Specifying expert priors for species delimitation

Graham Jones

2015-10-23, April 18, 2016

Most attempts to infer species delimitations use genetic data only. It is usually the case that geographical and morphological information is available as well (Zhang et al., 2014). One way to incorporate extra information is to model the evolution of morphological characters and the changes in locations. An alternative approach is for experts in the organisms to formalize their knowledge in the form of a prior on the space of all possible clusterings. This note focuses on this second approach. I will refer to these experts as 'taxonomists', although they may not classify themselves like that.

It is unreasonable to expect taxonomists to provide a probability for each possible clustering, unless the number of individuals is very small: there are 203 clusterings of 6 individuals and the number of clusterings increases rapidly for more individuals. Furthermore, there are few taxonomists who also have a good understanding of probability theory. Here, we assume that for each pair of individuals, an estimated probability that they both belong to the same species has been provided, and describe a way of specifying a prior distribution on the space of all possible clusterings which respects the pairwise probabilities.

1 The method

The individuals are labelled $1, 2, \dots n$. For each i, j with $1 \leq i < j \leq n$, a number $p_{ij} \in [0, 1]$ is assumed to be given, with $p_{ji} = p_{ij}$ and $p_{ii} = 1$. It is sometimes more convenient to use the complementary probabilities $\bar{p}_{ij} = 1 - p_{ij}$. These are the estimated probabilities that the two individuals belong to different species, and could more loosely be called dissimilarities or distances. We say that such a set of values p_{ij} is **consistent** if there is probability distribution over the space of all possible clusterings which results in them.

We do not assume that the provided values are consistent with one another. For example, suppose $n = 3$ and $p_{12} = 1.0$, $p_{13} = 1.0$, $p_{23} = 0.0$ were given. This is inconsistent, since if 1 and 2 belong to the same species, and so do 1 and 3, then 2 and 3 must as well. Any value for p_{23} other than 1 would be inconsistent with the other two values. It is less obvious, but the values $p_{12} = 0.8$, $p_{13} = 0.9$, $p_{23} = 0.6$ are also inconsistent. There is no probability distribution over the 5 possible clusters 123, 1+23, 2+13, 3+12, 1+2+3 which results in these pairwise probabilities. In the case of triplets, it is easy to test for consistency. The rule is that for each triplet i, j, k the triangle inequality must be satisfied for the complementary probabilities:

$$\bar{p}_{ij} + \bar{p}_{ik} \geq \bar{p}_{jk}$$

or equivalently,

$$p_{jk} \geq p_{ij} + p_{ik} - 1.$$

However, things rapidly become more complex as the number of individuals increases, and it does not seem feasible to insist that taxonomists provide consistent estimates of the pairwise probabilities. (It might be sensible to check triplets and quartets and warn about inconsistencies as this might catch some errors.)

One way of representing the set of pairwise values is as the edge weights in the complete graph whose nodes are are the set $S = \{1, 2, \dots, n\}$. See Figure 1. If we only consider pairs which correspond to a set of edges T which do not form a cycle, then no inconsistencies can arise. Given a particular clustering X (one that the MCMC chain could visit) we can assign it a value as follows.

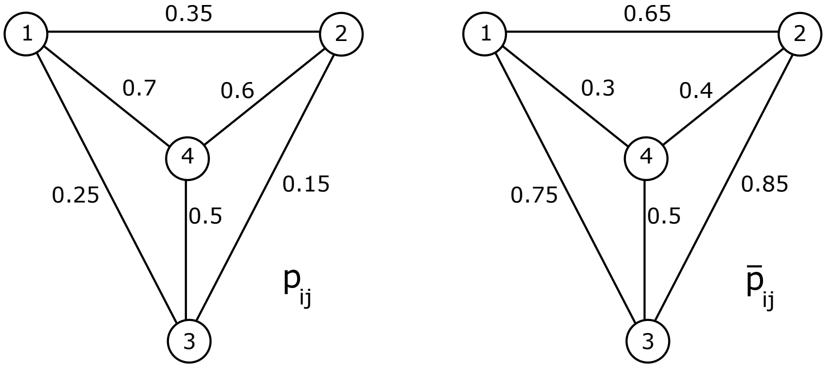

Figure 1: An example of pairwise values for 4 individuals labelled 1,2,3,4. On the left are probabilities p_{ij} that the pairs belong to the same species, as might be estimated by a taxonomist for each of the 6 pairs. On the right are the complementary probabilities $\bar{p}_{ij} = 1 - p_{ij}$.

Let $x_{ij}(X)$ be 1 if i and j belong to the same cluster in X, and zero otherwise. We regard the pairwise values for edges in T as independent. Then define

$$f(X, T) = \prod_{ij \in T} 1 - |x_{ij}(X) - p_{ij}|. \tag{1}$$

The term $1 - |x_{ij} - p_{ij}|$ is p_{ij} if $x_{ij} = 1$, and $\bar{p}_{ij}$ otherwise. Thus $f(X, T)$ is the probability that would naturally be given to X if the only information available was from those pairs in T. Note that the sum of $f(X, T)$ over all X will not sum to 1 in general so this is an unnormalised distribution.

Each set of edges T gives a 'restricted view' of the complete set of pairwise values. In order to get as big a view as possible while avoiding the possibility of inconsistencies, we need the **spanning trees** of the graph. See Figure 2. (They should not be confused with phylogenetic trees. They are the same type of mathematical object as an unrooted and perhaps multifurcating phylogenetic tree. But in these trees, all the nodes, internal as well as tips, are individuals.) Denote the set of all spanning trees on S as $\mathcal{T}(S)$. A simple way of combining the 'views' is to average over all spanning trees. We would therefore like to evaluate

$$f(X) = \sum_{T \in \mathcal{T}(S)} \prod_{ij \in T} f(X, T) = \sum_{T \in \mathcal{T}(S)} \prod_{ij \in T} 1 - |x_{ij}(X) - p_{ij}| \tag{2}$$

This appears difficult, but there is a theorem to help us. It is Kirchhoff's matrix tree theorem (`https://en.wikipedia.org/wiki/Kirchhoff's_theorem`), or more precisely a generalisation of this. It is called the 'Souped-Up Matrix-Tree Theorem' in these lecture notes (`http://www.math.ku.edu/~jmartin/mc2004/graph1.pdf`) from Jeremy L. Martin, University of Kansas.

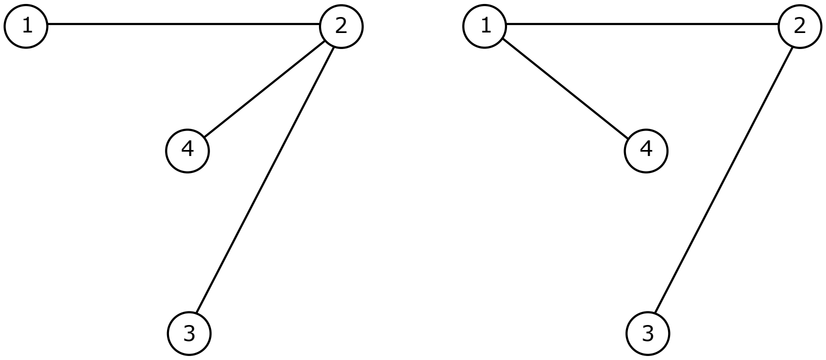

Figure 2: Two spanning trees for the complete graph with 4 nodes.

It turns out that all we need to do is take the $n \times n$ matrix with entries $|x_{ij}(X) - p_{ij}| - 1$ for $i \neq j$, set the diagonal terms so that each row (and therefore column) sums to zero, remove the first row and column, and calculate the determinant of the remaining $n - 1 \times n - 1$ matrix.

2 Example

Matrix of pairwise values p_{ij}, as in Figure 1.

```
   1    2    3    4
1  1   0.35 0.25 0.70
2 0.35  1   0.15 0.60
3 0.25 0.15  1   0.50
4 0.70 0.60 0.50  1
```

Matrix of values $x_{ij}(X)$ for the clustering X=12+34.

```
   1    2    3    4
1  1    1    0    0
2  1    1    0    0
3  0    0    1    1
4  0    0    1    1
```

Matrix of values $|x_{ij}(X) - p_{ij}| - 1$, with diagonal filled in.

```
    1     2     3     4
1  1.4  -0.35 -0.75 -0.30
2 -0.35  1.6  -0.85 -0.40
3 -0.75 -0.85  2.1  -0.50
4 -0.30 -0.40 -0.50  1.2
```

$f(X)$ is then the determinant of

```
  1.6  -0.85 -0.40
 -0.85  2.1  -0.50
 -0.40 -0.50  1.2
```

which is 2.089.

Again, this produces an unnormalised distribution. The value of the normalisation factor is constant for the analysis, but depends on the values of the p_{ij}. This is not a problem for a single analysis, but would be an issue if, for example, you wanted to compare two analyses with different p_{ij} values using MCMC methods like path sampling or stepping-stone. When the values of $f(X,T)$ for all 15 clusterings are found, and the results normalised, the distribution looks like this.

```
  X      f(X,T)
1234    0.031
1+234   0.052
2+134   0.077
3+124   0.111
4+123   0.015
12+34   0.061
13+24   0.061
14+23   0.06
12+3+4  0.061
13+2+4  0.05
14+2+3  0.119
34+1+2  0.081
24+1+3  0.099
23+1+4  0.041
1+2+3+4 0.081
```

From this, we can calculate new values for the p_{ij}.

```
   1    2    3    4
1  1   0.28 0.23 0.40
2 0.28  1   0.20 0.35
3 0.23 0.20  1   0.30
4 0.40 0.35 0.30  1
```

They are quite different from the originals, although they do have the same order from smallest to largest.

References

Chi Zhang, B Rannala, and Z Yang. Bayesian species delimitation can be robust to guide-tree inference errors. *Syst. Biol.*, 0:1–12, 2014. doi: 10.1093/sysbio/syu052.