USING THE ONCOTREE PACKAGE

ANIKO SZABO, KENNETH BOUCHER AND LISA PAPPAS

Abstract. This paper shows a short example of building and exploring oncogenetic trees using the Oncotree package. A detailed description of the theory of oncogenetic trees can be found in


A short introduction is given in doc/Oncotree.pdf.

We start by loading a dataset. The package contains an example dataset:

```r
library(Oncotree)
> data(ov.cgh)
> str(ov.cgh)
'data.frame': 87 obs. of 7 variables:
$ 8q+  : int 0 0 1 1 0 1 1 0 0 1 ... 
$ 3q+  : int 0 0 1 0 0 1 1 1 1 0 ...
$ 5q-  : int 0 0 1 0 0 1 1 0 1 ...
$ 4q-  : int 0 1 1 0 0 1 1 0 0 1 ...
$ 8p-  : int 0 0 0 0 0 1 1 0 0 1 ...
$ 1q+  : int 1 1 0 0 0 0 0 0 0 1 ...
$ Xp-  : int 0 0 0 0 0 0 1 0 1 1 ...
```

Based on these data, we construct the oncogenetic tree using the default $\ell_2$-distance error function to estimate the false-positive and false-negative error rates.

```r
> ov.tree <- oncotree.fit(ov.cgh)
> ov.tree
Oncogetic tree from 7 events
Parent function:
  8q+  <- Root
  3q+  <- 8q+
  5q-  <- Root
  4q-  <- 5q-
  8p-  <- 5q-
  1q+  <- Root
  Xp-  <- 8p-
Estimated error rates: epos= 0.2084556 , eneg= 0.0267696
```

The fitted tree can be examined several ways: printing it produces a quick summary, but the result of plotting is easier to interpret (the plots are shown in Figure 1).

```r
> ov.tree
```

We can compare the observed and fitted marginal occurrence frequencies of the mutations (the distance between these two was minimized for the error-rate estimation). The plot is shown in Figure 2.

```r
> print(obs <- colMeans(ov.tree$data))
   Root 8q+ 3q+ 5q- 4q- 8p- 1q+ Xp-
1.0000000 0.7011494 0.5517241 0.5287356 0.5057471 0.4712644 0.4367816 0.4252874
> print(est <- marginal.distr(ov.tree, with.errors=TRUE))
```

1
Figure 1. Fitted oncogenetic tree for the ov.cgh data set.

Figure 2. Observed and fitted frequencies of occurrence of each event.

In addition to the marginal frequencies, it is possible to estimate the entire joint distribution generated by the tree:

```
> dd <- distribution.oncotree(ov.tree, with.errors=TRUE)
> head(dd)
```

| Root 8q+ 3q+ 5q− 4q− 8p− 1q+ Xp− Prob |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 1     | 1     | 0     | 0     | 0     | 0     | 0     | 0     | 0.029222901 |
| 2     | 1     | 1     | 0     | 0     | 0     | 0     | 0     | 0.027992097 |
| 3     | 1     | 0     | 1     | 0     | 0     | 0     | 0     | 0.009202964 |
| 4     | 1     | 1     | 1     | 0     | 0     | 0     | 0     | 0.062160896 |
| 5     | 1     | 0     | 0     | 1     | 0     | 0     | 0     | 0.008323722 |
| 6     | 1     | 1     | 0     | 1     | 0     | 0     | 0     | 0.007973145 |
Using the overall joint distribution, it is straightforward to obtain marginal joint distributions (2- or higher way) if needed (the plot is shown in Figure 3).

\[
\begin{array}{cccccccc}
8q^+ & 3q^+ & 5q^- & 4q^- & 8p^- & 1q^+ & Xp^- \\
8q^+ & 0.7011494 & 0.4834457 & 0.3707227 & 0.3465855 & 0.3320996 & 0.3062492 & 0.3005693 \\
3q^+ & 0.4834457 & 0.5550202 & 0.2934589 & 0.2743523 & 0.2628855 & 0.2424226 & 0.2379265 \\
5q^- & 0.3707227 & 0.2934589 & 0.5287356 & 0.3884206 & 0.3683135 & 0.2309420 & 0.3245777 \\
4q^- & 0.3465855 & 0.2743523 & 0.3884206 & 0.4943105 & 0.3393380 & 0.2159057 & 0.2992688 \\
8p^- & 0.3320996 & 0.2628855 & 0.3683135 & 0.3393380 & 0.4736503 & 0.2068817 & 0.3130649 \\
1q^+ & 0.3062492 & 0.2424226 & 0.2309420 & 0.2309420 & 0.4367816 & 0.1872399 & 0.1872399 \\
Xp^- & 0.3005693 & 0.2379265 & 0.3245777 & 0.2992688 & 0.1872399 & 0.4286807 & 0.4286807 \\
\end{array}
\]

Another way to evaluate goodness-of-fit is through bootstrap resampling of the data. Two approaches are implemented: a parametric bootstrap that assumes that the model is correct and a non-parametric bootstrap. The plot is shown in Figure 4.
> set.seed(43636)
> ov.boot <- bootstrap.oncotree(ov.tree, type="nonparam", R=1000)
> ov.boot

Out of the 1000 replicates there are 309 unique trees with frequencies from 83 down to 1
   The bootstrap process found the original tree 83 times

> opar <- par(mfrow=c(3,2))  # the plot is in Figure 4
> plot(ov.boot, minfreq=45)
> par(opar)

The non-parametric bootstrap gives an estimate of the reconstruction confidence: the original tree was obtained
   83 times out of 1000 resamples, so the estimated confidence is 8.3%.

We can look at the frequency of edge occurrences in the bootstrapped trees:

> ov.boot$parent.freq

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
<th>Root</th>
<th>8q+</th>
<th>3q+</th>
<th>5q-</th>
<th>4q-</th>
<th>8p-</th>
<th>1q+</th>
<th>Xp-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Root</td>
<td>0</td>
<td>997</td>
<td>69</td>
<td>519</td>
<td>225</td>
<td>4</td>
<td>807</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>8q+</td>
<td>0</td>
<td>929</td>
<td>89</td>
<td>162</td>
<td>409</td>
<td>24</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3q+</td>
<td>2</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5q-</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>522</td>
<td>399</td>
<td>9</td>
<td>169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4q-</td>
<td>0</td>
<td>0</td>
<td>275</td>
<td>0</td>
<td>143</td>
<td>50</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8p-</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>84</td>
<td>0</td>
<td>4</td>
<td>599</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp-</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>45</td>
<td>12</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is clear that some edges are really stable: Root \rightarrow 8q+, 8q+ \rightarrow 3q+, root \rightarrow 1q+, all with confidence > 80%,
   while other edges are less stable (for example, 8p- is the child of 8q+ about as often as of 5q-).
Figure 4. The most frequently occurring bootstrap trees.