Statistical modelling of compositional problems involving finite probability distributions

JOHN AITCHISON

Department of Statistics, University of Glasgow, Scotland;
john.aitchison@btinternet.com.

Abstract

Finite probability distributions and compositional data are mathematically similar, consisting of D-dimensional positive vectors with sum 1. Despite this similarity the meaningful forms of analysis in these different areas may require substantially different concepts and statistical modelling. This paper highlights these differences, but also poses the question of how such differences may contribute to understanding in the different areas. At CoDa workshops we have become so accustomed to, even obsessed with, modelling all compositional data problems within a simplex sample space together with its algebraic-geometric Hilbert space structure. The context of this Hilbert sample space is certainly often relevant to the formulation of a number of compositional data problems, but its mathematical elegance should not override appropriate meaningful statistical modelling to resolve the real compositional problem. In this paper I illustrate some relevant modelling by consideration of how a variety of persons differ in their ability to perform inferential tasks particularly in the process of differential diagnosis.

1 Introduction

My first encounter with the simplex in statistical analysis was not in compositional data analysis as we tend to understand it at CoDa workshops, but in the study of finite probability distributions in clinical medicine as attempts at differential diagnosis between D possible forms of a disease. Typically the clinician starts with a prior probability distribution, his or her view of the incidence rate of patients referred to the clinic. The diagnostic path then consists of selecting from a series of diagnostic tests and updating the probability distribution after observing the result of each test. At each stage of such a process there are several questions to be asked. What is the clinician's degree of uncertainty about the form of the disease? Has the clinician chosen the most informative test at each stage? How far does the updating deviate from the normative Bayesian move, the appropriate compositional perturbation of the current distribution by the likelihood. We examine these questions in an attempt to provide a performance profile for each clinician faced with an individual patient. We describe two studies, one real situation in a non-toxic goitre clinic in Glasgow, the other in a teaching situation comparing different types of students and others in their abilities to make statistical inferences.

All this work started as a consultative problem in 1970 from a medical colleague T.R. Taylor who was investigating how six clinicians in a non-toxic goitre clinic at Glasgow Royal Infirmary arrive at a differential diagnosis of the three types of non-toxic goitre: simple goitre, Hashimoto's disease, thyroid carcinoma. The preliminary modelling and results of the study were reported in Taylor et al. (1971) and further refinements and developments of the modelling and interpretation can be found in Aitchison (1974, 1981); Aitchison and Kay (1973, 1975); Aitchison et al. (2004). This paper will concentrate on the details of the various aspects of assessment of performance leaving all the pictorial aspects to the workshop presentation. Any impatient reader will find pictorial aspects in the works just referenced.

First it is important to realise that a clinician's view of a patient's diagnosis in this non-toxic goitre clinic can be represented by a point in our familiar ternary diagram with vertices 1 (simple goitre), 2 (Hashimoto's disease), 3 (thyroid carcinoma). In the study each clinician faced with a new patient was asked to start the diagnostic process at a position, the clinician's view of the incidence rate of the three types of non-toxic goitre referred to the clinic. The clinician was then asked to select which of the 30 possible tests should be carried out, the result for the patient was provided and then the clinician changed the diagnostic position to a new point within the triangle. Then a further choice of a test and a move to a new diagnostic point and so on until the clinician thought that a reasonable position to

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

make a diagnosis had been reached. Thus for each clinician in the study and for each patient we have a diagnostic path within the triangle 123. Such a diagnostic path results from the clinician subject starting from a perceived incidence rate and after each test changing the set of differential probabilities being placed on each form and heading towards a diagnosis.

The diagnostic path so far envisaged is on a flat triangle but there is an extra important dimension to the problem. It is clear that for any diagnostic point within the triangle there is a degree of uncertainty about the diagnosis and this can be conveniently expressed in terms of the Shannon (1948) measure.

$$H(p) = -\sum_{i=1}^{3} p_i \log p_i \,,$$

a form easily extendible to higher dimensions. Note that for any $p_i = 0$, $p_i \log p_i$ is set to 0. See also Khinchin (1957) for an excellent exposition of such a concept.

The highest degree of uncertainty is clearly at the centre $[1/3, 1/3, 1/3]$ of the triangle and is equal to $\log 3 = 1.093$. Where opinion is divided equally between two forms, say 1 and 2, and so at point $[1/2, 1/2, 0]$, the degree of uncertainty is $\log 2 = 0.693$. For an intermediate point, say $[0.7, 0.2, 0.1]$, the degree of uncertainty is 0.802. If the differential probability vector is at a vertex, say at $[1, 0, 0]$, then we have certainty with a degree of uncertainty 0. Thus the diagnostic path is really within a triangular bowl of uncertainty with depth $H(p)$ below the flat triangle at any point p. This degree of uncertainty plays an important role in our analysis of subjective inference. As far as I know it has played no important role in compositional data analysis, but later in this paper we may see a possible use.

2 Creating a profile of subjective inference

Interesting questions in such investigations are how good are subjects in making reasonable inferences, how good are they at choosing informative tests. Are they liable to underuse the information available or have they a tendency to read too much into the data. If we can answer these questions we should be able to build up, for a subject, a performance profile.

2.1 Degree of uncertainty

At the beginning of each step in a subject's diagnostic process at position p the degree of uncertainty $H(p)$ can be recorded and this gives a view of how the subject is hopefully decreasing the uncertainty and progressing towards a diagnosis.

2.2 Measure of inference discrepancy

At the beginning of each step in a diagnostic path the subject has reached a position $p = [p_1, p_2, p_3]$ within the triangle. Suppose that for a particular chosen test the outcome is x. We can then form the appropriate likelihood function, $l = [l_1, l_2, l_3]$ associated with x for the appropriate Bayesian inference. The mechanism for the inference is exactly the same as the basic group operation of perturbation as we understand it in compositional data analysis. The resultant position reached in the triangle will be r, given by

$$r = l \oplus p = [l_1 p_1, l_2 p_2, l_3 p_3]/(l_1 p_1 + l_2 p_2 + l_3 p_3) \,.$$

The subject's inferential move may have taken him to a point $s = [s_1, s_2, s_3]$ in the triangle. So what we require as a measure of inference discrepancy is a measure of how far the subject's s is from the *target* or *correct* inference r. The appropriate measure $I(r, s)$ here is the well known *directed* divergence measure of Kullback and Leibler (1951), defined as follows:

$$I(r, s) = \sum_{i=1}^{3} r_i \log \frac{r_i}{s_i} \,.$$

This measure has the obvious required properties that

$$I(r,s) > 0, \quad \text{if } r \neq s\,, \qquad I(r,s) = 0, \quad \text{if } r = s\,,$$

and that, roughly speaking, the further s is from r the greater the inference discrepancy.

2.3 Information gain index

We can quantify such notions as 'underusing the information available´, 'reading too much into the data´, 'going contrary to the evidence´ in terms of an information gain index $J(p,r,s)$, where p is the subject's diagnostic position before receiving the result of the chosen test, s is the position moved to instead of the correct position r. Let us define the information gain index as

$$J(p,r,s) = \frac{H(p) - H(s)}{H(p) - H(r)}\,,$$

and consider the various possibilities.

Suppose that $H(p) - H(r) > 0$, so that the Bayesian inference has removed $H(p) - H(r)$ of uncertainty or equivalently gained this amount of information. The subject on the other hand has gained an amount $H(p) - H(s)$ of information in the move from p to s. If $J(p,r,s) > 1$ then the subject has removed more uncertainty than the Bayesian move and so we can say that subject is acting liberally or reading too much into the data. If $0 < J(p,r,s) < 1$ the subject is acting conservatively or underusing the data. If $J(p,r,s) < 0$ then the subject is increasing the uncertainty when it ought to be decreasing and so we can say that the subject is moving contrary to the evidence.

The same kind of argument applies to $J(p,r,s)$ when $H(p) - H(r) < 0$ and in the special circumstances when $H(p) - H(r) = 0$ we set $J(p,r,s) = +\infty, 1, -\infty$, according to whether $H(p) - H(s)$ is positive, zero or negative.

2.4 Test selection discrepancy

An interesting and important concept in statistical analysis, particularly in experimental design, appeared in Lindley (1956) in the definition of the expected gain of information from an experiment. The concept is Bayesian, which fits into the context of our problem here, and we shall define this in terms of the choice of tests in the non-toxic goitre diagnostic process. Let us denote $\theta = [1,2,3]$ as the unknown disease form or parameter. Suppose that a subject has reached an inferential position $p(\theta)$ in the diagnostic process, is thus with a remaining degree of uncertainty $H(p(\theta))$ and that the set of unused tests still available is E. If the subject chooses test $e \in E$ and observes result x then the correct inference would be based on the likelihood function $(l|x) = [l(1|x), l(2|x), l(3|x)]$ taking the subject to a new inferential position, say $p(\theta|x)$ and degree of uncertainty $H(p(\theta|x))$. Since, before the performance of test e the outcome x of the test is not known the only way to assess the informativeness of the test is to compute the expectation of the gain in information $H(p(\theta)) - H(p(\theta|x))$,

$$G = \sum_x \{H(p(\theta)) - H(p(\theta|x))\} p(x) = \sum_x \sum_\Theta p(\theta,x) \log \frac{p(\theta,x)}{p(\theta)p(x)}\,.$$

Thus the optimum procedure would be to compute for each $e \in E$ the expected gain of information $G(e)$ and choose

$$e^* = \arg\max_E G(e)\,.$$

If we write $e_* = \arg\min_E G(e)$, the worst possible choice of test, we can then obtain an obvious measure of test selection discrepancy, namely

$$T(e) = \frac{G(e) - G(e_*)}{G(e^*) - G(e_*)}\,.$$

It is obvious that $0 \leq T(e) \leq 1$ with 0 corresponding to the worst possible choice and 1 to the optimum choice.

All those ideas will be illustrated by examining a number of actual profiles at the workshop.

3 Some comments

It is clear that to make the Bayesian or normative move within the simplex it is necessary to have the likelihood function. This was easy in the non-toxic goitre case since Boyle et al. (1966) had investigated the 30 tests involved and had decided that they could be regarded for all practical purposes as independent. For more complex situations it would be necessary to obtain a reasonable model for the density functions $p(x|\theta)$ associated with each θ so that the likelihood function $l(\theta|x)$ for given observation x would be available.

It will be interesting at the workshop to find if any participants see uses of the less commonly used compositional concepts in this paper. For example, the degree of uncertainty $H(p)$ might have a possible use as a measure of diversity in situations where a single measure of biodiversity is needed. Also the inference discrepancy measure could possibly have a use where the purpose of experimentation is to attempt to move a substance with composition p to a more desirable composition r. Such experiments seem to appear in food engineering. If an experiment led to composition s then a suitable measure of failure would be the Kullback-Leibler directed measure.

References

Aitchison, J. (1974). Hippocratic inference. *IMA Bulletin 10*, 48–53.

Aitchison, J. (1981). Some distribution theory related to the analysis of subjective performance in inferential tasks. In C. Taillie, G. P. Patil, and B. A. Baldessari (Eds.), *Statistical Distributions in Scientific Work*, Volume 5, pp. 363–385. D. Reidel Publishing Co., Dordrecht (NL), 455 p.

Aitchison, J. and J. Kay (1973). A diagnostic competition. *IMA Bulletin 9*, 382–383.

Aitchison, J. and J. Kay (1975). Principles, practice and performance in decision-making in clinical medicine. In K. C. Bowen and D. G. White (Eds.), *Proceedings of the 1973 NATO Conference on The Role and Effectiveness of Decision Theories in Practice*, London (GB). English Universities Press.

Aitchison, J., J. Kay, and I. Lauder (2004). *Statistical Concepts and Applications in Clinical Medicine*. London, Chapman & Hall Ltd. / CRC. 360 p.

Boyle, J. A., W. R. Greig, D. A. Franklin, H. R. M, B. W. W, and E. M. McGirr (1966). Construction of a model for computer-assisted diagnosis: Application to the problem of non-toxic goitre. *Quarterly J. Medicine 35*(4), 565–588.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York, NY (USA). 120 p.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*(1), 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics 27*(4), 986–1005.

Shannon, C. (1948). A mathematical theory of communication. *System Technical Journal 27*, 379–423 and 623–656.

Taylor, T. R., J. Aitchison, and E. M. McGirr (1971). Doctors as decision-makers: a computer-assisted study of diagnosis as a cognitive skill. *British Medical Journal 3*, 35–40.