A goodness of fit test for *qpGraph*

Robert Maier, Pavel Flegontov, Nick Patterson

June 18, 2020

1 Theory

We describe our test for goodness of fit, and also show how our methodology can be used to compare two models which need not 'nest'.
We have two graphs G_1, G_2.

1. We compute blocks as we usually do for the Jackknife in ADMIXTOOLS. *The default block size is* 0.05 *(units Morgans). For large number of populations in a big graph we recommend using a smaller block size. Say blgsize: 0.005.*

   Randomly assign blocks to I (in sample) or O (out of sample) . Half the blocks are assigned to I.

2. Fit G_1, G_2 to the insample data I, getting fitted f-stats f_1, f_2. We also have the empirical stats f_I on I, f_O on O.

3. Let Q be the estimated error covariance of f_O. Scores S_1, S_2, S_I are now given by
$$S_1 = (f_1 - f_O)' Q^{-1} (f_1 - f_O)$$
   etc. We are interested in testing whether $S_1 < S_2$, significantly. If so this is evidence in favor of G_1. This can be tested with the block jackknife on O.
   Specifically we can test by computing $f_O^{[j]}$, deleting the j-th block and then calculating:
$$\begin{aligned} d &= S_1 - S_2 \\ d_j &= (f_1 - f_O^{[j]})' Q^{-1} (f_1 - f_O^{[j]}) - (f_2 - f_O^{[j]})' Q^{-1} (f_2 - f_O^{[j]}) \end{aligned}$$
   And then applying the weighted block (scalar) Jackknife to estimate a standard error for d.
   (As often in ADMIXTOOLS (for example *qpAdm*) we really ought to recalculate Q in the Jackknife, but this seems unnecessarily complicated without important gain. We calculate Q using all the data as a default. See *fullvar* below.)

Notice that we can test S_I and S_i in the same way, computing $S_I - S_i$. This gives a test for goodness of fit of the model G_i.
We unfortunately do not know the distribution of $S_I - S_1$ even under the hypothesis that G_1 is the true phylogeny that generated the data. However a large negative value and Z-score is strong evidence against G_1.

2 Implementation

This is now implemented in the current release. This is version 7300 and upwards. Another implementation will appear in an R package.
New parameters:

```
halfscore:  YES
## default NO -- get old behavior
halfjackname:  <fname>
## summary file for block jackknife.  Required for model comparison (see below)
```

It is strongly recommended that if *halfscore* is set the user also sets *seed*.

```
seed: <nonzero integer>
```

otherwise it will be impossible to compare *halfscore* on two different models.

2.1 Goodness of fit score

We compare on the outsample blocks (not used to fit the model):

a Fitted model

b The f-statistics from the insample blocks.

We argue that if the model has the true topology, the fitted f-statistics should have lower noise than the empirical.
See

```
Zfit:
```

in the log file (*stdout*). A large negative score indicates a failure to fit.

2.2 Comparing two models

Suppose we fit *grapha* and *graphb* outputting jackknife summaries (*halfjackname*) *jacka* and *jackb*. Now run

```
jackdiff -a jacka -b jackb [-o jackdiff] ## jackdiff is a Perl script
## some scripts in .../perlsrc may need modifying to find Perl path on your system.
```

A large Z-score is strong evidence to prefer *grapha* over *graphb*. $-o$ flag is optional but if present outputs a jackknife summary file for model difference scores.

2.2.1 fullvar

For this half scoring to work well it is important to get good estimates of the f-statistic covariance matrix. By default we use all the data to estimate the covariance. If this is not desired code

```
fullvar: YES.
```

when just the out of sample blocks are used.

2.3 Caveat

This implementation is new and only minimally tested. Both the code and theory are experimental and I would welcome feedback.

References

[1] N. Patterson, D. C. Petersen, R. E. van der Ross, H. Sudoyo, R. H. Glashoff, S. Marzuki, D. Reich, and V. M. Hayes. Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.*, 19:411–419, Feb 2010.