Tina Memo No. 2010-008
Internal.

Can we use Pattern Recognition for Science?

N.A.Thacker

Last updated
1 / 3 / 2011

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

Can we use Pattern Recognition for Science?

N.A. Thacker 19/6/10

Abstract

The purpose of this document is to lay out ideas for the use of pattern recognition in scientific studies. The arguments presented are based upon, and related to, the use of Monte-Carlo simulation techniques as used in the physical sciences. It is shown that for pattern recognition methods built with 'honest' use of probability, it should be possible to make quantitative use of the output of the system. A general theoretical model of a pattern recognition method based upon Bayes Theory is used, which allows us to identify the differences which would be encountered when using "probability density" and "decision boundary" based systems. The arguments presented are intended to provide constraints for the future development of systems which quantify the content of medical images and satellite images of Mars.

Introduction

Pattern recognition is used as a fundamental component of application driven computer vision systems. The use of these approaches for large scale recognition tasks is now common, with performance characterisation on reference databases. This characterisation is often performed as a *scenario evaluation*, or 'shoot out' and used as the basis for evaluation of new algorithmic techniques in comparison to previous methods. It might therefore be reasonable to assume that the published literature contains a sufficient level of information to allow the use of these techniques in scientific studies. In fact, there is often very little in a 'shoot out' style publication to support this process. This is because scenario evaluation is specific to the composition of the dataset on which it has been performed and does not translate to any others. *Technology evaluation* on the other hand, where attention is paid to the statistical nature of the input/output transform, is specific to the algorithm, and can possibly be used as a starting point for a scientific study. Even here the evaluation must generate data and information which is pertinent to the task. Below we investigate this issue via a series of questions which take us from the use of Monte-Carlo simulation (a proven solution) through to use of technology evaluation, and quantitative use of the probabilities estimated by pattern recognition algorithms.

Analysis

Scientific analysis requires the comparison of data with theory and a meaningful comparison must be quantitative, taking appropriate account of the uncertainties in measurement. This then gives us the ability to reject or corroborate theories, ie: the scientific method. Here we will take two specific examples to illustrate this. Let us say that a theory predicts a particular distribution for data in the world, for example, volumes of biological structures in clinical MR scans or the sizes of craters on the surface of Mars. Our task then is to interpret image data and construct a sample distribution in such a way that we can perform meaningful quantitative tests between the sample and theory.

Analysis of images is often performed using segmentation methods, generally based upon pattern recognition. So here is a question;

Question 1: If we generate sample distributions from data labelled using pattern recognition can we ever understand the uncertainties in the data sufficiently well to answer scientific questions?

Often, the main objective in constructing pattern recognition systems is to drive down the mis-classification rates. Indeed, if we can build a *perfect* system then comparisons between theory and data can be performed immediately using conventional Poisson based sample statistics (Chi-square tests, Fisher exact, Kolmogorov-Smirnov, and other hypothesis tests). However, in practice machine vision systems do not achieve high performance figures. Thus any sample distributions constructed will have systematic effects due to efficiencies and misidentification which will compromise statistical tests to the point that they may be uninformative (or worse misleading).

The obvious way to correct for deficiencies in a pattern recognition system is to construct a simulation of the data. If we can make the simulation sufficiently representative of the real world, then any systematic effects, likely to

arise due to using a particular form of analysis, can be estimated. Once estimated the original data distributions, and so scientific interpretation, can be corrected. So the answer to our question is;

Yes, if we can construct a suitable Monte-Carlo simulation of the data.

However, construction of statistically equivalent simulation of data is difficult. Indeed, the process of constructing realistic simulations seems to be an unrealistic solution for arbitrary scenes. It requires a tuning process which matches any key distribution variables to the real data. So that it is always the real data, in some sense, which becomes the reference for its own interpretation. This is an important observation which suggests a method for avoiding the simulation process altogether.

The process of simulation involves two components, first generating the quantities of specific objects and features, and then a statistical perturbation. The first of these must be made to agree with the data under study, though the second can come from an understanding of general variation. These two aspects of modelling fall into the categories of "scenario" and "technology" evaluations. It seems to me a possibility, that if enough were known regarding a particular patten recognition technique, then it should be possible to estimate the corrections needed for quantitative tasks during the process of data analysis (which has already the required frequency of object components), so avoiding the separate construction and analysis of simulation data. Our next question is therefore;

Question 2: Can we use the probabilities estimated by pattern recognition techniques in order to quantitatively interpret sample distributions?

For example, we wish for the case of Martian craters, to use performance characterisation to construct size distributions and associated quantitative error bars from analysis of real images. To get us started I will now provide a straw man analysis of one way this might be done.

Let us imagine that the pattern recognition process is based upon quantitatively meaningful processes. We can then assume there to be a finite number of data generation mechanisms ($j \in J$) which give rise to characteristic distributions $P(X|j)$ (which normalise to unit volume for convenience) for some multi-dimensional pattern space X. We will define X to be a representation computed from a local image region. We can now say that we can fit a density model to any specific regional sample of data (I) such that the distribution of this data is modelled by

$$Q_I(X_i) \;=\; \sum_{j}^{J} P(X_i|j) Q_I(j) \qquad (1)$$

If we assume that densities $P(X|j)$ and quantities $Q(j)$ are both fixed we can use fixed boundaries estimated from sample data (eg. SVM, Kernal PCA, Boosting, Random Forests). However, this implies immediate performance limitations if we know we wish to apply the system to a variety of data with varying composition.

Alternatively, we can assume that densities $P(X|j)$ are fixed but quantities $Q(j)$ change (eg. independent component analysis). If the $P(X|j)$ can be measured from data samples a-priori then we can adjust the $Q_I(j)$ using an iterative procedure such that

$$Q'_I(j) \;=\; \sum_{i}^{I} P(j|X_i) \qquad (2)$$

where

$$P(j|X_i) \;=\; P(X_i|j) Q_I(j) / \sum_{k}^{K} P(X_i|k) Q_I(k) \qquad (3)$$

This can now be recognised as Bayes Theorem but is still restricted by the assumption of fixed $P(X|j)$ in application to real world tasks.

Finally, we can take a generalised view and assume that both densities and quantities change and need to be estimated from sample data (eg. expectation maximisation). This will require us to identify a class of density functions which constrain allowed solutions to those corresponding to physical possibilities. Even though this is selected as a straw man, if you believe (as I do) that all methods in pattern recognition can be associated either with density estimates followed by Bayes theorem, or direct estimation of decision boundaries based upon an implicit choice of models and prior quantities $Q(k)$, then any problems we have using the straw man approach for science will be present for all pattern recognition methods in general.

Now it turns out that equation (2) generates the Likelihood estimate of $Q_I(j)$ needed to match the data density to the assumed model, ie: the value needed to generate the required fractions of each model component in observed data, such as we would need in a Monte-Carlo simulation. This expression can be generalised by introducing s

as some subset partition of the original class membership (eg: crater size). Once the system of estimates have converged we can therefore modify (2) to

$$Q_I(j,s) \;=\; \sum_i^I P(j|X_i,s)$$

although if the probability densities used are uncorrelated (do not change due to invariance properties of the X) with the quantity we wish to measure ($P(j|X_i,s) \;=\; P(j|X_i)$) then any distribution we care to construct can be estimated by weighting the entry of data into the sample histogram by the corresponding $P(j|X)$ from the original pattern recognition system.

This is an important result. What we have demonstrated is that in an 'honest' classification system the estimated probabilities correspond to the technology evaluation of the recognition modules performance. Further, we can estimate the expected error on $Q_I(j)$ by remembering that the Likelihood function we minimised when fitting a density model (via E.M. for example) is;

$$F \;=\; -\sum_i log[Q_I(X_i)]$$

Assuming that the individual density distributions are well determined from training data, then the main cause of error during estimation will come from our ability to establish the normalisation terms $Q_I(j)$. For the case where $P(j|X_i)$ is also a source of error see Appendices B abd C. The second derivative of F w.r.t. $Q_I(j)$ therefore gives the Cramer-Rao bound on the expected error. It is then possible to show that the error on $Q_I(j)$ is approximated by

$$var(Q_I(j)) \geq \frac{Q_I(j)^2}{\sum_i P(j|X_i)^2}$$

As $\sum_i P(j|X_i)^2 \leq \sum_i P(j|X_i)$, this gives a (best case) Poisson estimate, ($var(Q_I(j)) \;=\; Q_I(j)$) for non overlapping distributions. This is the *perfect* classifier mentioned earlier. More generally the inverse covariance for a set of classification quantities is given by

$$C_Q^{-1} \;=\; \sum_i D(X_i)^T \otimes D(X_i)$$

where the vector $D(X_i)^T \;=\; [P(j|X_i)/Q_I(j),..P(J|X_i)/Q_I(J)]$, summarises the information (in a the Fisher sense) contributed to each j for each datum (Appendix A). This covariance can then be used as required subject to estimates of any quantity based upon linear combinations (such as a set of specific density functions which define a more general class), such as $Q_I(S) \;=\; S.Q_I$, using

$$Q_I(S) \;=\; S^T C_Q S$$

Note that in this approach, the priors (generally assumed to be fixed inputs for pattern recognition methods based upon fixed decision boundaries), become output estimates of the system with associated statistical error. The prior knowledge is therefore encoded not in the term we conventionally label as a 'prior', but in the set of allowed density functions $P(X_i|j)$. These functions describe the expected correlations between observed data.

This completes the theory for Likelihood estimates of quantities and their statistical errors when applying pattern recognition for quantitative estimation.

So the answer to our second question is;

Yes, provided that we are using probability quantitatively, so that we can correct estimated distributions and estimate statistical errors.

Returning to our specific examples, the corrected distribution of Martian crater sizes or tissue volumes would be constructed as follows.

1. Identify a set of independent generation mechanisms for local regions of image data, and construct data density models for each $P(X|j)$.
2. For the region of the image where we wish to take measurements iterate equations (2) and (3) until we converge on quantitative estimates of each regional image class $Q_I(j)$.
3. Now fill a histogram H_1 of sizes (structural volume for MR and size for Martian craters) weighting each entry with $P(j|X)$ and another H_2 with weights $P(j|X)^2$.

Following this procedure H_1 is an unbiased estimate of the size distribution corrected for efficiency and mismatch error. While H_2 is the histogram of factors required for estimation of sample variance [1].

[1] We may also wish to take into account other processes such as parameter stability.

Of course, this solution is dependent upon the validity of the original formulation, implicit in equation (1). Although we might wish to believe that meaningful density distributions $P(X|j)$ can be defined for specific objects in an image (such as pure and partial volumes in MR), it seems this is a rather simplistic interpretation. Given the nature of general image formation processes it is clear that many quantities which relate information in images (representations) will be too variable and not satisfy this requirement. In addition, although we can quantify the performance of a pattern recognition algorithm using Monte-Carlo simulation, if (1) is not valid the results obtained from any recognition method will be, in some way flawed. The most general requirement for constructing a quantitative recognition system is therefore exemplified by the assumption that equation (1) is a valid description of representational data density.

We therefore generate our final question;

Question 3: Is it possible to represent image data in such a way that predictable density distributions can be used as a quantitative description of image content?

The problem of applying pattern recognition generically in the scientific interpretation of image content seems to be entirely dependent upon finding an answer to this question. In fact, the issue is related to the concept of "stationarity", which has been well understood in pattern recognition for decades. What appears to be less well understood is any solution. One viewpoint is that these distributions can never be known and we should forget about trying to make these methods quantitative. It should be possible to make some rational comments regarding this issue, either to aid in finding candidate solutions, or else to prove conclusively that it is impossible (see the Discussion below).

If we consider the problem of estimating $P(X|j)$ we can see that complications arise when dealing with mixtures of generating processes. Different processes n and m may not generate equivalent distributions $P(X|j_n) \neq P(X|j_m)$. However, we only need to be able to find the set of $P(X|j_n)$ which can account for all previous examples of data. This might be done using independent component analysis. For one case where we have already applied quantitative segmentation (MR image analysis) we can see how this might work. Independent component analysis of multiple example image regions would have to generate both pure tissue and partial volume tissue distributions in order for equation (1) to accord with the known physics. As we need two distributions (triangles convolved with a Gaussian) for the partial volume terms, and they are expected to correlate, this might be difficult. We are aided here in one important respect. It is legitimate to use any linear combination of densities which defines the required sub-space. Unique identification of the true generator distributions seems unnecessary, provided the data generation process is truly linear.

Our corresponding algorithm for crater counting looks like this;

1. Use multiple image regions to perform an independent component analysis and estimate $P(X|j_n)$ for some N which allows adequate representation of data density.
2. For the region of the image where we wish to estimate a crater size distribution iterate, equations (2) and (3) until we converge on quantitative estimates of each regional image class $Q_I(j) = \sum_n^N Q_I(j_n)$.
3. Now fill one histogram H_1 of measured crater sizes weighting each entry with $P(j|X)$ and another H_2 with weights $P(j|X)^2$.

As before H_1 is now an unbiased estimate of crater size and H_2 is the histogram needed for the corresponding sample variance.

Notice that the first step in this process is simply used to establish a family of density functions which can be used to model the allowable variations in an effective $P_I(X|j) = \sum_n^N P(X|j_n)Q_I(j_n)$. Subcomponents of this model are unlikely to be orthogonal, but strictly all that is needed is that they allow adequate approximation of $P_I(X|j)$ with a small number of degrees of freedom.

The only two alternatives to answering question 3 are to use Monte-Carlo analyses or to make the recognition system work so well that quantitative correction is unnecessary (ie: near perfect performance). Though both of these alternatives seem to appear as important aspects of the computer vision literature, the idea of trying to make pattern recognition methods quantitative seems to be very rare. It should be noted that much recent work in this area (such as SVM, boosting, random forest ands Kernal PCA) focuses on computational tractability rather than quantitation.

Discussion

When trying to characterise the distributions in a pattern recognition system we must accept that it is impossible to obtain perfect results. Distributions based upon finite samples are inevitably approximations. In the light of this and the analysis above some might wish to argue that it is reasonable to conclude that we can never know the correct distributions for a computer vision task. They may further argue that there is no need to attempt to match assumed distributions to those present in data, and even that this will not deliver performance benefits. As this can be described as unscientific they may further state that computer vision is "engineering" rather than "science". We can answer this view with the following argument.

Firstly, we have to define what we mean by using correct distributions. On this issue we need only require that the distributions match the "true" distribution within the limits of our ability to distinguish between them using statistical tests. That is that any systematic bias due to use of the model must be less than that detectable in the presence of statistical sampling variation. In this case any difference between the performance of the system with the approximate distribution and the true one will be (by definition) quantitatively insignificant. This falls into the area of Bayesian sensitivity analysis.

For example, when building tissue classification models for MR images one finds that the distributions of grey level values for pixels corresponding to air (zero signal) are strictly Rician. Although it is a significantly different distribution to a Gaussian, the Gaussian assumption is appropriate if the calculation of $P(j|X)$ for data in the air space is going to be statistically equivalent regardless of which model we choose. Whenever there is only one possible generator of the observed data Bayes theory gives $P(j|I) = 1$. This is the case in MR data for the parts of a Gaussian which match poorly to a Rician. Another way of explaining this is to say that the details of assumed distributions only affect the output when there is an overlap between alternative generator distributions $P(X|j)$. It is in these regions that the relative proportions of the assumed distributions establish the detailed behaviour of classification probabilities (including associated errors on summary statistics) which define the shape and location of decision boundaries. Although we may like to define the concept of having a"correct" distribution for cases where physical argument provide the theory, our concept of what constitutes the "best" approximating distribution is necessarily defined according to the overall proportions of classes in any test sample (Appendix C).

Good engineering is involved with choosing approaches which give the best results. For object classification we have a well defined idea of what this means. Given our pattern classification system we wish to minimise the number of classification errors in a performance test. Even though we may not know the appropriate density parameters with which to construct a Bayes decision system at the beginning of a test, we know them at the end. The best performance will be set by the Bayes error rate. This is only achievable using parameters for the density model which match those from the total sample of data in the experiment. So we do need to match the assumed distributions to those in data and the closer the assumptions match the data the closer we will approach Bayes error performance. This is just good engineering.

There is an associated point here which is interesting. If we split a large dataset in two, and these two sets are selected so that they do not have equivalent distributions, we know that the set of optimal parameters for the overall dataset must be different to those for the two subsets. Given the choice, we would expect to do better by splitting data into dissimilar groups and using the optimal parameters for each, rather than using large groups. There is a caveat that the process we use to generate subsets must be uncorrelated with the distributions used for classification. The limit of this process will be set by the ability to accurately determine the required parameters. This gives us a way of introducing regional segmentation algorithms into analysis, prior to application of the pattern recognition process. Trivially, optimal performance would be obtained if we could split the input data into the appropriate classes before classification[2].

Though the required sets of parameters (or distribution assumptions) will be required to change, according to the specific dataset we wish to apply it to, for any given set of data the best we can expect is to be self consistent in our assumptions. In these circumstances the probabilities used in the decision system will be quantitatively honest, and we can attempt to use them for scientific studies as described above. For the case of MR image segmentation this corresponds to fitting the data density distributions to the data we are analysing, and setting any prior tissue parameters to those which match the quantities of these tissues in the image. The assertion that we can **never** know the appropriate distributions is contradicted by this example. Of course this does not prove that this will be the case for any formulation for image analysis. For example, we can see that the use of spatial priors for tissue segmentation is meaningless in the context of a single data set. Spatial priors optimise the classification decisions across an entire group of images for the case where this group has the same spatial distribution. However, it does not optimise the tissue classification for any single image. There is insufficient data in a single image to allow us to

[2]More realistically, we might expect that any segmentation processes which might exist in biological systems, can be quantitatively refined in order to optimise subsequent pattern recognition based decision making.

confirm that this distribution is appropriate. Indeed, any measurements made from a single image will illustrate a bias toward the mean distribution, which for medical applications are likely to impede our ability to characterise abnormality.

For Martian crater analysis we have a more complex problem. We have no arguments, based upon a general understanding of image formation, from which to define characteristic functions for $P(X|j)$. We must therefore perform research in order to see if such distributions can be inferred from sample data. Although this process might seem somewhat arbitrary (with regard to selected regions etc.), the only point which matters here is that over the region I the estimated distributions constructed by the specified weighting process is unbiased. The differences between alternative versions of an analysis should only differ within the implied statistical confidences, ie: they should be equivalent statistical summaries of the data. Good and bad approaches can therefore be distinguished on the basis of small variance predictions which match the quantitative reproducibility of the method.

Conclusions

This document has considered how the theory of pattern recognition might be used to allow us to understand the quantitative behaviour of recognition systems. This has been illustrated via consideration of application of these techniques to MRI data and satellite images of Mars. We have explained how the process of probability weighted counting

$$Q(k) \;=\; \sum_d P(k|X_d)$$

(as opposed to simple lablel counting) is consistent with a likelihood estimate, corrected for efficiency and ambiguity losses.

We can further say that for a fixed set of data in which the proportions of underlying generation processes remain fixed, the variance on estimated quantities $Q(k)$ due to statistical fluctuation is proportional to $Q(k)$, and the constant of proportionality can be predicted by the associated probability theory (Appendicies A and B). The systematic error on this quantity due to errors in assumed distribution functions is proportional to $Q(k)$ (Appendix C). This term is more difficult to esimate, requiring tests of the level of agreement between predicted and observed error on representative data sets. The two terms can be summarised as

$$Q(k) \pm \alpha\sqrt{Q(k)} \;\pm \beta Q(k)$$

The aim of good engineering should be to choose ways of representing data density ($p(X|k)$) so that the systematic error term (β) is as small as necessary to ensure the statistical errors dominate for the quantities of data involved in a given scientific study. In this way, scientific conclusions will be dominated by the observed evidence rather than the assumptions inherrent to the pattern recognition process. Techniques which assume fixed prior quantities (by modelling decision boundaries) will not only preclude use of probabilistic weighting but will also make significant irreducible systematic errors and are consequently best avoided.

Appendix A: Quantitative Counting Errors

The inverse variance can be estimated from the Cramer-Rao bound.

$$\frac{1}{var(Q_I(j))} \leq -\sum_{i\in I} \frac{\partial^2 log\ Q_I(X_i)}{\partial Q_I(j)^2}$$

using the following results.

$$\frac{\partial log\ Q_I(X_i)}{\partial Q_I(j)} \;=\; \frac{p(X_i|j)}{\sum_k p(X_i|k)Q_I(k)}$$

$$\frac{\partial^2 log\ Q_I(X_i)}{\partial Q_I(j)\partial Q_I(k)} \;=\; \frac{p(X_i|j)p(X_i|k)}{[\sum_k p(X_i|k)Q_I(k)]^2} \;=\; \frac{p(j|X_i)p(k|X_i)}{Q_I(j)Q_I(k)}$$

So that

$$C_Q^{-1}(j,k) \;=\; \sum_{i\in I} p(j|X_i)p(k|X_i)Q_I(j)Q_I(k) \;=\; \sum_{i\in I} D_j(X_i)D_k(X_i)$$

or equivalently

$$C_Q^{-1} \;=\; \sum_{i\in I} D(X_i)\otimes D(X_i)$$

Appendix B: Density Estimate Errors

We have assumed in the main text that the effect of errors on $P(X|k)$ are small in comparison to those due to $P(j)$. In fact in the worst case we can estimate each $P(X|k)$ separately using a histogram (brute force look up) rather than assuming a low parameter parametric form. In this case we can estimate the contribution to the variance on $P(j)$ as follows. In order to treat the process of correlation between samples correctly, we start by defining the measured quantity of k as a sum over all distinguishable patterns X rather than samples X_d, where

$$Q(k) \;=\; \sum_X P(k|X)Q(X) \;=\; \sum_d P(k|X_d)$$

and $Q(X)$ is the number of patterns in bin X. Defining $Q(\bar{k})$ as the number of all other classes other than k we have

$$var(Q(k)) \;=\; \sum_X \left[(\frac{\partial Q(k)}{\partial P(X|k)})^2 var(P(X|k)) \;+\; (\frac{\partial Q(k)}{\partial P(X|\bar{k})})^2 var(P(X|\bar{k})) \right] Q(X)^2$$

$$= \sum_X \left(\frac{Q(k)}{P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k)} \;-\; \frac{P(X|k)Q^2(k)}{(P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k))^2} \right)^2 var(P(X|k))Q(X)^2$$

$$+ \left(\frac{P(X|k)Q(k)Q(\bar{k})}{(P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k))^2} \right)^2 var(P(X|\bar{k}))Q(X)^2$$

we assume that the density distributions are determined from a set of training data of total quantity $TQ(k)$ (we have trained with T times more data than we now count) then $var(P(X|k)) \;=\; P(X|k)/(TQ(k))$ and $var(P(X|\bar{k})) \;=\; P(X|\bar{k})/(TQ(\bar{k}))$

$$= \sum_X \frac{\left(\left[Q(k)P(X|\bar{k})Q(\bar{k}) \right]^2 P(X|k)/Q(k) \;+\; \left[P(X|k)Q(k)Q(\bar{k}) \right]^2 P(X|\bar{k})/Q(\bar{k}) \right) Q(X)^2}{T(P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k))^4}$$

$$= \sum_X \frac{\left(\left[P(X|\bar{k})Q(\bar{k}) \right]^2 P(X|k)Q(k) \;+\; \left[P(X|k)Q(k) \right]^2 P(X|\bar{k})Q(\bar{k}) \right) Q(X)^2}{T(P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k))^4}$$

$$= \sum_X \frac{P(X|\bar{k})Q(\bar{k})P(X|k)Q(k)Q(X)^2}{T(P(X|\bar{k})Q(\bar{k}) \;+\; P(X|k)Q(k))^3}$$

We can now convert back to a sum over individual data vectors d as

$$var(Q(k)) \;=\; \sum_d \frac{P(X_d|\bar{k})Q(\bar{k})P(X_d|k)Q(k)Q(X_d)}{T(P(X_d|\bar{k})Q(\bar{k}) \;+\; P(X_d|k)Q(k))^3}$$

Now for a model which fits the data $(P(X_d|\bar{k})Q(\bar{k}) \;+\; P(X_d|k)Q(k)) \;=\; Q(X)$ so that

$$var(Q(k)) \;\approx\; \sum_d \frac{P(X_d|\bar{k})Q(\bar{k})P(X_d|k)Q(k)}{T(P(X_d|\bar{k})Q(\bar{k}) \;+\; P(X_d|k)Q(k))^2} \;=\; \frac{1}{T}\sum_d P(\bar{k}|X_d)P(k|X_d)$$

Now as $P(\bar{k}|X_d) \;+\; P(k|X_d) \;=\; 1$ we can write this as

$$var(Q(k)) \;\approx\; \frac{1}{T}\{\sum_d P(k|X_d) \;-\; \sum_d P(k|X_d)^2\} \;=\; \frac{1}{T}\{Q(k) \;-\; \sum_d P(k|X_d)^2\}$$

In conclusion, the variance due to sampling errors on the likelihood density has a maximum value of $Q(k)/T$, reducing to zero in circumstances of no distribution overlap and with increasing training sample quantity. Whereas the contribution to variance expected due to the estimation of $Q(k)$ using likelihood is always at least $Q(k)$ and larger in circumstances of distribution overlap. Thus by normal rules of variance estimation the contribution to variance, even in the worst case situation of non-parametric bin sampling, is expected to be negligible in comparison to the CRB likelihood contribution. For reasonable parametric density models the contribution to variance will be smaller than this provided the overall sample size is not so big as to show up systematic errors in the assumed parametric form (Appendix C).

Appendix C: Probability Modelling Errors

As we have seen, the correction of data for scientific purposes requires accurate calculation of conditional probabilities $P(j|X)$. We can therefore assess the difference between the theoretical correct distribution $P(X|j)$ and an approximation $P'(X|j)$ in terms of the error generated for sample distributions weighted by this quantity. Defining

$$Q(X) \;=\; \sum_k P(X|k)Q(k)$$

The sensitivity of $P(j|X)$ to a change in $P(X|j)$ is given by

$$\frac{\partial P(j|X)}{\partial P(X|j)} \;=\; \frac{Q(j)\sum_{k\neq j} P(X|k)Q(k)}{Q(X)^2} \;=\; Q(j)P(\bar{j}|X)/Q(X)$$

from which we can see that if sum term is exactly zero then there will be no error on $P(j|X)$. This occurs when there is no overlap between j and the other distributions, as expected. Also that the error on a weighting quantity is

$$P(j|X) - P'(j|X) \;=\; [\frac{\partial P(j|X)}{\partial P(X|j)}](P(X|j) - P'(X|j))$$

remembering that the errors generated for a specific value of X are correlated we can compute a quality function which estimates our ability to accurately correct sample data, in terms of the expectation of the systematic error on samples from a specific class j

$$S(Q(j)) \;=\; E < (P(j|X) - P'(j|X)) > \;=\; \sum_X Q(X)P(j|X)[\frac{\partial P(j|X)}{\partial (P(X|j)}](P(X|j) - P'(X|j))$$

so that when summing instead over d

$$S(Q(j)) \;=\; Q(j)Q(\bar{j})\sum_d P(j|X_d)P(\bar{j}|X_d)\left[P(X_d|j) - P'(X_d|j)\right]/Q(X_d)$$

For fixed $P(X_d|j)$ and $P'(X_d|j)$ and $Q(\bar{j})/Q(X_d)$ ratio, we can see that $S(Q(j)) \;\propto\; Q(j)$, i.e. we will get a fixed proportional error on any estimated quantity due to a fixed error on the density functions. This is in contrast to the variance terms computed in Appendices A and B, which predict statistical errors which scale as $\sqrt{Q}(j)$. For quantitative use, the sample size must remain below a value where standard sampling errors dominate. The better the density estimation process the larger this value will be. This suggests a way of testing parametric assumptions, by checking on the level of divergence from predicted variance as a function of sample size.