Classifying with Confidence using Bayes Rule and Kernel Density Estimation

Tom Fearn[a,*], Dolores Pérez-Marín[b], Ana Garrido-Varo[b],
José Emilio Guerrero-Ginel[b]

[a]Department of Statistical Science, UCL, Gower Street, London WC1E 6BT,UK
[b]Department of Animal Production, E.T.S.I.A.M., Universidad de Córdoba,
14071 Córdoba, Spain

March 27, 2019

Abstract

An example in which near infrared spectroscopic data are used to classify animal feed ingredients is used to make the case for the value of probabilistic approaches to classification problems. The accuracy of probabilities given by linear and quadratic discriminant analysis and by a more flexible kernel density approach are examined, and the effect on these probabilities of the use of different tuning criteria is explored. The example involves the classification of multiple particles in a sample, and detailed probability calculations bearing on the inference for both the sample and its parent population are presented.

Keywords

Classification; Probability; Kernel density estimation; Near infrared spectroscopy

1 Introduction

Although the statistical literature on classification includes extensive discussion of the uncertainty in assignments of unknowns to classes, see [1] for example, it could be argued that this topic is generally given less emphasis in chemometrics, perhaps because many of the algorithms used do not easily provide these uncertainties. An example in which high-dimensional near infrared (NIR) spectroscopic data are used to classify animal feed ingredients is studied here with the aim of demonstrating some methodologies that do provide uncertainties. In particular we will argue for the use of approaches that are based on probabilistic modelling

[*]Corresponding author. E-mail address: t.fearn@ucl.ac.uk

of the within-class distributions of the predictors because these lead very naturally to probabilities as outputs. The example involves both the measurement of multiple particles and considerations of sampling variability, and the probability calculations that enable statements of uncertainty in such a context will be illustrated. The focus is on the one example, but the methodologies employed and the probability calculations illustrated are far more widely applicable.

2 Data

The dataset used to demonstrate the methodology comes from a study designed to develop a method based on NIR-microscopy for the detection, and ideally the quantification, of animal protein by-products in compound feeds [2]. As part of this work, a database was built of spectra from 17570 individual particles of avian (2128 spectra from 41 samples), porcine (2382 spectra from 40 samples), bovine (4657 spectra from 8 samples), ovine (2560 spectra from 4 samples) and fish (5843 spectra from 65 samples) meals, and it is this database that is studied here. It was chosen for two reasons: there is substantial uncertainly in some of the classifications, and the unusually large number of spectra means that it is possible to make reliable assessments of how realistic are the probabilities produced by the methods studied.

The spectra were collected using an auto image microscope connected to a Fourier transform NIR spectrometer (Perkin Elmer, Wathham MA, USA). This instrument measures spectra on individual particles taken from the milled samples of animal meal and spread carefully on a measuring stage. The average number of particles per sample was around 100. The spectral range after truncation of noisy regions was 1500-2280nm, with a wavelength interval of 4nm. The spectra were pretreated by transforming to second derivative using a Savitzy-Golay filter [3] with a 15 point window, then applying a standard normal variate (SNV) scatter correction [4], and finally removing the main water peak between 1824 and 1976nm. This treatment was one of several tried for these data in [2], where it proved the best choice for both K-nearest-neighbours (KNN) and SIMCA. It was the only one employed here. Figure 1 shows the mean spectra for the 5 classes both before and after the truncation and pretreatment. As is common with NIR applications, the baseline and scale variations in the raw spectra are not useful for the classification because the within-class variability in these is much larger than the between-class variability seen in the left-hand plot, which is why the use of pretreatments that minimise these variations is beneficial.

Of interest are both the five-class problem of distinguishing between all five species, and a two-class problem in which the challenge is simply to discriminate between fish and the other four species. The term species is used rather loosely here: the avian and, especially, the fish samples will include multiple species. In Figure 1, the fish spectrum, which is shown as a darker line than the others, does have distinctive features in some regions whilst the other four are much more similar to each other. Not surprisingly the two-class problem, which is important because of the possibility that fish but not the other proteins might at some stage be permitted for use in compound feed for ruminants, is by far the easier one, though if it looks too easy it should be remembered that the plotted spectra are means over thousands of particles.

It could be argued, see for example [5] or [6], that the two-class problem as defined here would be better tackled as a one-class problem, since 'not-fish' is not a well-defined class.

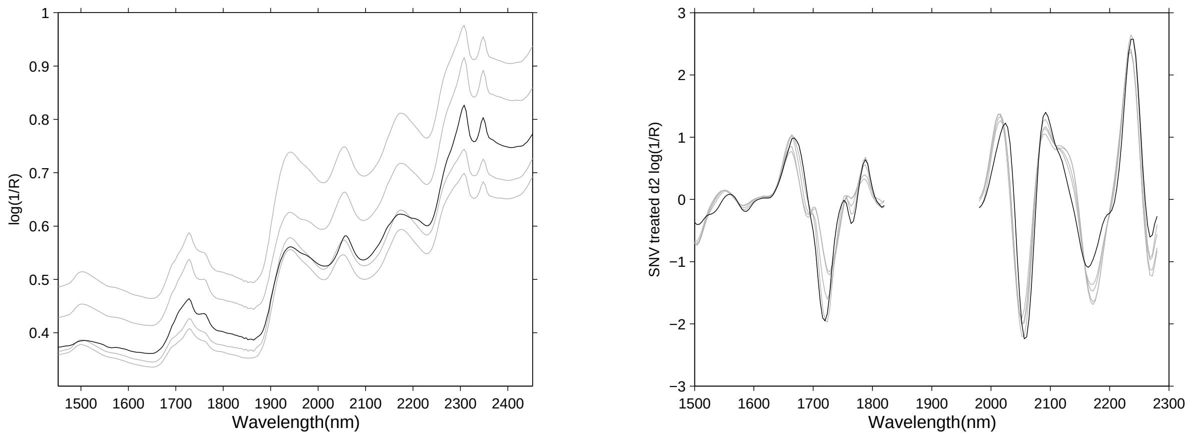

Figure 1: Mean spectra for the 5 classes. Left, raw spectra, right, truncated and pretreated spectra. In each case the darker spectrum is that of the fish

The introduction of a third ‘unknown’ class in Section 4.3 is an attempt to meet at least one of the criticisms of the use of classification methods in applications like this. Another issue is that the pooling of classes may make it less likely that the distributional assumptions underlying some of the methods used will hold, though one of the methods, the kernel density approach, does not make any such assumptions.

The aim in reanalysing these data is not to beat the classification accuracies achieved by KNN and SIMCA in [2], but to investigate whether comparable performance can be achieved with methods that give meaningful probabilities of group membership and to demonstrate the value of these probabilities.

3 Methodology

Approaches to classification can themselves be classified into three types [7, Chapter 1]. The generative approach models the within-class distributions of the predictors and uses Bayes theorem to derive the probability of class membership given predictors as described in Section 3.1. The discriminative approach, the best-known example of which is logistic regression [1, Chapter 8], directly models this probability of class membership given predictors. Approaches of the third type, sometimes called algorithmic, do not model probabilities at all but simply provide classification rules. KNN is an example of this type. There is a good deal of blurring between types: linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) can be presented as examples of both generative and algorithmic approaches, and partial least squares discriminant analysis (PLSDA), which is essentially algorithmic, sometimes has its predictions converted to a 0-1 scale and interpreted as probabilities [8]. Here we adopt the generative approach. We want to work with probabilities throughout rather than add them at the end of the process; and we choose generative rather than discriminative partly because modelling the distribution of spectra given type of sample seems the more natural way to proceed and partly because the explicit appearance of prior probabilities for classes in this setup is preferable to their implicit derivation from the training samples that is a feature of regression methods. See [9] for a more detailed discussion of this last point in the context of quantitative calibration. Oliveri and Downey [10] also argue for the use of the generative approach in food authenticity problems. The methodology used here is Bayesian, see [11] for a good introduction.

3.1 Classification using Bayes rule

The approach taken here is to use the training samples to model the within-class distributions of the spectral data and use Bayes theorem to derive probabilities of class membership given the spectral data for an unknown. With the classes labelled from 1 to m the classification step takes the form

$$P(\mathit{class} = c|\mathit{spectrum}) = \frac{P(\mathit{spectrum}|\mathit{class} = c).P(\mathit{class} = c)}{\sum_{i=1}^{m} P(\mathit{spectrum}|\mathit{class} = i).P(\mathit{class} = i)}, \quad c = 1, \ldots, m. \tag{1}$$

The ingredients here are the prior probabilities $P(\mathit{class} = i)$ that the sample to be classified belongs to class i and the probability $P(\mathit{spectrum}|\mathit{class} = i)$ of observing the spectrum associated with the unknown given that it belongs to class i. The result of the calculation is a set of m probabilities quantifying the uncertainty about the class membership of the unknown given that we know its spectrum. These probabilities can be used to assign the unknown to a class. Here we will simply assign to the class with the highest probability, though if the costs associated with different types of misclassification were unequal it would be preferable, and simple, to combine the probabilities with these costs to find the classification that minimises expected loss [1, Chapter 1].

The prior probabilities are those we would assign to the class membership of the unknown sample before measuring its spectrum. The most obvious default options are uniform probabilities of $1/m$ for each class, or probabilities derived from the proportions of each class in the training set. This second option is the one used in the cross-validations of Sections 4.1 and 4.2 below. In a more general prediction context the distribution of samples amongst classes in a training set does not always reflect that in the population to be predicted, and if the latter can be estimated it would be a better basis for the prior probabilities.

The probabilities $P(\mathit{spectrum}|\mathit{class} = i)$ are derived from multivariate probability distributions fitted to the spectral data in each class. Using continuous distributions these quantities will actually be probability densities evaluated for the observed spectrum rather than probabilities, but the distinction is only a technical one. To make this fitting feasible, the spectral data need to be reduced in dimension. Here principal component analysis (PCA) of the spectra for all classes combined was used for this purpose, with the number of retained components p chosen by a double cross-validation procedure as described below. Other possible approaches include using the scores from PLSDA or using a small number of selected wavelengths.

The most obvious distribution to use is the multivariate normal. If a separate multivariate normal distribution is fitted to the spectral data for each class by plugging in estimates of the means and covariance matrices obtained from the training samples, then the approach described here corresponds to QDA in the sense that assigning unknowns to the class with the highest probability gives the same classifications as QDA, though formulating it as above also gives us probabilities. Similarly, if the means of the fitted multivariate normal distributions are allowed to differ but the covariance matrix is constrained to be the same for each class and estimated by a sample covariance matrix pooled over classes the result is LDA [1, Chapter 3]. If the training set is not large we should account for the uncertainty in the parameter estimates rather than simply plug them in, see [1, Chapter 2] for example, but with the data set used here this added complication is not worth the effort.

Although it is rare for real data to follow a multivariate normal distribution, it is known [1, Chapter 5] that LDA and QDA often perform very well in terms of correct classification rates even when the true distributions are non-normal. What may not work so well when the normality assumption is incorrect are the probabilities generated by Equation 1. When the tails of a distribution are used to calculate probabilities the distributional form becomes critical, and an inappropriate normal distribution can give unrealistic probabilities. For this reason we also explore a more flexible approach to modelling the within-class distributions of the spectral data: kernel density estimation.

3.2 Kernel density estimation

When the sample sizes available for estimation are large, as they are in the example studied here, it is possible to avoid restrictive assumptions about the forms of the within-class distributions and instead model them in a flexible way using kernel density estimation [7, Chapter 2] [1, Chapter 9], often called class modelling using potential functions in the chemometric literature [12, 13, 10]. The idea, which has been used with spectroscopic data in a quantitative context in [9], is very simple. The probability distribution within a class is estimated by placing a spherical multivariate normal distribution, known as a kernel, at the location of each training sample in the class and then averaging these distributions. Thus the density evaluated where the $q \times 1$ vector of spectral spectral data has value $\mathbf{x}$ is

$$P(\mathbf{x}|c) = \frac{1}{n_c} \sum_{i=1}^{n_c} (2\pi\sigma^2)^{-q/2} \exp\left\{ -\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{2\sigma^2} \right\}, \tag{2}$$

where the sum is over all n_c training samples in class c, $\mathbf{x}_i$ is the $q \times 1$ vector of spectral data for the ith training sample in the class and σ is a tuning parameter controlling the spread of the kernels. The use of spherical kernels means that the relative scaling of the q spectral variables is important. The principal component scores used here were scaled to have standard deviation 1.

3.3 Tuning and validation

When the kernel densities are used there are two parameters to be tuned: the number of factors p retained in the dimension reduction and the parameter σ of the kernels. For LDA and QDA just p needs to be tuned. There is no separate validation set, so both tuning and assessment need to be based on the samples in the database. Although there are large numbers of spectra for each class, some classes have only a few samples, and any splitting of the database needs to be done at the level of samples to maintain independence between training and prediction cases. A double cross-validation leaving out individual samples would be too expensive in computing time, so the following scheme was adopted. The database was split into 30 blocks of about 600 spectra each, with all the spectra in a block belonging to the same class. In some cases, and in particular for the 12 blocks containing the bovine and ovine spectra, the blocks comprised a single sample, in other cases they included between 3 and 19 smaller samples. To tune and assess the various methods a double cross-validation was carried out by leaving out each of the 30 blocks in turn, tuning the method on the other 29 blocks using leave-out-one-sample cross-validation to select the best parameter value (in the case of LDA or QDA) or combination (in the case of the kernel density) from a grid, and then predicting the left-out block using the selected parameter values. Both the dimension reduction by PCA and the evaluation of prior probabilities from the training set were done inside the innermost loop of the cross-validation.

Two criteria were used in the tuning process. In one run the parameters were chosen to maximise the number of correctly classified particles in the inner cross-validation. In another they were chosen to optimise a score calculated from the probabilities assigned to the correct classes of the particles in the inner cross-validation. Of the several possible scores [14] we chose to use the log probability score,

$$L = \frac{1}{n}\sum_{i=1}^{n} \log_e(p_i), \tag{3}$$

where the sum is over the n particles being predicted and p_i is the probability assigned by the classification rule to the true class of the ith particle. The best possible score would be 0, achieved by assigning probability 1 to the correct class in each case. To give some idea of the scale of this score, assigning probabilities of $1/m$ to each of m classes, including of course the correct one, would give a score of $\log_e(1/m)$. This is -1.61 when $m = 5$ and -0.69 when $m = 2$. This scoring rule heavily penalises over-confidence in the probabilities, for example assigning a probability of 0.0001 to the true class incurs a penalty of -9.2, so its use in tuning should encourage models that produce realistic probabilities.

4 Results

As explained in Section 2 both the obvious five-class problem and a simpler two-class problem are of interest. We examine them in turn.

4.1 Five classes

The kernel Bayes classifier for the five-class problem was tuned using the double cross-validation scheme described in Section 3.3 and a rectangular grid with ranges of 3 to 7 for the number of factors p retained in the dimension reduction and values of the kernel parameter σ from 0.1 to 0.5 in steps of 0.1. The procedure was run twice, once with total number of correct classifications as the criterion to be optimised, and again using the log probability score for this purpose. The outcomes were different.

When the tuning parameters were chosen to maximise the number of correct classifications, the values $p = 6$ and $\sigma = 0.2$ were chosen for 24 of the 30 outer blocks, with the others either having $p = 5$ or $p = 7$ with $\sigma = 0.2$ or in one case $p = 6$ with $\sigma = 0.1$. The number of classification errors in this double cross-validation was 4865, which is an error rate of 28%. This drops to 4626 or 26% in the double cross-validation if the parameters are fixed at $p = 6$ and $\sigma = 0.2$ for the prediction of each outer block, and falls further to 4247 or 24% in a simple cross-validation with the parameters fixed. Thus the over-optimism of a simple cross-validation that both tunes and assesses is around 4%. None of these error rates is impressive, but they are similar to those obtained with KNN.

When the tuning parameters were chosen to maximise the log probability score, the values $p = 6$ and $\sigma = 0.4$ were chosen for 20 of the 30 outer blocks, with the other 10 having $p = 5$ and $\sigma = 0.3$. The number of classification errors in the double cross-validation was 5506, which is an error rate of 31%. This is, as might be expected, a little higher than the 28% that can be achieved by tuning to minimise the error rate. In contrast, the log probability score is -0.76 here compared with -1.02 above.

The result of classifying any sample using this method is a set of 5 probabilities, one for each class. To investigate how well calibrated these probabilities are, the selected models were run using simple leave-out-one-sample cross-validation, the 17570×5 probabilities produced by each model were grouped into bins of width 0.01 and the proportion of 'correct' classes in each bin plotted against the bin centre. For example, with $p = 6$ and $\sigma = 0.4$, 546 classes were assigned probabilities lying between 0.17 and 0.18. Of these 546, 116 were actually the true class of the particle concerned. At 21.2% this is a little too high - ideally we want 17.5% to be correct - but not by much. Figure 2 shows these plots for the two models tuned using different criteria. It is clear from these plots that tuning using the log probability score results in a model that gives more accurate probabilities.

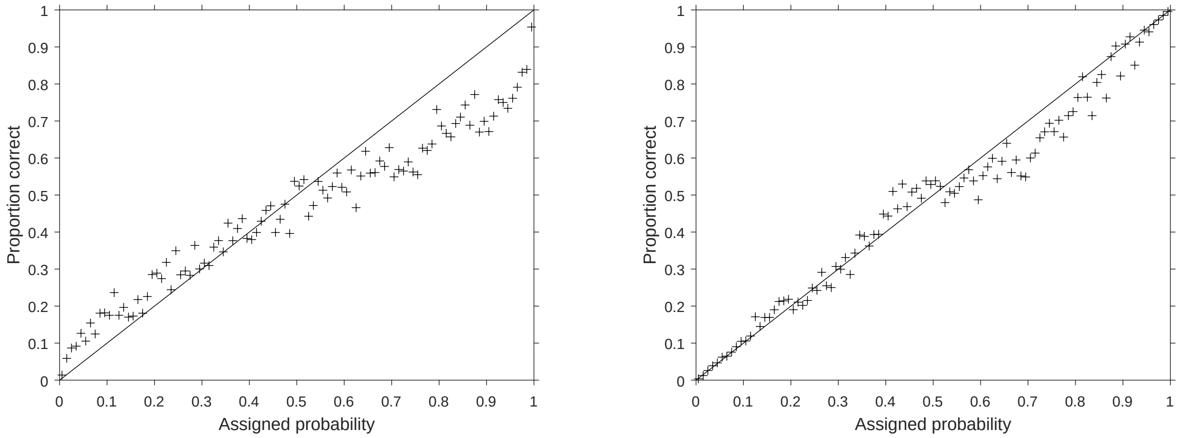

Figure 2: Calibration of probabilities for the five-class case, kernel densities. Left, $p = 6, \sigma = 0.2$, right $p = 6, \sigma = 0.4$

For comparison both QDA and LDA were tuned using the double cross-validation scheme, trying from 2 to 25 factors, p. When QDA was tuned using number of correct classifications as the criterion $p = 24$ factors were selected for most (23/30) blocks and the total number of classification errors was 4461 (25%). When the log probability score was used, $p = 4$ was selected for all but two blocks and the number of errors was 5934 (34%). The log probability scores for the two cases were -1.46 and -1.08 respectively. Thus QDA, if trained to maximise correct classifications, beats the kernel density method on this criterion. However the probabilities it produces are much less accurate, as can be seen from Figure 3. LDA gave both a larger number of errors and less accurate probabilities than the kernel density method and the results will not be described in detail.

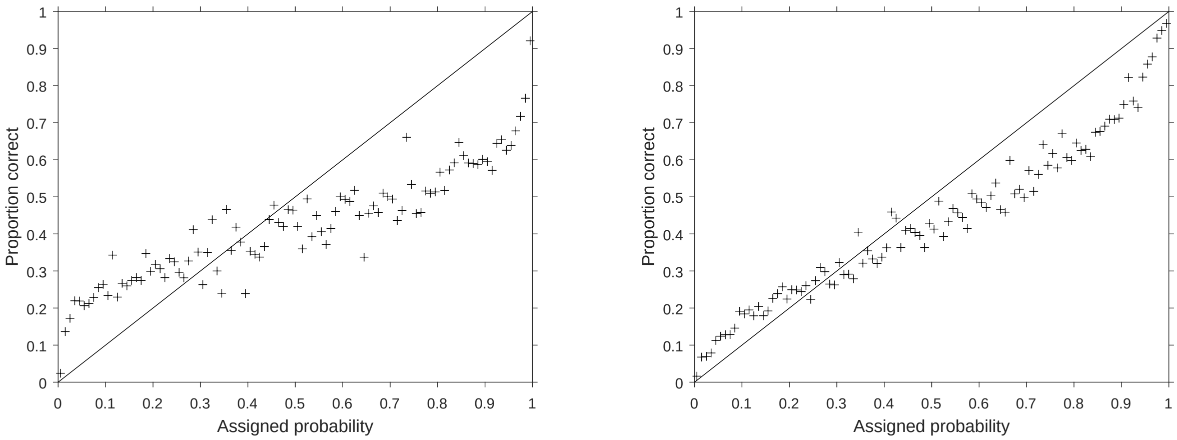

Figure 3: Calibration of probabilities for the five-class case, QDA. Left, $p = 24$, right $p = 4$

4.2 Two classes

Tuning the kernel density method for the two-class problem using the double cross-validation scheme with log probability score as the criterion resulted in the selection of $p = 5$ and $\sigma = 0.3$ for 29 of the 30 blocks, 410 (2.3%) classification errors and an overall probability score of -0.071. Tuning using numbers of correct classifications as the criterion also selected $p = 5$ and $\sigma = 0.3$ for most blocks but, interestingly, resulted in more errors (515) as well as a worse score (-0.093). There is no contradiction here. Using this criterion there were three blocks for which $p = 5, \sigma = 0.3$ was not selected and for one of them the selected combination $p = 6, \sigma = 0.3$ made quite a few more errors predicting the block than did the $p = 5, \sigma = 0.3$ selected by the log probability score. Fixing the parameters at the optimal choice and running a simple leave-out-one-sample cross-validation reduced the number of errors to 366 (2.1%) and produced the probabilities plotted in Figure 4. This 2.1% error rate is slightly worse than the 1.5% error rate achieved in [2] using KNN and the same cross-validation scheme, but this may a price worth paying to achieve probabilities as good as those in the left panel of Figure 4. With only two classes the assigned probabilities occur in pairs adding to 1, so this plot has a symmetry around 0.5. Also, because the problem is easier, there are fewer probabilities in the middle of the range, and so the bins between 0.1 and 0.9 have been widened to 0.1 to get more stable estimates of success rates.

When LDA and QDA were tuned in the same way the best outcome was for LDA tuned with the log probability score. This selected $p = 11$ for 24 of the 30 blocks, made 504 (2.9%) classification errors and obtained a log probability score of -0.099. The right hand panel of Figure 4 shows the probabilities for LDA with $p = 11$.

4.3 Artificial mixtures

To illustrate how the probabilities might be used in practice in the context of the detection of contaminating particles, we constructed two series of artificial mixtures by removing, in each case, one fish sample and one bovine sample from the database and randomly sampling spectra from these two samples to generate samples of 200 particles containing 0, 2, 4, 10 and 20 bovine particles, corresponding to contamination levels of 0, 1, 2, 5 and 10%. The same

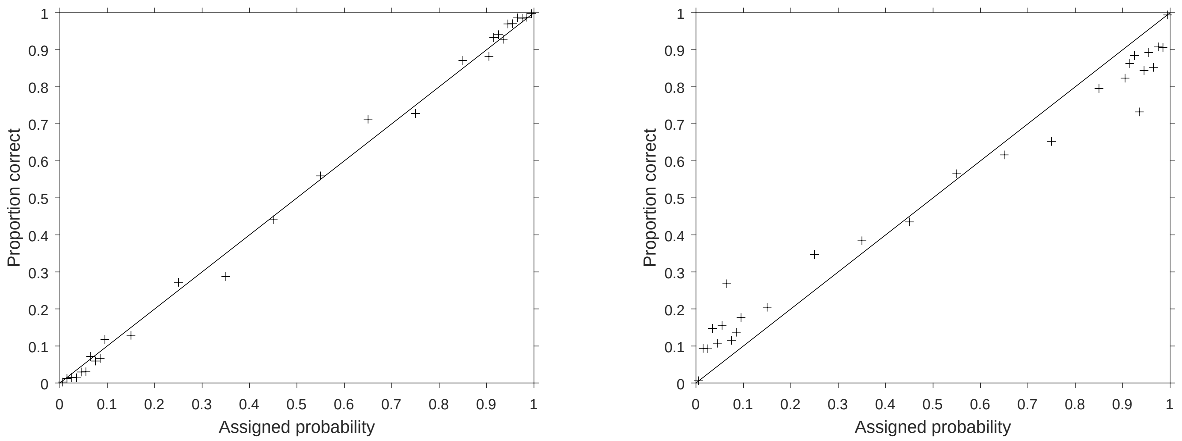

Figure 4: Calibration of probabilities for the two-class case. Left, kernel densities, $p = 5, \sigma = 0.3$, right, LDA $p = 11$

fish sample was used for each series; the bovine samples were different, and were chosen on the basis of the numbers of misclassifications recorded for those samples in the cross-validations described above. The bovine sample for the first series was easy to separate from fish, the bovine sample for the second was more difficult. These artificial contaminated samples were then compared with the rest of the database and the particles in them classified using the two-group model with $p = 5$ and $\sigma = 0.3$ and prior probabilities of 0.5 for fish and 0.5 for non-fish. It needs to be stressed that this exercise is not intended as a demonstration that the spectroscopic methodology is able to detect contamination at the 1% level but simply to illustrate the value of having probabilities. Mixing the measured spectra is by no means the same as measuring spectra on mixtures, in which particles may stick together and both sampling variability and possible sampling bias complicate matters, and indeed the performance on real mixtures was found to be disappointing in [2].

If the model were to be used on genuinely unknown samples, it would need to be extended to allow a spectrum to be classified as not like anything in the database. In this setup this can be achieved by adding an extra class whose within-class probability distribution is uniform over the whole region. Experimenting with the database shows that in this case if the probability associated with this class is set at 0.5×10^{-6} approximately 1% of the database is classified as “unknown”, and this was the value chosen. A similar approach could be used to adapt the methodology to the single-group or non-targeted case, i.e. is this sample authentic or is it something else, with the something else unspecified.

Assigning the spectra to the class with highest probability gave the results in Table 1. In series 1 the assignments were all correct except that one fish particle, the same one in each sample, was classified as unknown. In series 2, around one fifth of the bovine particles present were misclassified as fish, and the outlying fish particle appears in 3 of the 5 samples. If the outlier class is removed the offending fish particle is classified as fish, not as non-fish, and to simplify matters we omit this class in the calculations in the rest of this section.

Rather than simply classifying individual particles, it is more informative to calculate and examine probability distributions relating to the whole sample. To do this it is necessary to specify a prior distribution that reflects the dependency between the particles, in the sense that they all belong to the same sample. The obvious assumption in most contexts would

Sample	True composition		Assignments, series 1			Assignments, series 2		
	Non-fish	Fish	Unknown	Non-fish	Fish	Unknown	Non-fish	Fish
1	0	200	1	0	199	1	0	199
2	2	198	1	2	197	0	1	199
3	4	196	1	4	195	1	4	195
4	10	190	1	10	189	0	8	192
5	20	180	1	20	179	1	16	183

Table 1: Classification of particles in two series of artificial mixtures

be that the $n = 200$ particles in the sample are a random sample from a large population of particles that has a proportion θ of contamination by non-fish particles. Of course this is not true for the samples analysed here, which are artificially constructed, but the idea is to examine what the inference would be if they were random samples. To allow learning about θ from the data we need not to fix it but to give it a prior distribution. Here we use a uniform prior $p(\theta) = 1$ over the range 0 to 1. Then, conditional on θ, each particle in the sample has probability θ of being non-fish, independently of the other particles. If we average over θ using the uniform prior $p(\theta)$ to get the marginal prior probability for one particle to be non-fish it is easily shown to be 0.5, so this setup is consistent with the probabilities used above. Marginally, i.e. after integrating over θ, the particles are no longer independent, reflecting the fact that what we learn about θ from one particle affects our belief about the others. This is why it is not appropriate to simply multiply together probabilities obtained from the classification of individual particles in a sample.

One probability distribution of interest is that for the number of non-fish particles R in the sample. Using Bayes theorem this is given by

$$p(R = r|X) = \frac{p(X|R = r)p(R = r)}{\sum_{i=0}^{n} p(X|R = i)p(R = i)} \ , \tag{4}$$

where X denotes the spectral data for all n particles in the sample and $p(R = r)$ on the right hand side is the probability before observing X that the sample will contain r non-fish particles. With random sampling plus a uniform prior on the population proportion this latter probability turns out to be $1/(n + 1)$ for each r from 0 to n and so cancels in Equation 4. The likelihood terms $P(X|R = r)$ are given by

$$p(X|R = r) = \frac{1}{\binom{n}{r}} \sum_{Sr} \left\{ \prod_{i \in Sr} p(\mathbf{x}_i|NF) \prod_{i \notin Sr} p(\mathbf{x}_i|F) \right\}, \tag{5}$$

where $\binom{n}{r}$ is the number of ways of choosing r of the n particles to be non-fish, and the sum is over all $\binom{n}{r}$ subsets Sr containing exactly r elements of the set $\{1, 2, \ldots, n\}$. For each term in the sum we multiply the probabilities of the observed spectral data for each of the particles, evaluating the probability for particle i using either $p(\mathbf{x}_i|NF)$ from the kernel density model for the non-fish particles or $p(\mathbf{x}_i|F)$ from the model for the fish particles depending on the subset chosen to be non-fish.

With $n = 200$ the number of subsets of size r rapidly becomes very large as r increases, so that evaluating Equation 5 by brute force is very slow for r bigger than 4, while in principle one needs probabilities for all r to evaluate the denominator in Equation 4. The distributions shown for sample 3 of each series in Figure 5 were computed using two shortcuts. First, only the probabilities for r up to 8 were computed, since they become small enough to neglect beyond this. The second short cut was to order the $p(\mathbf{x}_i|NF)$ and only compute the terms in the sum that correspond to the 30 largest of these. This is enough to give an accurate approximation in this particular case, as judged by varying the 30.

The bars in Figure 5 show the resulting probabilities. For each of these samples 4 particles reached the probability threshold of 0.5 to be classified as non-fish, so the results in Table 1 are the same for the two samples. However, we can see from Figure 5 that the probability distributions tell quite different stories. For series 1, the message is that it is very likely that there are either 3 or 4, and most probably 4, non-fish particles in the sample. For series 2 there is more uncertainty, with three possibilities receiving substantial probabilities, and the most likely number of particles is deemed to be 1, not 4. For both samples we can be very confident there is at least one non-fish particle present, the probability of zero non-fish particles is approximately 10^{-4} in both cases.

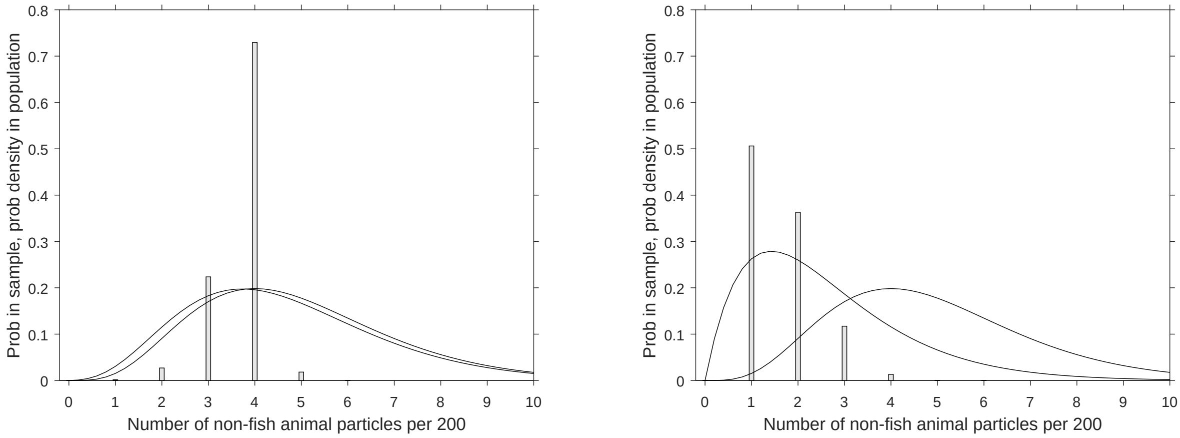

Figure 5: Probability distributions of numbers of non-fish animal particles for mixture 3. Left, series 1, right, series 2. The bars are the probability distributions for numbers of non-fish particles in the samples, the curves are probability densities for the numbers per 200 in the population from which the sample is hypothetically drawn.

These probabilities relate to the sample, rather than the population it came from, and so the uncertainty described by them does not include sampling variability. It is arguably more informative, and actually computationally easier, to compute a probability distribution for the proportion θ of non-fish particles in the population given the observed spectral data on the sample of 200. Again using Bayes theorem, this is given by

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_0^1 p(X|\theta)p(\theta)d\theta} \ . \tag{6}$$

Since θ is a continuous parameter Equation 6 defines a probability density and the normalising sum in the denominator of Equations 1 and 4 becomes a normalising integral here. One could in principle evaluate the integral analytically, but it is easier to discretise θ using a grid of values θ_i from 0 to 1 with steps of $\delta\theta = 0.001$ and evaluate the sum $\sum_{i=1}^{1000} p(X|\theta_i)p(\theta_i)\delta\theta$

instead. Keeping the uniform prior $p(\theta) = 1$, we need simply to evaluate $p(X|\theta)$ for each value of θ in the grid and then divide these values by the sum. The likelihood $p(X|\theta)$ is given by

$$p(X|\theta) = \prod_{i=1}^{n} \left\{ p(\mathbf{x}_i|NF)\theta + p(\mathbf{x}_i|F)(1-\theta) \right\}. \tag{7}$$

Here the probability of observing the spectral data $\mathbf{x}_i$ for the ith particle is the probability θ that the particle is non-fish, multiplied by the probability of observing $\mathbf{x}_i$ for a non-fish particle, plus the probability $1-\theta$ that the particle is fish multiplied by the corresponding probability for $\mathbf{x}_i$. Compared with the expression in Equation 5 this is much more easily computed.

The resulting probability densities for θ for sample 3 of series 1 and 2 are plotted in Figure 5. To enable direct comparison with the distributions for the numbers of non-fish particles in the samples, $p(\theta|X)$ has been rescaled to be a probability density for $\phi = 200\theta$, the number of non-fish particles per 200 in the population. In each case the left hand curve, slightly to the left for series 1, distinctly to the left for series 2, is the relevant one. The right hand curve is the probability distribution we would get if we regarded the results in the sample as 4 definite non-fish particles and 196 definite fish particles, thus ignoring the uncertainty in the classification. This is available as a standard result [11, Chapter 2] or derivable by putting $p(\mathbf{x}_i|NF) = 1$ and $p(\mathbf{x}_i|F) = 0$ for 4 particles, and vice-versa for the other 196, in Equation 7. The result is $p(X|\theta) = \theta^4(1-\theta)^{196}$, which normalises to be a beta distribution with parameters 5 and 197. Comparing the two curves allows us to see the effect of taking into account the uncertainty in the classification of the particles. Comparing the left hand curve with the discrete probability distribution shows how classification plus sampling uncertainty compares with just classification uncertainty. Once again, the picture is quite different for series 1, in the left panel of Figure 5, and series 2 in the right panel. For series 1 we can see that most of the uncertainty arises from the sampling variability, with the two distributions for ϕ having much wider spread than that for the number of particles in the sample. Taking account of the uncertainty in the classification of the particles shifts the the distribution for ϕ slightly to the left, but not by much, and hardly broadens it. For series 2 the effect of taking account of the uncertainty in the classification is substantial; the distribution for ϕ moves well to the left, aligning with the discrete distribution though with much greater spread than the latter.

5 Discussion and Conclusions

The aim of this work was to demonstrate the value of a probabilistic approach to classification, as exemplified by the comparison between Table 1 and Figure 5, where the latter is much more informative. The context of the example analysed, food and feed authenticity, is an area where probabilities are particularly important, but they are always useful. The Bayesian framework adopted here is the most natural way to work with probabilities, and in particular it provides a straighforward way of combining sampling uncertainty with classification uncertainty, as shown in Section 4.3.

The example chosen for this demonstration is unusually large, having over 17,000 spectra. There is little doubt that this is one of the reasons why the kernel density method is so successful. With datasets comprising, say, a hundred spectra it seems very likely that it would be hard to avoid over-fitting with the kernel density approach, and the use of LDA or QDA would be preferred. Further investigations with a range of data sets of different natures and sizes will be necesssary to establish at what point the kernel density approach becomes a realistic option.

.

Acknowledgement

The animal feed dataset is the result of work carried out within the frame of the European Project SAFEED-PAP (FOOD-CT-2006-036221), belonging to the 6th FP.

References

[1] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[2] D. Pérez-Marín, T. Fearn, J. E. Guerrero, and A. Garrido-Varo. A methodology based on NIR-microscopy for the detection of animal protein by-products. *Talanta*, 80:48–53, 2009.

[3] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[4] R. J. Barnes, M. S. Dhanoa, and S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43:772–777, 1989.

[5] P. Oliveri. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - A tutorial. *Analytica Chimica Acta*, 982:9–19, 2017.

[6] O.Y. Rodionova, A.V. Titova, and A.L. Pomerantsev. Discriminant analysis is an inappropriate method of authentication. *Trends in Analytical Chemistry*, 78:17–22, 2016.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

[8] N. F. Pérez, J. Ferré, and R. Boqué. Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometrics and Intelligent Laboratory Systems*, 95:122–128, 2009.

[9] T. Fearn, D. Pérez-Marín, A. Garrido-Varo, and J. E. Guerrero-Ginel. Inverse, classical, empirical and non-parametric calibrations in a Bayesian framework. *Journal of Near Infrared Spectroscopy*, 18:27–38, 2010.

[10] P. Oliveri and G. Downey. Multivariate class-modelling for the verification of food authenticity. *Trends in Analytical Chemistry*, 35:74–86, 2012.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis.* CRC Press, Boca Raton, FL, third edition, 2013.

[12] D. Coomans and D. L. Massart. Potential methods in pattern recognition part 1. Classification aspects of the supervised method ALLOC. *Analytica Chimica Acta*, 133, 1981.

[13] M. Forina, C. Armanino, R. Leardi, and G. Drava. A class-modelling technique based on potential functions. *Journal of Chemometrics*, 5, 1991.

[14] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.