To appear in: Maximum Entropy and Bayesian Methods, Eds. K. Hanson, R. Silver, 1995 (in press 1995)

DETERMINING WHETHER TWO DATA SETS ARE FROM THE SAME DISTRIBUTION

David H. Wolpert
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501 USA (dhw@santafe.edu)

ABSTRACT. This paper presents two Bayesian alternatives to the chi-squared test for determining whether a pair of categorical data sets were generated from the same underlying distribution. It then discusses such alternatives for the Kolmogorov-Smirnov test, which is often used when the data sets consist of real numbers.

1. Introduction

Let d_1 and d_2 be two sets of elements from a space X, with cardinalities N_1 and N_2, respectively. View d_1 and d_2 as "samples" of two distributions over X, p_1 and p_2 respectively. Based on d_1 and d_2, do we believe that p_1 and p_2 are equal (or at least close approximations of each other), and how confident are we in this belief?

When X is finite, the traditional approach to this common problem is the chi-squared test. When X is the set of real numbers, the traditional approach is instead the Kolmogorov-Smirnoff test. (See [2] and references therein for discussions of both tests.) Both of these tests can be viewed as types of null-hypothesis tests. Accordingly, they suffer from a number of defects: they are average-data rather than this-data; they can (sort of) rule out the null, but not "rule it in"; they are very dependent on issues like the "power" and "size" of the statistic, etc.

Clearly a Bayesian alternative to these tests, properly constructed, would be preferable. Some have confused such an alternative with, for example, proofs (like in [1]) that in the proper limit the chi-squared test approximates a Bayesian procedure of some sort. What is instead needed is a first-principles Bayesian approach to the problem, which in general may have no relation to tests like chi-squared.

In this paper two such approaches are worked out in detail for the finite X case, and some possible approaches are mapped out for the uncountable X case. See also [3] for related work.

2. Finite X—the posterior expected difference approach

Consider the case where X has a finite number of possible values, m. For this scenario, both p_1 and p_2 consist of m real numbers, all of which are non-negative, and which sum to 1. I will indicate the i'th component of p_j by $p_j(i)$. Also I will indicate the histograms over the X values induced by d_1 and d_2 as $d_1(i)$ and $d_2(i)$ respectively. So $d_j(i)$ is the number of elements in d_j that have the i'th X value.

Let $S(p_1, p_2)$ be a measure of the similarity between p_1 and p_2. For simplicity, in this paper I will concentrate on the quadratic distance measure, $S(p_1, p_2) \equiv \sum_{i=1,m}(p_1(i) - p_2(i))^2$. However the calculations presented here can be applied to any analytic S by expanding that S in a Taylor series. Moreover, as is illustrated below, for many non-quadratic S simple tricks allow one to perform the calculations without resorting to such an expansion.

In this section I will show how to calculate the posterior expected value of S, S_1, and the posterior expected value of S^2, S_2. The formula for S_1 provides a measure of how much the data indicates that p_1 and p_2 differ, and in $\sqrt{S_2 - (S_1)^2}$ we have an error bar for that measure.

First note that

$$\begin{aligned} S_1 &\equiv \int dp_1 dp_2 \; S(p_1, p_2) \; P(p_1, p_2 \mid d_1, d_2), \quad \text{and} \\ S_2 &\equiv \int dp_1 dp_2 \; S^2(p_1, p_2) \; P(p_1, p_2 \mid d_1, d_2). \end{aligned}$$

By Bayes' theorem $P(p_1, p_2 \mid d_1, d_2) \; \propto \; P(d_1 \mid p_1) \; P(d_2 \mid p_2) \; P(p_1, p_2)$. Now assume the d_i were created by IID sampling of p_i: $P(d_i \mid p_i) \; \propto \; \prod_{j=1}^{m} \; p_i(j)^{d_i(j)}$. All that remains to fully fix the integrals for S_1 and S_2 is the prior, $P(p_1, p_2)$.

2.1. Calculating the first moment

Given the preceding, up to an overall proportionality constant, S_1 is given by

$$J \equiv \int dp_1 dp_2 \; [\prod_{j=1}^{m} \; p_1(j)^{d_1(j)} \; p_2(j)^{d_2(j)}] \; P(p_1, p_2) \; S(p_1, p_2). \tag{1}$$

The proportionality constant is set by normalization and equals

$$K \equiv \int dp_1 dp_2 \; \prod_{j=1}^{m} \; p_1(j)^{d_1(j)} \; p_2(j)^{d_2(j)} \; P(p_1, p_2). \tag{2}$$

To proceed further we must specify the prior, $P(p_1, p_2)$. Write

$$\begin{aligned} P(p_1, p_2) \; = \; & F(p_1, p_2) \times \delta\{(\sum_{j=1}^{m} p_1(j)) - 1\} \; \times \; \prod_{j=1}^{m} \theta(p_1(j)) \\ & \times \delta(\{\sum_{j=1}^{m} p_2(j)) - 1\} \; \times \; \prod_{j=1}^{m} \theta(p_2(j)), \end{aligned} \tag{3}$$

where the (Dirac) delta functions force each p_i to have its components sum to 1, and the (Heaviside) theta functions force all such components to be non-negative.

Consider the case where $F(p_1, p_2)$ is analytic, i.e., where it can be written as a sum of products of powers of the components of p_1 and p_2. Note that the likelihood term in our integrals is simply a product of powers of the components of p_1 and p_2. Accordingly, if we know how to calculate S_1 and S_2 for the case where $F(p_1, p_2)$ is a constant, by using appropriate modifications of the d_i we can calculate S_1 and S_2 for any analytic F. (I.e., the Dirichlet prior is the conjugate prior for this problem.) Accordingly, without loss of generality, from now on I will take $F(p_1, p_2) = 1$.

The integrals J and K are relatively straight-forward to evaluate [4]:

$$K(d_1, d_2) \;=\; \frac{\prod_{i=1}^{m} \Gamma(d_1(i)+1)\ \Gamma(d_2(i)+1)}{\Gamma(N_1+m)\ \Gamma(N_2+m)}. \tag{4}$$

where "$\Gamma(.)$" is the gamma function. (For current purposes, where the $d_j(i)$ are integers, the gamma function is just a factorial.) Next, use the expansion $S(p_1, p_2) \;=\; \sum_{i=1}^{m} [p_1(i) - p_2(i)]^2 \;=\; \sum_{i=1}^{m} \{[p_1(i)]^2 \;+\; [p_2(i)]^2 - 2p_1(i)p_2(i)\}$. This gives

$$J \;=\; \sum_{i=1}^{m} \{K[d_1 + 2(i), d_2] \;+\; K[d_1, d_2 + 2(i)] \;-\; 2K[d_1 + 1(i), d_2 + 1(i)]\}. \tag{5}$$

where by "$d_i + t(j)$" is meant the histogram d_i with t extra counts added to the j'th bin. Note that the total number of counts in $d_i + t(j)$ is $N_i + t$; this must be taken into account when plugging the formula for K into the formula for J.

With Eq.'s (4) and (5) we can calculate the posterior expected value of S, J/K.

2.2. Calculating the second moment

S_2 is calculated in a similar way to the calculation of S_1. It can be written as $S_2 \;=\; \frac{Z_1+Z_2}{K}$, where K is the same as in Eq. (4), and Z_1 and Z_2 are linear combinations of K's. To evaluate this ratio first define $g(x,y) \equiv (x+y)!/x!$ and $n_i \equiv N_i + m - 1$. Then

$$\begin{aligned}\frac{Z_1}{K} \;=\; & \sum_{i=1}^{m} \Big\{ \frac{g(d_1(i),4)}{g(n_1,4)} - 4\frac{g(d_1(i),3)\ g(d_2(i),1)}{g(n_1,3)\ g(n_2,1)} \;+ \\ & 6\ \frac{g(d_1(i),2)\ g(d_2(i),2)}{g(n_1,2)\ g(n_2,2)} \;-\; 4\frac{g(d_1(i),1)\ g(d_2(i),3)}{g(n_1,1)\ g(n_2,3)} \;+\; \frac{g(d_2(i),4)}{g(n_2,4)} \Big\}.\end{aligned} \tag{6}$$

Note that this can be broken up into five separate sums - doing that, you only need to perform five separate divisions (no denominator term involves the summation variable i). Moreover, the $g(.,.)$ terms can be pre-calculated.

A similar calculation gives the following:

$$\begin{aligned}\frac{Z_2}{K} \;=\; 2\sum_{i<j} \Big\{ & \frac{g(d_1(i),2)\ g(d_1(j),2)}{g(n_1,4)} \;+\; 2\frac{g(d_1(i),2)\ g(d_2(j),2)}{g(n_1,2)\ g(n_2,2)} \;+ \\ & \frac{g(d_2(i),2)\ g(d_2(j),2)}{g(n_2,4)} \;-\; 4\frac{g(d_1(i),2)\ g(d_1(j),1)\ g(d_2(j),1)}{g(n_1,3)\ g(n_2,1)} \;- \\ & 4\ \frac{g(d_2(i),2)\ g(d_1(j),1)\ g(d_2(j),1)}{g(n_2,3)\ g(n_1,1)} \;+ \\ & 4\ \frac{g(d_1(i),1)\ g(d_2(i),1)\ g(d_1(j),1)\ g(d_2(j),1)}{g(n_1,2)\ g(n_2,2)} \Big\}.\end{aligned} \tag{7}$$

As in evaluating Z_1/K, it makes sense to break up this expression into a set of sums (to reduce the number of divisions to six) and precompute quantities like g's. We would still seem to have an m^2 calculation though.

To get around this, use the following identity:

$$\sum_{i \neq j} u(j)\ v(j)\ U(i)\ V(i) \;=\; [\sum_j u(j)\ v(j)]\ [\sum_i U(i)\ V(i)] \;-\; \sum_i u(i)\ v(i)\ U(i)\ V(i). \quad (8)$$

Next note that each of the g's in the numerators in Eq. (7) is a product of no more than two terms. This allows us to evaluate all products of those g's using Eq. (8), and thereby make the entire calculation linear in m. More precisely, for

i) the first term in Eq. (7), set $u = U, v = V$, $u(j) = d_1(j) + 2$, and $v(j) = d_1(j) + 1$;
ii) the second term, set $u(j) = d_2(j) + 2, v(j) = d_2(j) + 1, U(j) = d_1(j) + 2$, and $V(j) = d_1(j) + 1$;
iii) the third term, set $u = U, v = V$, and $u(j) = d_2(j) + 2, v(j) = d_2(j) + 1$;
iv) the fourth term, set $u(j) = d_1(j) + 1, v(j) = d_2(j) + 1, U(j) = d_1(j) + 2$, and $V(j) = d_1(j) + 1$;
v) the fifth term, set $u(j) = d_1(j) + 1, v(j) = d_2(j) + 1, U(j) = d_2(j) + 2$, and $V(j) = d_2(j) + 1$;
vi) the sixth term, set $u(j) = d_1(j) + 1, v(j) = d_2(j) + 1, U(j) = d_1(j) + 1$, and $V(j) = d_2(j) + 1$.

The details of plugging all of this into Eq. (7) are not too illuminating, and in the interests of space are left as an exercise for the reader.

2.3. Comments

Adding data doesn't change the number of operations needed to calculate S_1 and S_2. On the other hand, despite the large degree of cancellation in our equations (e.g., when one divides Eq. (5) by Eq. (4)), things do get more expensive as one increases m, the number of bins. This is because we have many products over bins. Here conventional tricks like operating in logarithm space (so products become sums) are needed to keep the computational time (not to mention underflow and overflow problems) tractable.

Finally, simple tricks allow evaluation of S_1 and S_2 for some non-quadratic choices of $S(p_1, p_2)$, without going to the trouble of Taylor expanding such an S. For example, to evaluate the posterior moment of the Kullback-Leibler distance between p_1 and p_2, one has to be able to evaluate integrals of the form $\int dp_1 dp_2 \ \ln(p_1(i)) \prod_{j=1}^m [p_1(j)^{\hat{d}_1(j)} \ p_2(j)^{\hat{d}_2(j)}] \ P(p_1, p_2)$ (where $\hat{d}_1$ and $\hat{d}_2$ are in general slight variants of d_1 and d_2). To do this we can use Eq. (4) and the simple identity $x^a \ln(x) \;=\; (\partial_n x^n)|_{n=a}$ to get logarithms into the integrands [4].

3. Finite X—the ratio of posteriors approach

The technique outlined above assigns measure 0 to the set of events $p_1 = p_2$. I.e., it says that it is impossible for p_1 to equal p_2, regardless of the data. It is possible to use a Bayesian technique that instead assigns comparable probabilities to the two models $M_1 \equiv \{p_1 = p_2\}$

and $M_2 \equiv \{p_1 \neq p_2\}$. To see how first write

$$P(d_1, d_2 \mid M_1) = \int dp_1 dp_2 \; P(d_1, d_2 \mid M_1, p_1, p_2) \; P(p_1, p_2 \mid M_1), \quad \text{and then}$$
$$P(p_1, p_2 \mid M_1) = P(p_2 \mid p_1, M_1) \; P(p_1 \mid M_1) = \delta(p_1 - p_2) \; G(p_1) \; stuff(p_1),$$

where "*stuff(p₁)*" is the usual expression forcing p_1 to be a probability distribution.

Write $P(d_1, d_2 \mid M_1, p_1, p_2) = P(d_1 \mid p_1) \; P(d_2 \mid p_2)$, so

$$P(d_1, d_2 \mid M_1) = \int dp_1 \; G(p_1) \; stuff(p_1) \prod_{j=1}^{m} p_1(j)^{d_1(j)+d_2(j)} \; \frac{N_1! \; N_2!}{\prod_{j=1}^{m} (d_1(j))! \; (d_2(j))!}.$$

As usual, to analyze the analytic G case it suffices to consider the case where G is a constant. Being careful to maintain normalization, for this case

$$P(d_1, d_2 \mid M_1) = \frac{(m-1)! \; (N_1)! \; (N_2)!}{(N_1 + N_2 + m - 1)!} \times \frac{\prod_{j=1}^{m} (d_1(j) + d_2(j))!}{\prod_{j=1}^{m} (d_1(j))! \; (d_2(j))!}. \tag{9}$$

Next write $P(d_1, d_2 \mid M_2) = \int dp_1 dp_2 \; P(d_1 \mid p_1) \; P(d_2 \mid p_2) \; F(p_1, p_2) \; stuff(p_1, p_2)$. Again, take F constant. (As an aside, if one wants non-constant F and G, it probably makes sense to have them "correspond" in some way.) This gives

$$P(d_1, d_2 \mid M_2) = \frac{[(m-1)!]^2 \; (N_1)! \; (N_2)!}{(N_1 + m - 1)! \; (N_2 + m - 1)!}. \tag{10}$$

This depends only on N_1, N_2 and m; no other aspects of the d_i are relevant.

Finally, use Eq.'s (9) and (10) to get the ratio of the posteriors of the models:

$$\frac{P(M_1 \mid d_1, d_2)}{P(M_2 \mid d_1, d_2)} = \frac{P(M_1)}{P(M_2)} \times \frac{P(d_1, d_2 \mid M_1)}{P(d_1, d_2 \mid M_2)}. \tag{11}$$

This posterior ratio for the uniform F and G case is extremely quick to evaluate and in many respects is at least as "reasonable" in its behavior as the traditional chi-squared test. Nonetheless, one may want to consider non-uniform F and G. In particular, non-uniform F raises/lowers the probabilities of $p_1 - p_2$ pairs for which $p_1 \neq p_2$ but which lie close to $\{p_1 = p_2\}$. So for example, if $d_1 = d_2$, then having F favor $p_1 - p_2$ pairs that lie close to $\{p_1 = p_2\}$ will "leach" some of the posterior probability of M_1 into the posterior probability of M_2. This is because F will be favoring $p_1 - p_2$ pairs that can reasonably explain the data.

4. The Uncountable X Case

For real-valued X, binning X (so that the techniques of the previous section can be applied) is sometimes problematic. That is because the final result can depend on the binning scheme used. One obvious potential solution to this problem is to take inspiration from the Kolmogorov-Smirnov test: have the statistic concern differences in the cumulative distribution functions (CDF's) rather than the density functions directly. For example, one might define $S(p_1, p_2) \equiv \sum_{i=1}^{m} [CDF_1(i) - CDF_2(i)]^2$. Since the CDF's tend to be relatively insensitive to the precise binning, with this scheme how you bin should not be a big problem.

Another possibility is to use a prior that favors smooth p_i, so that $p_i(j)$ is close to $p_i(k)$ if bin j is close (in X) to bin k. Such a prior can be used with either of the posterior ratio or statistic moments approaches. For the latter a CDF-based statistic is not needed; a conventional (e.g., quadratic) S could be used.

A third possibility is not to bin, but rather consider a parameterized set of p_i. Under this scheme one could use either of the posterior ratio or moments of S approaches. However now the integrals would be over the parameters of the p_i rather than over the p_i directly.

Finally, there are some schemes that involve neither binning nor parameters. For example, one could define a new space $Y \equiv d_1 \cup d_2$ and do the analysis in that space. So the p_i are now distributions over Y, and the values in the histograms of the d_i are all 0's and 1's (assuming there are no delta functions in $P(p_1(X), p_2(X))$, so there are no duplicates in $d_1 \cup d_2$). The idea would be to have the prior favor smooth p_i, where the degree to which $p_i(j)$ is pushed towards $p_i(k)$ depends on the distance between the X values corresponding to elements j and k of Y.

Future work involves comparing these schemes to other Bayesian procedures (F. Ruggeri—private communication) related to the Kolmogorov-Smirnov test.

ACKNOWLEDGMENTS: I would like to thank Tony Begg and especially Hank Vaccaro for helpful discussions. This work was supported in part by TXN.

References

[1] D. Lindley, *Introduction to probability and statistics 2*, Cambridge University Press. (1965).

[2] W.H. Press et al, *Numerical Recipes in C*, Cambridge University Press. (1992).

[3] D.R. Wolf, *Mutual Information as a Bayesian Measure of Independence*, send email to "comp-gas@xyz.lanl.gov" with subject "get 9511002".

[4] D.H. Wolpert, D.R. Wolf, *Estimating functions of probability distributions from a finite set of samples*, Physical Review E, in press. (1995).