

©Lucy, D. 2001

cite as Lucy, D. 2001 Method and equations for non-parametric discrimination in dietary analysis - with contributions by Aykroyd, R and Millard, A., unpublished informal report.

Method and equations for non-parametric discrimination in dietary analysis

D. Lucy

Department of Archaeological Sciences, University of Bradford, Bradford, BD7 1DP

Abstract A method is proposed whereby probabilities may be assigned to a nominal variable from knowledge of one or more (preferably lowish dimensions) variables which may be on a continuous or ordinal level of measurement. Furthermore the observable measurement may be known with a level of error, the method taking into account uncertainty in the measurement of both the training set, and the targets. This method may be extended into making estimates for a continuous, or ordinal, variable, and, can be used to estimate a least probability for non-membership of any of the groups represented in the training-set.

Keywords: Keywords: Kernel density estimation, Bayesian estimate, Maximum likelihood estimate.

Introduction

For decades application areas such as provenance analysis from trace elements (Ag) and isotopic composition (Ag, Sn and Cu), dietary analysis from other isotopic compositions (principally $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$), and provenance analysis from Pb, have relied upon plots of data in a relatively low Euclidian space and line of eye to pick out groups, and assign individual cases to groups. Usually rectangular areas have been described in the data-space to enclose groups, if the practitioners are being especially clever then ellipses used. Even when more esoteric means have been employed such as discriminant function analysis (eg: Stos-gale) the techniques employed have been badly misapplied to fewer than four dimensions leading to sometimes acrimonious disputes concerning the reliability of group assignments^{3,4}. Usually there is nothing too much amiss with the above except:

- Reliance on point estimates in the data-space, no account taken of the fact that measurements always have an associated error.
- Unreliable boundaries drawn around and between groups as the rectangles and ellipses represent an unquantized probability.

- No reliable method available to decide which group, or more correctly the split of probability between groups, point measurements belonging within an area of overlap between two or more groups.
- Few means by which to decide whether measurements which don't fall neatly within an area of the data-space belong to any of the groups represented in the training-set. If a discriminant function is employed then it will always deliver probabilities for membership of each group, but cannot produce a probability for 'no group'

The reasons why 'eyeball' type estimation has remained in use for so long are manifold, the strongest being:

- Most groups † are reasonably well distinguished, having but small areas of overlap with their neighbours.
- Except in the case of some applications of trace element analysis few variables are measured ‡, which means that the data-space is amenable to visual examination without having to use some data reduction technique and examination of composite variables. This makes most scientists far more comfortable with, and sure about, the results.

With the flourishing of dietary analysis in archaeological applications in recent years, and the uptake of isotopic information in many areas of applied science it would be a wise move to review the use of numerical procedures in those particular application fields, and put them on a more solid foundation than they currently are.

Simple cases

Let us suppose a highly simplified case where measurements had been made for one ordinal variable x which could be classified into i levels, and a grouping variable G , of which there were h categories. From the cross-table of group and variable one could calculate the Bayesian probability of membership of group $Pr(G_h)$ given that a measurement x_i had been made, from the following:

$$Pr(G_h|x_i) = \frac{Pr(G_h).Pr(x_i|G_h)}{\sum_{h=1} [Pr(G_h).Pr(x_i|G_h)]} \quad (1)$$

which is a perfectly conventional statement of Bayes theorem. For example, were variable x categorised into five levels, and there to be two groups, we could examine the first group ($G_{h=1}$) and record x for nine individuals thus:

†In dietary analysis there are good biological reasons why '*trophic level*' is in discrete non-overlapping units, see reference 5 - Andrew Millard points out that this does not apply to humans as they are omnivours with diets which can, in effect, fall between groups.

‡Again dietary analysis usually has $\delta^{18}\text{O}$ and $\delta^{15}\text{N}$

$x_i = (1, 1, 2, 3, 2)$ where $x_{i=1}$ is the leftmost, and $x_{i=5}$ the rightmost. In the same fashion we could record x_i for $G_{h=2}$ as: $x_i = (2, 3, 2, 1, 1)$. So, were we to observe an x where $x_{i=4}$, then the probability of membership of group one ($Pr(G_{h=1})$) would be:

$$Pr(G_{h=1}|x_{i=4}) = \frac{9/18 \times 3/9}{(9/19 \times 3/9) + (9/18 \times 1/9)} = 0.75$$

where $Pr(G_h)$ is $9/18$ when $h = 1$ as we measured 9 individuals from group one from a total of 18, and $Pr(G_h)$ is $9/18$ for the same reason, but for group two. $Pr(x_{i=4}|G_{h=1})$ is equal to $3/9$ as three out of the nine fell into category four from group one, and $Pr(x_{i=4}|G_{h=2})$ is equal to $1/9$ as only one was recorded as being in category four from a sample of nine from group two. Performing the calculation for group two:

$$Pr(G_{h=2}|x_{i=4}) = \frac{9/18 \times 1/9}{(9/19 \times 3/9) + (9/18 \times 1/9)} = 0.25$$

which, as the sums of the probabilities are one, is entirely expected.

Taking measurement error into account

In the example above we considered a single point estimate as our observation of $x_{i=4}$, in fact what we could have done is expressed our value of x in terms of all levels of x , i.e. $x_{i=1,2,\dots,5}$, which would be a distribution expressed by the vector $x = (0, 0, 0, 1, 0)$. Having done this we could have performed each of the calculations above 10 times to assess the probability of our observation of x belonging in either group for all $x_{i=5}$. Above we had no need to as according to equation 1 all values of zero in x would evaluate to zero, thus being wasted effort. In fact we needn't have calculated the probability of membership of group two as we know that both probabilities must sum to unity, so a single calculation would suffice. What were this not the case, and our observation of x had been half in one level of x and half in another, let us say for example $x_{i=3}$. The vector above would become $x = (0, 0, 1/2, 1/2, 0)$. We could then recalculate the example above, so for $G_{h=1}$:

$$Pr(G_{h=1}|x_{i=3}) = \frac{9/18 \times 2/9}{(9/19 \times 2/9) + (9/18 \times 2/9)} = 0.50$$

and as above for $G_{h=1}$:

$$Pr(G_{h=1}|x_{i=4}) = \frac{9/18 \times 3/9}{(9/19 \times 3/9) + (9/18 \times 1/9)} = 0.75.$$

Though this time each probability[§] is split in proportion to the probability of the observation, that is:

$$Pr(G_{h=1}|x_{i=3}) = 0.50 \times 1/2 = 0.25, \text{ and, } Pr(G_{h=1}|x_{i=4}) = 0.75 \times 1/2 = 0.375$$

If we do the same for $G_{h=2}$ then we get:

[§]I'd coin a phrase like partial probability as it sounds impressive, and is accurate, but has probably already been used.

$$Pr(G_{h=2}|x_{i=3}) = \frac{9/18 \times 2/9}{(9/19 \times 2/9) + (9/18 \times 2/9)} = 0.50$$

$$Pr(G_{h=2}|x_{i=4}) = \frac{9/18 \times 1/9}{(9/19 \times 3/9) + (9/18 \times 1/9)} = 0.25.$$

Performing the same probability split:

$$Pr(G_{h=1}|x_{i=3}) = 0.50 \times 1/2 = 0.25, \text{ and, } Pr(G_{h=1}|x_{i=4}) = 0.25 \times 1/2 = 0.125.$$

Therefore it must be the case that the probability of group membership is the sum of the two individual (partial) probabilities which in this case $Pr(G_{h=1}|x) = 0.25 + 0.375 = 0.625$, and, $Pr(G_{h=2}|x) = 0.25 + 0.125 = 0.375$, the sum being one which is expected. Equation 1 can be rewritten as:

$$Pr(G_h|x) = \sum_{i=1}^{i=i} \frac{Pr(G_h) \cdot Pr(x_i|G_h)}{\sum_{h=1}^{h=h} Pr(G_h) \cdot Pr(x_i|G_h)} Pr(x_i) \quad (2)$$

where:

$Pr(x_i)$ is the probability of x for state i for the observation of x for the unknown group.

A note from Robert Aykroyd (Mathematics and Statistics, Leeds)

you will be pleased to hear that your formulae can be turned into mathematically correct equations which will say what you mean... And your calculation looks correct. Its your notation (and a bit of logic) which is out. (I do have a few minor comments about the rest).

The augument will go something like (forgetting your prob. of not belonging to any of the known groups):

- i) A sample belongs to one of N groups, $g=1, \dots, N$
- ii) Suppose we have a "gold standard" measure, X (say) and that we know (or can estimate) the probability $p(x|g)$
- iii) From an observed x we would then calculate group membership probabilities as usual:

$$p(g|x) = p(x|g)p(g) / \sum_g p(x|g)p(g)$$

- iv) Now suppose that in practice X is not observable, but a noisy version, Y , is measured, with distribtion $p(y|x)$ - this might me additive noise $y=x+e$ - then the group member probabilities are:

$$p(g|y) = p(y, g)/p(y) \text{ with } p(y, g) = \sum_x p(y, g, x)$$

$$= \sum_x p(y|g, x)p(x|g)p(g)$$

$$= \sum_x p(y|x)p(x|g)p(g)$$

is y is conditionally independent of g given x

$$\text{Now } p(y) = \sum_g p(y, g)$$

$$= \sum_g \sum_x p(y|x)p(x|g)p(g)$$

and the overall result:

$$p(g|y) = \sum_x p(y|x)p(x|g)p(g) / \sum_g \sum_x p(y|x)p(x|g)p(g)$$

I hope this helps!!

A more realistic applied example

The variables in isotopic dietary analysis are generally $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$. We need not concern ourselves in this article with how these particular isotopic systems come to participate in non-equilibrium reactions leading to their fractionation. Sufficient to say that they vary according to whether the animal from which the hard tissue was extracted subsisted mostly on animal or vegetable protein, and if animal whether this was of marine or terrestrial origin.

Sample probability density function

Given two continuous variables and a nominal (grouping) variable it would be reasonable to estimate the joint probability distribution function using kernel density techniques, generating for each stage of the nominal a three dimensional structure of x ($\delta^{13}\text{C}$), y ($\delta^{15}\text{N}$), and z being the density for x, y , (equation 3).

$$\hat{f}(x, y; H) = \frac{1}{n} \sum_{i=1}^n K_H(x - x_i, y - y_i) \quad (3)$$

where K is a bivariate kernel function and H a symmetric positive definite 2×2 bandwidth matrix. Here the kernel chosen is the bivariate Gaussian as in the case of isotopic methods this distribution accurately represents the error structure of the actual measurements. Other distributions could be employed as circumstances and the particular application demanded.

$$K_H(x, y) = \frac{1}{2\pi|H|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [x, y] H^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right\}. \quad (4)$$

The general bandwidth matrix has three independent parameters, h_x^2 , h_y^2 and h_{xy} , defining the variances of X and Y , and the correlation between X and Y . This means that the kernel can have elliptical contours with alignment determined by h_{xy} [¶]. The choice of h_x and h_y can be determined in a number of ways. A

[¶]The kernel definitions have been lifted from the AS paper

probability distribution function can be obtained for the sample by using the laboratory estimates for the error on the individual isotopic measurement as the values of smoothing parameter in each continuous dimension. An alternative approach would be to employ some optimised value⁷ were one trying to calculate the value for the probability density function of the population. An appropriate policy might be to use the maximum of the two above as this would allow an empirical estimate of the p.d.f in cases where large samples were available to the investigator, and an estimate to be used where needed. The third kernel parameter h_{xy} could be determined empirically, or suggested from the background theory. In the case of isotopic dietary analysis there is no reason to believe $\delta^{13}\text{C}$ is in anyway related to $\delta^{15}\text{N}$ so in this case h_{xy} could probably be set to zero.^{||}

Observation density function

The p.d.f for the observed value can be generated in the same way as above using equations 3 and 4. Only this time it would be more appropriate to use the stated laboratory error for h_x and h_y because the kernel in the xy plane represents a measurement with uncertainty rather than an estimate for a population.

Group membership estimation

Following the group notation G with h levels used earlier and the notation used in equations 3 and 4, equation 2 giving the estimated probability of group membership given x and y can be rewritten as:

$$\hat{Pr}(G_h|x, y) = \int_0^x \int_0^y \frac{Pr(G_h)\hat{f}_h(x, y; H)}{\sum_{h=0}^h Pr(G_h)\hat{f}_h(x, y; H)} \hat{f}_{xy}(x, y; H) \quad (5)$$

where $\hat{f}_{xy}(x, y; H)$ is the p.d.f of the observed value**.

The null group hypothesis

An important question which can be asked is whether an observation belongs in any of the groups represented in the *training set*. Most discriminant analyses (above being no exception) make the initial assumption that the observation belongs in one in a represented group, the procedure is then to see which of the groups the observation is most likely to have come from. However, in provenance analysis for example, not all possible source sites may have been sampled. The kernel based approach above provides an opportunity to examine the distribution of the observation in relation to that of the *training set*. Anderson² approaches the problem using a test statistic of the integrated squared difference (basically a measure of overlap) between the two distributions. Anderson² concludes that

^{||}Apparently there is an approx 1per.mil shift in $\delta^{13}\text{C}$ per trophic level, and a 3 per.mil shift in $\delta^{15}\text{N}$ (Millard)
 - this will inevitably lead to some correlation.

**Not quite sure whether to return the *training set* to it's pristine state, ie. with each surface proportional to the number of cases in the group. I think strictly one doesn't have to as the likelihood is $\hat{f}_h(x, y|G_h)$ the joint normalised across every surface in this instance.

the asymptotic distribution is Gaussian, but as most empirical distributions are of small size a bootstrapped hypothesis test is more appropriate. **Done nothing on this yet apart from a little on Mary Lewis's data on child long bone lengths.**

Obligatory example

Three bi-variate normal distributions were sampled the parameters of which are given in Table 1. Additionally a single point was selected from a fourth distribution representing an unknown quantity which may have been drawn from one of the represented groups.

Table 1: Distributional parameters for example groups

| Sample | size | mean x | mean y | SD x | SD y |
|---------|------|----------|----------|--------|--------|
| Group 1 | 25 | 25 | 20 | 5 | 4 |
| Group 2 | 30 | 40 | 30 | 6 | 3 |
| Group 3 | 30 | 25 | 45 | 2 | 3 |
| Unknown | 1 | 30 | 30 | 4 | 4 |

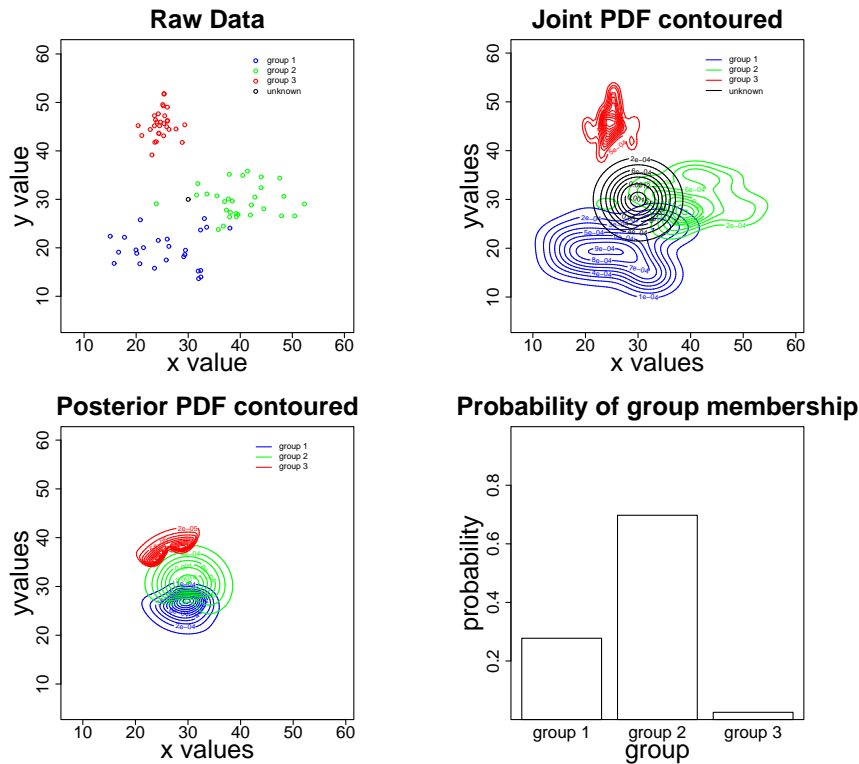
In order to examine which group the unknown measurement is most likely to have come from bi-variate joint kernel density distributions were calculated for each group, and posterior probabilities were calculated from Equation 5. Graphically this process is illustrated in Figure 1.

In this case it has been assumed that the variables x and y are uncorrelated, hence a for parameter kernel (Equation 3). However, it is unlikely that variables will be truly independent, so a five parameter kernel could be employed instead (Equation 4). In the case illustrated it is clear that were the unknown to be derived from one of the represented groups then that group would be Group 2. Additionally it would be of utility to calculate the probability that the measurement came from one of the groups as discussed above.

Discussion

The kernel-based technique described is flexible in that unlike most modelled approaches deviations from normality and mixed variable types can be handled with no special extensions to the method. There is no reason in principle why it's application should be restricted to bi-variate distributions, although at the moment its usage is more suited to low dimensional problems such as dietary analysis. However, it is possible that some technique based on partialling⁶ may provide a short to intermediate term solution to the problem, and may allow some way round the small sample sizes encountered in archaeological studies. A number of drawbacks are to be noticed though. Firstly the technique does not provide a feature rule¹, or at least any simple feature rule for discrimination. Instead output consists solely of a set of type probabilities. Whether this

Figure 1: Contoured distributions for the example. *Top-left*: raw data for each group, *Top-right*: each joint p.d.f. *Bottom-left*: The posterior distributions calculated from the left half of Equation 5. *Bottom-right*: barplot of integrals for the previous plot



is a major problem remains to be seen. Another drawback in a dietary application is that it cannot give answers which are of a 45% marine, 55% C4 variety (unless those proportions have been included as a group), although this property is a feature of all discriminant type analyses.

Conclusions

The method outlined above has the following advantages over more conventional discriminant approaches:

1. non-parametric and unconstrained by data type
2. measurement errors taken into account
3. intuitive for low dimensional data
4. some means of assigning a probability to whether the case comes from the training set
5. good at handling applications where the discriminating variables may be correlated, which frankly, applies most of the time.

Although the following disadvantages are evident:

1. difficulty at the moment with greater than two dimensions - partialling may be a way round this
2. no feature rule established - type probabilities only
3. needs extension into giving proportions of dietary group when used for dietary analysis or some other application where the answers may be on a spectrum rather than absolutely in groups

Andrew Millard (Archaeology - Durham University) adds

One disadvantage is that people using isotopic techniques to look at diet, particularly human diet don't just need to assign group (ie. marine, terrestrial vegetarian etc.) because most humans live on a variety of sources they really need the proportion of each in diet - Andrew Millard came up with this:

How to estimate dietary proportions came to me in a brainwave.

Assume 2 dietary end members G_A and G_B with kdes of pdfs $\hat{f}_A(x, y; H)$ and $\hat{f}_B(x, y; H)$, following your ms. Then the pdf of the a diet with proportions g of diet A and $(1-g)$ of diet B is:

$$\hat{f}_g(x, y; h) = \frac{1}{n_A n_B} \sum_i = 1^{n_A} \sum_i = 1^{n_B} K_H(x - (1-g)x_j - gx_i, y - (1-g)Y_j - gY_i)$$

(Hope I got that TeX right!) Hence it is easy to find $p(g) \forall 0 \leq g \leq 1$ at any point x, y , and to generalise to $p(g)$ integrated over the pdf of an observation. It also generalises to give the joint probability for the proportions of three dietary componenets.

$$\hat{f}_g(x, y; h) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} K_H(x - (1-g)x_j - gx_i, y - (1-g)Y_j - gY_i)$$

References

1. Aitchison J, and Aitkin CGG (1976) Multivariate binary discrimination by the kernel method. *Biometrika* **63**(3):413-420.
2. Anderson NH, (1994) A test statistic for comparing two kernel density estimates.
3. Budd P, Pollard AM, Scaife B, and Thomas RG (1995) The possible fractionation of lead isotopes in ancient metallurgical processes. *Archaeometry* **37**(1):143-150
4. Budd P, Haggerty R, Pollard AM, Scaife B, and Thomas RG (1996) Rethinking the quest for provenance. *Antiquity* **70**:168-174.
5. Colinvaux P (1980) *Why Big Fierce Animals Are Rare*. London: Penguin Books.
6. Lucy D, Aykroyd RG, and Pollard AM (Unpublished) Non-parametric calibration for age estimation. *Applied Statistics*.
7. Silverman BB (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.