; What region (R) should we define? (the one which minimises measurement error?)

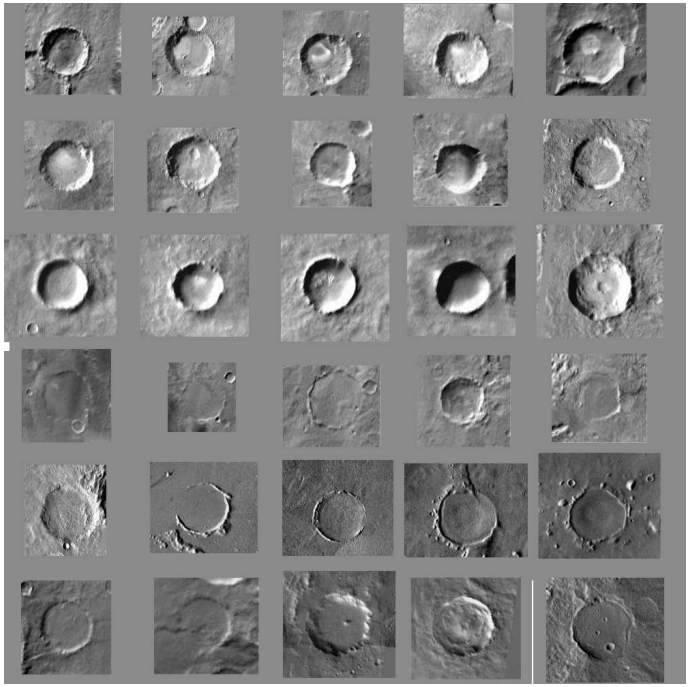
(See: Tina memos 2003-007, 2007-005)

White Box Construction of Density Distributions

What happens when we cannot predict the density ditributions?

- We can model density distributions non-parametrically (e.g. linear models of histograms).
- These models can be constructed from data samples using a version of ICA.
- We seek the model which approximates the variations in distributions seen across multiple samples.
- We can adjust these models to 'best' account for incoming data during analysis.
- These models have estimation errors 'locked in' at the time of training.

Example: Crater Recognition

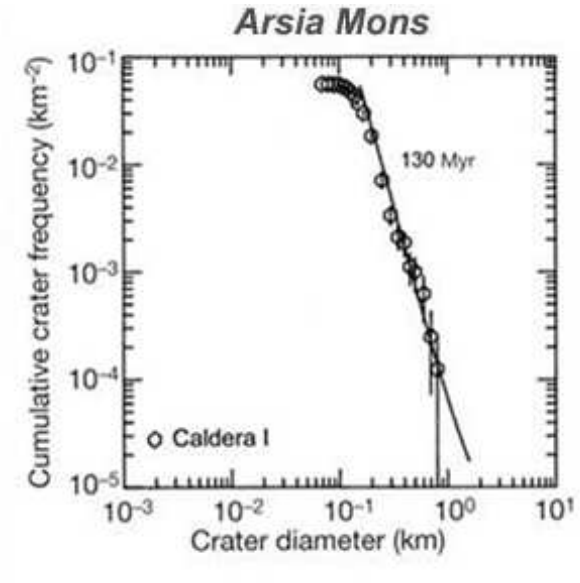


Define a set S of probability densities which describe the appearance of prototypical craters $p(\mathbf{X}|\omega_m)$ with $m \in S$.

This has an associated mixture density

$$p(\mathbf{X}|crater) \;=\; Q(S)\sum_{\omega\in S} p(\omega_m)p(\mathbf{X}|\omega_m)$$

Crater Recognition



(From Neukum *et al.*, 2004, *Nature*, v. 432, p. 972.)

This task now conforms to our earlier definition of scientific measurement.

What are the errors on measured quantities $Q(\omega)$?

White Box estimation of Statistical Error

Start with Likelihood and compute the Cramer-Rao Bound.

$$-\frac{\partial^2\ log(L)}{\partial Q(\omega_m)^2}\ =\ \frac{1}{var(Q(\omega_m))}$$

$$L\ =\ \Pi_{i\in R}\sum_m^M Q(\omega_m)p(\mathbf{X}_i|\omega_m)$$

Assume that $p(\mathbf{X}_i|\omega_m)$ are known (no free parameters).

$$var(Q(\omega)) \geq \frac{Q(\omega)^2}{\sum_i P(\omega|\mathbf{X}_i)^2}$$

As $\sum_i P(\omega|X_i)^2 \leq \sum_i P(\omega|X_i)$, this gives a (best case) Poisson estimate, for **non overlapping** distributions. $var(Q(\omega))\ =\ Q(\omega)$

For a weak classifier we need to consider the off-diagonal terms in the inverse covariance matrix and as $P(\omega|X_i)\ \rightarrow\ 0.5$ then the off-diagonal terms become comparable to the diagonal ones and

$$var(Q(\omega)) \rightarrow \infty$$

This has implications for the evaluation of pattern recognition systems, as the true accuracy of a classication assessment may be many time worse than that predicted by sampling (Poisson or Binomial) statistics.

White Box Correction of Distribution Assumptions

The observed distribution can be modelled as a linear combination of sub-classes. The inverse covariance for a set of classification quantities is then given by

$$C_Q^{-1} \;=\; \sum_{i \in R} \mathbf{D}(\mathbf{X}_i)^T \mathbf{D}(\mathbf{X}_i)$$

with

$$\mathbf{D}(\mathbf{X}_i)^T \;=\; [P(\omega_1|\mathbf{X}_i)/Q(\omega_1), ..P(\omega_M|\mathbf{X}_i)/Q(\omega_M)]$$

We can define a set of subset of classes $m \in S$ as a quantitative measure which relates the theory. s.t.

$$Q(S) \;=\; \sum_{m \in S} Q(\omega_m) \;=\; \mathbf{S}.\mathbf{Q}$$

Then

$$var(Q(S)) \;=\; \mathbf{S}^T C_Q \mathbf{S}$$

(See: Tina memo 2010-008, 2011-003)

Conclusions

Methods which label data using fixed decision boundaries will be sub-optimal.

A reliance on Monte Carlo methods only shifts the problem from design of the pattern recognition system to construction of the simulation.

Automatic construction of a simulation for arbitrary data samples requires solution of the very problem which pattern recognition is supposed to solve.

Many popular pattern recognition techniques and published performance details (e.g. ROC) are unsuitable for direct use in scientific tasks.

For scientific use it is far better to understand the quantitation errors in analysis than to optimise mis-identification rates.

Many of the problems which arise when trying to apply pattern recognition to quantitative tasks are due to 'Black Box' methodologies.

Conclusions: White Box vs Black Box

The White Box approach supports;

- Optimal adjustment of distributions to incoming data.
- Goodness of fit measures to test data conformity (representiveness).
- Estimation of statistical errors for incoming data.
- Estimation of systematic errors arising during training.
- Quantitative tests for self-consistency [Haralick].

Remaining problems with methodology are associated with 'gold standard' definitions and selection of training/test data.

'Black Box' testing is simple and quick resulting in rapid publication. It can be performed by students with little training.

'White Box' analysis requires an understanding of both methods and statistics.

Acknowledgements

MRI segmentation project;
Paul Bromiley.

Satellite images of Mars;
Paul Tar, Jamie Gilmore, Merren Jones.

References

A. Niculescu-Mizil, R. Caruana, Obtaining Calibrated Probabilities from Boosting, Proc. 21st Conf. Uncertainty in Artificial Intelligence, AUAI press, 2005.

K.Chatfield et al. The Devil is in the Details, An Evaluation of Recent Feature Encoding Methods. BMVC Dundee, 2011.

M. Saerens et al. (2001), Adjusting the Outputs of a Classifier to new a priori Probabilities may Significantly Improve Classification Accuracy: Evidence from a Multi-class Problem in Remote Sensing, Proc. 18th ICML, pp. 298-305.

R.M. Haralick, On the Use of Error Propagation for Statistical Validation of Computer Vision Software, (with Xufei Liu and Tapas Kanungo), IEEE Pattern Analysis and Machine Intelligence 27, No. 10 (2005), pp. 1603- 1614.

L Shamir, Evaluation of Face datasets as a tool for Assessing the Performance of Face Recognition Methods. IJCV, 79,3, 225-230, 2008.

All Tina memos available from www.tina-vision.net