Computing Likelihood Functions for High-Energy Physics Experiments when Distributions are Defined by Simulators with Nuisance Parameters

Radford M. Neal
Dept. of Statistics, University of Toronto

Abstract
When searching for new phenomena in high-energy physics, statistical analysis is complicated by the presence of nuisance parameters, representing uncertainty in the physics of interactions or in detector properties. Another complication, even with no nuisance parameters, is that the probability distributions of the models are specified only by simulation programs, with no way of evaluating their probability density functions. I advocate expressing the result of an experiment by means of the likelihood function, rather than by frequentist confidence intervals or p-values. A likelihood function for this problem is difficult to obtain, however, for both of the reasons given above. I discuss ways of circumventing these problems by reducing dimensionality using a classifier and employing simulations with multiple values for the nuisance parameters.

1 The Problem

I will discuss a class of problems that I hope at least resemble those encountered in high-energy physics experiments, such as searches for the Higgs Boson with the LHC. The solutions that I examine will have much in common with some present practice, though I will not attempt to provide comprehensive references. I hope that my discussion will clarify the role of existing techniques, such as the training of classifiers for 'signal' vesus 'background' events, and also point to possible new approaches.

In this paper, I deal with experiments where we will observe O events, indexed by $i = 1, \ldots, O$, that are described by variables, v_i, computed from the raw observational data. Events can either be from the 'background' or (if it exists) from the 'signal' — for instance, an event in which a previously-unobserved particle appears. I assume that simulation programs for background and signal events exist, which stochastically generate the variables from either a background distribution, which has probability density function $p_0(v)$, or a signal distribution, which has probability density function $p_1(v)$. The real events come from a *mixture* of signal and background distributions, with an unknown proportion, f, of signal. We may be most interested in whether or not f is zero — since $f > 0$ may, for instance, correspond to the existence of a previously-unknown particle.

Our first difficulty is that no explicit formulas for $p_0(v)$ and $p_1(v)$ exist. We may 'know' p_0 and p_1 in some sense, or we couldn't have written the simulator programs, but we have no way of translating this knowledge into a practical method for computing these density functions.

Our second difficulty is that, typically, we don't actually know p_0 and p_1 exactly. The simulators for generating from these distributions have some parameters — relating either to the physics or to the behaviour of the detector — whose values are not known precisely. Call these parameters ϕ. I'll assume that although ϕ is not known, we have a suitable prior distribution for ϕ, with density $p(\phi)$. Note that ϕ is a 'nuisance' parameter, since our only real interest is in f. The fact that ϕ is unknown is just an annoyance. (Though ϕ might be of interest to other people, such as the designers of the detector.)

2 The Role of the Likelihood Function

The *likelihood function* is the probability (or probability density) of the observed data, seen as a function of the model parameter(s). The likelihood function is defined only up to an arbitrary constant factor, and hence only ratios of likelihoods for different values of the parameters are meaningful.

When there are no nuisance parameters, the likelihood function for our problem (assuming independent observations) is a function of f alone:

$$L(f) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i) \;+\; (1-f) p_0(v_i) \Big] \tag{1}$$

Here, $fp_1(v_i) \;+\; (1-f)p_0(v_i)$ is simply the probability density for obtaining the observation v_i from either the signal distribution (with probability f) or the background distribution (with probability $1-f$). When there are nuisance parameters, ϕ, the likelihood is a function of both f and ϕ. I defer consideration of nuisance parameters to Section 4.

According to the *likelihood principle* (see, for example, the discussion by Cox and Hinkley [1], Section 2.3), the likelihood function contains all the information from the experiment that is relevant to inference for the parameters. So inference should not depend on aspects of the data that do not enter into the likelihood function. (An exception is that checks of the appropriateness of the model on which the likelihood function is based may utilize other aspects of the data.) Note that the likelihood function is itself a function of the data, and sometimes (not always!) depends only on some low-dimensional statistic computed from the data, such as the sample mean and/or the sample variance. A quantity computed from the data that can be used to compute the likelihood function is known as a 'sufficient statistic'. The likelihood function itself is a 'minimal sufficient statistic', containing no irrelevant information.

This 'weak' form of the likelihood principle is accepted by most statisticians. The 'strong' form, which is not universally accepted, says that the same conclusions should be drawn from two experiments (involving the same parameters) if they produced the same likelihood function. Bayesian inference obeys the strong likelihood principle, since it simply combines the likelihood with a prior distribution (which presumably does not vary with the choice of experiment). The strong likelihood principle is accepted by some non-Bayesian statisticians as well, however, partly because it follows from the weak likelihood principle together with a form of the principle that one should condition on an ancillary statistic (whose distribution does not depend on the parameters).

Classical (ie, non-Bayesian, frequentist) confidence intervals and p-values (other than those used for model checking) often violate the likelihood principle. For example, consider observations of n independent binary events, of which k turned out to be 1, with the remaining $n-k$ being 0. If we are interested in inferring the probability, θ, that an event is 1 (assumed the same for all events), the likelihood function will be $L(\theta) = \theta^k(1-\theta)^{n-k}$. In particular, if $n = 10$ and $k = 1$, the likelihood function is $L(\theta) = \theta(1-\theta)^9$. This likelihood function is the same regardless of whether we had decided to observe $n = 10$ events and found that $k = 1$ of them were 1, or we had decided to observe events until $k = 1$ of them were 1, and found that this was reached when $n = 10$. Hence, according to the likelihood principle, our conclusions should be the same in these two scenarios. However, the one-sided p-value for testing the null hypothesis that $\theta = 1/2$ versus the alternative that $\theta < 1/2$ is $P(k \leq 1) = (10+1)\,2^{-10}$ when the number of events is fixed at $n = 10$, but is $P(n \geq 10) = 2^{-9}$ when the number of 1 events is fixed at $k = 1$. Of the many arguments why such differing results should not be accepted, I will mention only consideration of an observer who knows everything that the experimenter does and sees, but doesn't know the experimenter's thoughts. Does this observer really need to ask the experimenter whether the stopping condition was $n = 10$ or $k = 1$ in order to draw an inference from the data? And would inference really be impossible if the experimenter had forgotten?

This issue arises also with Feldman and Cousins' [2] method for constructing confidence intervals from data, n, that is a sum of Poisson-distributed counts of 'signal' events (with unknown mean, μ) and 'background' events (with known mean, b). (We will see in Section 3 that this problem can arise as a much-reduced form of the problem discussed in this paper.) Their method (as well as some others) produces different confidence intervals for μ from an observed count of zero depending on the mean number of background events — the interval is tighter (with smaller upper limit) when the mean number of background events is higher. This violates the strong likelihood principle, since the likelihood functions for a count of zero from experiments with different background means differ only by a constant factor (which is irrelevant for likelihoods):

$$L(\mu) \;=\; \exp(-(b+\mu)) \;=\; \exp(-b)\,\exp(-\mu) \;\propto\; \exp(-\mu) \tag{2}$$

It makes intuitive sense that the inference drawn when the count is zero should not depend on the mean background — with a count of zero, we *know* that no background events occurred, so how many would occur on average if we were to repeat the experiment is of no relevance.

Feldman and Cousins are aware of this issue, but in responding to it, appear to have lost track of the scientific purpose of a statistical analysis, as is not uncommon in such discussions. They say that "for making decisions, [Bayesian inference] is probably how many scientists do (and should) think", but that "[classical] confidence intervals provide the preferred option for publishing numerical results of an experiment in an objective way. However it is critical not to interpret them as Bayesian intervals, i.e., as statements about $P(\mu_t|x_0)$". They remark with respect to the dependence of their intervals on the expected background when the observed count is zero that "We find that objections to this behaviour are typically based on a misplaced Bayesian interpretation of classical intervals, namely the attempt to interpret them as statements about $P(\mu_t|n_0)$." In further discussion of this situation, and in particular their method's production of confidence intervals for a count of zero that are tighter when the experiment is more poorly designed (with higher mean background), they say "The origin of these concerns lies in the natural tendency to want to interpret these results as the probability ... of a hypothesis given data rather than what they really are related to... It is the former that a scientist may want to know in order to make a decision, but the latter which classical confidence intervals relate to." In their discussion, they say nothing about what actual scientific use their classical confidence intervals might have, leaving (at least to me) the impression that they believe classical confidence intervals should be computed and reported simply as a ritual activity.

Fortunately, Feldman and Cousins do say that "it is important to publish relevant ingredients to the calculation so that the reader... can (at least approximately) perform alternative calculations or combine the result with other experiments". The "relevant ingredient" is in fact the likelihood function. Nothing more is needed, and nothing less than the full likelihood function would allow (for example) any Bayesian with any prior to make inferences.

When there is only a single parameter, such as f, the result of an experiment can easily be communicated fully by a plot of $L(f)$ versus f. In general, such a plot contains more information than a classical confidence interval, or any other interval that attempts to summarize the result. However, in many situations, the likelihood function approaches the exponential of a quadratic function as the amount of data increases (see [1], Section 10.6), and for simple Gaussian models, the log likelihood may be a quadratic function even for small samples. In such situations it is possible to specify the likelihood function using only two numbers (recall that the likelihood is defined only up to a constant factor). The end-points of a classical confidence interval can sometimes serve this purpose. If the parameter space is also unbounded, it is possible to develop intuitions about the meaning of classical confidence intervals that reflect what really matters — the likelihood function — explaining (in my view at least) how their use has survived.

However, when the log likelihood is not approximately quadratic, or when the parameter space is bounded, as in the example above where $L(\mu) \propto \exp(-\mu)$, with $\mu \in (0,\infty)$, applying these intuitions about confidence intervals, developed in another contexts, is dangerous. One simply cannot represent the various forms that a likelihood function can in general take using only two numbers. Obtaining the full likelihood function, or a good approximation to it, is therefore a crucial objective of statistical inference.

3 First Difficulty: We Can't Compute the Likelihood

Consider again our model with no nuisance parameters, with likelihood function given by equation (1). For a model like this with only a single scalar parameter, the full result of the experiment can easily be communicated by simply plotting the likelihood function. In typical problems, one can also easily find the maximum likelihood parameter estimate, as well as various Bayesian inferences, such as the posterior density obtained when some prior distribution is assumed.

But for our problem, we don't know how to compute the likelihood! So we can't easily produce a plot of $L(f)$ versus f. The likelihood involves p_0 and p_1, which are known only through simulation programs. If the v_i are low-dimensional (not more than around four dimensional), we could generate many points from p_0 and p_1, and use them to get good estimates for these density functions, but for high-energy physics experiments, it seems that it is more typical for each event to be described by dozens or hundreds of values. It might be possible to compute the p_0 and p_1 densities by using techniques similar to those used to compute free energies from Monte Carlo simulations, but these techniques would likely be too slow for this application, since the number (O) of observations for which these densities would need to be computed is typically quite large.

Fortunately, we only really need to compute the ratio $p_1(v_i)/p_0(v_i)$ for each observation. Since constant factors in the likelihood can be ignored, we can reduce the likelihood as follows:

$$\begin{aligned} L(f) &= \prod_{i=1}^{O} \Big[f p_1(v_i) \;+\; (1-f) p_0(v_i) \Big] && (3) \\ &= \prod_{i=1}^{O} p_0(v_i) \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (4) \\ &= \Big[\prod_{i=1}^{O} p_0(v_i) \Big] \cdot \prod_{i=1}^{O} \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (5) \\ &\propto \prod_{i=1}^{O} \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (6) \end{aligned}$$

Here, we can ignore the product of the $p_0(v_i)$ since factors in the likelihood not depending on f can be ignored. So we can look for a way to compute $p_1(v)/p_0(v)$ without having to compute $p_0(v)$ and $p_1(v)$.

One way to compute $p_1(v)/p_0(v)$ is to produce a classifier to distinguish signal and background events, training it on many simulated signal and background events, drawn according to p_1 and p_0. This is commonly done, as illustrated, for example, by [3]. The classifier could be based on neural networks, decisions trees, or many other methods, though I will assume here that the classifier can produces probabilities for the two classes, not just a guess at the class, with no indication of how likely it is to be correct. Suppose that the fraction of simulated events used to train such a classifier that are from the signal distribution is s. If we manage to train an excellent classifier, the probability it outputs that an event described by variables v is a signal event (call this $c(v)$) will match the true probability that such a simulated event is signal, so that

$$c(v) \;=\; \frac{s \, p_1(v)}{(1-s) \, p_0(v) \;+\; s \, p_1(v)} \tag{7}$$

Once we have this classifier, we can find the desired ratios as follows:

$$\frac{p_1(v)}{p_0(v)} \;=\; \frac{c(v)}{1-c(v)} \, \frac{1-s}{s} \tag{8}$$

If we really trust our classifier, we can now compute the likelihood function for f, and present a plot of $L(f)$ as the result of the experiment.

If we don't totally trust our classifier, we can still use it to get good results. We just treat it as a way of reducing the dimensionality of the data — from the multidimensional measurements, v_i, to the scalar $r_i = c(v_i)$ produced using the classifier. If the classifier were perfect, this reduction would not lose any useful information (since the r_i determine the likelihood function). If it's not perfect, it will throw away a bit of information, but the reduction to a scalar allows us to easily estimate $p_0(r)$ and $p_1(r)$ from simulation data, and use them to compute a likelihood function given the r_i. The results will be valid (ie, not systematically misleading), since this likelihood captures what can be learned from the experiment if one insists on reducing dimensionality in this way. However, the results may not be as precise (ie, as informative) as would have been obtained using a perfect classifier, which would have produced the true likelihood given all the information in the data.

One could reduce the data further by binning the r_i values, but this loses information. Using a fairly large number of bins might be OK, however, if it loses little information, and makes estimating the probabilities easier. If only two bins are used (ie, the output of the classifier is thresholded at some value), we get a Poisson count with background problem, of the sort discussed in Section 2 — assuming there are many events and signal events are rare, the number of events in the "signal" bin will be Poisson distributed, with some being real signal events and some being mis-classified background events. But such a drastic reduction of the data might throw away quite a bit of relevant information.

4 Second Difficulty: Nuisance Parameters for the Physics and the Detector Behaviour

In practice, we probably don't know p_0 and p_1 exactly. The simulators for generating from these distributions will have some parameters, ϕ, relating either to the physics or to the behaviour of the detector, which are not known precisely. (As a convenience, we can assume that ϕ is the same for simulating p_0 and p_1, since we can let some components of ϕ be used by only one of these simulators.)

We have to assume that these ϕ parameters are known to some degree, or there's no hope of solving the problem. I'll assume that based on theory or previous experiments, a prior distribution for ϕ is available, with density $p(\phi)$. It's unlikely that this prior will be perfect — eg, it might assume independence of components of ϕ when it really ought not to. We must hope that the results are not too sensitive to this — formally checking whether this is true is a difficult problem.

Once there are ϕ parameters, the likelihood is a function of both f and ϕ:

$$L(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i|\phi) \;+\; (1-f) p_0(v|\phi) \Big] \tag{9}$$

where $p_0(v|\phi)$ and $p_1(v|\phi)$ denote probability densities for generating v from the background and signal simulators with parameters set to ϕ.

This is a high dimensional function (since ϕ is typically high dimensional), and hence will be difficult to visualize. Just plotting $L(f,\phi)$ will *not* be a feasible way of presenting the results of the experiment. Many ways of dealing with this type of problem have been proposed. For instance, we might look at the "profile likelihood", a function of f alone defined as $\sup_\phi L(f,\phi)$. This ignores many aspects of the likelihood function, however. A Bayesian approach is to instead integrate $L(f,\phi)$ with respect to a prior distribution for ϕ, to obtain a *marginal likelihood function* for f alone:

$$\underline{L}(f) \;=\; \int L(f,\phi)\, p(\phi)\, d\phi \tag{10}$$

We often could compute this fairly easily by simple Monte Carlo (sampling from the prior for ϕ), if we could compute $L(f,\phi)$. Since the marginal likelihood is one-dimensional, we would then be able to present the result of the experiment by simplying plotting $\underline{L}(f)$, if we could compute it.

The role of the prior, $p(\phi)$, in producing a marginal likelihood is worth examining. When there are no nuisance parameters, the likelihood function $L(f)$ is an 'objective' presentation of the experimental result (if one ignores subjectivity in the choice of model). Inferences can then be drawn using this likelihood in various, possibly 'subjective', ways. There is no need for the experimenters to draw such inferences (though they may of course do so if they wish), and hence no need for them to choose a prior for the parameter of interest, f. The situation is different for the nuisance parameters, ϕ, since many components of ϕ will relate to experimental details about which the experimenters are much more knowledgeable than anyone else. It therefore seems most sensible for the experimenters to decide on a suitable prior, $p(\phi)$, and use this to produce a marginal likelihood, $\underline{L}(f)$, that can be interpreted by others.

Unfortunately, actually computing $L(f,\phi)$, and from it $\underline{L}(f)$, is at least as difficult as computing $L(f)$ when there are no nuisance parameters. We might try, as in Section 3, to rewrite the likelihood in terms of ratios of probabilities:

$$L(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i|\phi) \;+\; (1-f) p_0(v_i|\phi) \Big] \tag{11}$$

$$= \;\Big[\prod_{i=1}^{O} p_0(v_i|\phi) \Big] \cdot \prod_{i=1}^{O} \Big[f\, \frac{p_1(v_i|\phi)}{p_0(v_i|\phi)} \;+\; (1-f) \Big] \tag{12}$$

However, unlike before, the first factor is now relevant, since it depends on the parameter ϕ. Properties of events that are irrelevant for classifying them as signal versus background may still be relevant for inferring ϕ, and hence indirectly for inferring f.

As we did in Section 3, we might try to avoid our difficulties by reducing the dimensionality of the data. If we can map the high-dimensional v_i to quantities r_i that are low-dimensional (and then possibly binned), it will be feasible to estimate $p_0(r_i|\phi)$ and $p_1(r_i|\phi)$ using a reasonable number of events generated by the simulators. We probably can't expect to reduce dimensionality in a way that preserves all relevant information, but we can hope to keep the loss of information small.

We can define a likelihood function based on this reduced data:

$$L_r(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(r_i|\phi) \;+\; (1-f) p_0(r_i|\phi) \Big] \tag{13}$$

Since we have likely lost information by going from v_i to r_i, this is not the same function as $L(f,\phi)$. But it can be used to make valid (though less efficient) inferences, provided that the mapping from v to r was not chosen based on the observed data. We can again define a marginal likelihood, integrating over the prior for ϕ:

$$\underline{L}_r(f) \;=\; \int L_r(f,\phi)\, p(\phi)\, d\phi \tag{14}$$

If we can compute this, plotting it will display the results of the experiment, as well as possible given our computational limitations, which forced the reduction from v_i to r_i.

To compute $\underline{L}_r(f)$, we could choose K values for ϕ from the prior, labelled $\phi_1,\ldots,\phi_K$, either randomly or by some quasi-Monte Carlo scheme, and then average $L_r(f,\phi_k)$ over these K values to approximate the integral above. Computing $L_r(f,\phi_k)$ will require simulating many events from the background and signal distributions with parameters ϕ_k, and then using these to estimate the probability densities $p_0(r|\phi_k)$ and $p_1(r|\phi_k)$ (or the bin probabilities, if the r_i were binned).

Though not easy, this computation seems to be feasible, provided a value for K in the hundreds or thousands is adequate, and the dimensionality of the r_i is small enough that for each ϕ_k the densities $p_0(r|\phi_k)$ and $p_1(r|\phi_k)$ can be adequately modeled using a few thousand events generated by each of the simulators. (Alternatively, one might try to build one general model for the conditional densities $p_0(r|\phi)$ and $p_1(r|\phi)$ using data generated with all values ϕ_k.) The total number of simulated events required would then be no more than a few tens of millions.

The choice of K is not easy, however. If one is sure that $L_r(f, \phi)$ does not vary drastically with ϕ, a value for K of a few hundred would suffice. However, if drastic variation is conceivable, a much larger value of K might be needed in order to be confident that the results are valid. Suppose, for example, that the actual observations are such that for most ϕ, $L(0, \phi)$ is small (compared to $L(f, \phi)$ for some $f > 0$), but that in some region of ϕ values with small but not negligible prior probability, $L(0, \phi)$ is very much larger, sufficiently so that this region dominates the integral defining $\underline{L}(0)$ — or to put it another way, if ϕ is in this region, the experiment will produce many more background events that look like signal events than for other values of ϕ. If none of the K values for ϕ_k that were chosen happen to lie in this region, the value for $\underline{L}(0)$ that is computed will be much smaller than the true value. If this situation is a possibility, one would need to use a value for K that is sufficiently large for the p-value, or other measure of confidence, that one aims to report for a discovery (certainly no smaller than the reciprocal of this p-value, and preferrably somewhat larger), in order to reduce to the required level the chance that such problematic values for ϕ might have been missed.

The most difficult problem is deciding how to reduce dimensionality. Training a classifier to distinguish signal from background still seems to be a useful way of isolating relevant information. This might be done in several ways, however, and we may also wish to preserve other information, relating to the first factor in equation (12).

We could train a classifier using background and signal events generated using a single value for the nuisance parameters (eg, the prior mean), reducing the data from v_i to $r_i = P(\text{signal}|v_i)$, as approximated by this classifier. This is much the same as we would do if there were no nuisance parameters (equivalently, if the correct ϕ were known). The predictions could of course be binned, and with only two bins, we would end up with Poisson-distributed counts in which the mean background is uncertain, but with a distribution that can be estimated from simulations, as described more generally above.

The danger of this approach is that the classifier that is trained may not work well for events generated with other values of ϕ, and in particular, may not work well for the true ϕ. If so, the number of background events misclassified as signal may be large, leading to a substantial loss of information (though not to misleading results, if the rest of the inference task is properly done).

Another simple approach is to generate events with many values of ϕ, drawn from the prior (a different ϕ_i for every v_i), and train a classifier with v_i as inputs on all of this data. The classifier might then learn how to distinguish signal from background in a robust way, that works for all ϕ. Of course, it could well be that there is no way to accurately classify without knowing ϕ, in which case this method will also lose much information.

When neither of these approaches work, it seems that one must rely on the data being informative about ϕ. I sketch here a scheme that may perhaps provide a feasible solution in this situation, based on reducing the dimensionality of both ϕ and v.

As before, we will train a classifier for signal versus background events, using data generated from the two simulation programs, with some fraction s of signal events, using values for ϕ drawn from its prior. This classifier will take both ϕ and v as inputs — ie, it learns to classify for any value of ϕ, provided that the correct ϕ is known. Furthermore, this classifier will contain "bottlenecks" for both ϕ and v, which force the classifier to learn to use reduced-dimension versions of these inputs. In detail, we specify some small dimensionality for ϕ^* and some small dimensionality for v^* (eg, perhaps both are two-dimensional), and then train a classifier that has the following functional form for the estimated probability that an event is from the signal distribution:

$$P(\text{signal}|v, \phi) \quad \approx \quad d(g(\phi), h(v)) \tag{15}$$

where $\phi^* = g(\phi)$ is the reduced form of ϕ and $v^* = h(v)$ is the reduced form of the variables describing the event. The functions d, g, and h are parameterized in some way, such as with a multilayer perceptron ('backprop') neural network. Training of the classifier is done by adjusting these parameters to match the data as sell as possible (eg, by some maximum penalized likelihood criterion). Dimensionality reduction schemes similar to this have long been used with neural networks (eg, see [4]).

We cannot use this classifier on real data, since we don't know the correct value for ϕ — nor for ϕ^*, which is all we would need. The only reason to train this classifier is to obtain the mappings $\phi^* = g(\phi)$ and $v^* = h(v)$, which, if the classifier is successful, must preserve most of the relevant information from ϕ and v. Note that we *can* obtain $v_i^* = h(v_i)$ for all the real events, since (once training is finished) h does not depend on the unknown value of ϕ. To obtain information about ϕ^* from the real events, we use simulated data to train a regression model (probably non-linear, perhaps a neural network) that approximates the expectation of ϕ^* given a single observation, v, by $e(v)$, where e is another parameterized function, learned from the data. Using data simulated only from the background distribution, p_0, may be sufficient, since even when the fraction of signal events is non-zero, it is typically quite small. This regression model also is merely used for dimensionality reduction, and will not be directly used to predict ϕ^*.

We are now in a position to produce the reduced data we will use for inference, consisting of the pairs $r_i = (h(v_i), e(v_i))$. If the dimensionality of these pairs is quite small, we can hope to build good models of the conditional densities $p_0(r|\phi)$ and $p_1(r|\phi)$ — most crudely, just by binning values for r, and using many simulated events with each value for ϕ, though more sophisticated methods may work better. Inference for f is then based on the marginal likelihood defined in equation (14), along with equation (13), computed by Monte Carlo, using a sample of values for ϕ drawn from its prior.

Note that the validity of the inference will depend only on the accuracy with which these densities are estimated, not on the quality of the classification and regression models described above. However, if these models are poor, much information may be lost, so the inferences may be uninformative. If instead our models are good, the $e(v_i)$ component of the r_i values will carry information about ϕ^*, with the result that most of the weight in the integral of (14) will be on values of ϕ close to the true value (or at least for which ϕ^* is close to its true value). For these values of ϕ, the likelihood will tend to favour values for f near the true value, because of the information carried in the $h(v_i)$ portion of r_i.

This scheme is untested, and may prove in practice to either lose too much information or be computationally infeasible — it may, for instance, prove necessary to train something more than a simple regression model in order to learn about ϕ^*. We can hope that success will be achievable using some such strategy for reducing dimensionality so that estimates for probabilities or probability density functions become feasible, allowing an approximation to the marginal likelihood function to be computed. When the parameter of interest (f in this problem) is one-dimensional (or more generally, of low enough dimension that a plot is intelligible), such a likelihood function is the most complete, and most useful, report of the experimental result.

Acknowledgements

I thank Nancy Reid, Zi Jin, and Byron Roe for helpful discussions. This research was supported by the Natural Sciences and Engineering Research Council of Canada. The author holds a Canada Research Chair in Statistics and Machine Learning.

References

[1] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman & Hall/CRC, 1974.

[2] G. J. Feldman and R. D. Cousins, "A unified approach to the classical statistical analysis of small signals", `http://arxiv.org/abs/physics/9711021v2`, 1997.

[3] H.-J. Yang, B. P. Roe, and J. Zhu, "Studies of Boosted Decision Trees for MiniBooNE Particle Identification", `http://arxiv.org/abs/physics/0508045`, 2005.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, vol. 313. no. 5786, pp. 504 - 507, 2006.