Documentation for *qpWave* and *qpAdm*

Nick Patterson

June 3, 2016

1 Introduction

We document 2 programs: *qpWave* and *qpAdm* based on a common set of ideas related to f_4 statistics [**?**] The first program, *qpWave* (formerly *qp4wave2*) emerged from work with David Reich on the peopling of the Americas [**?**]. The second, *qpAdm* is more recent, and is an attempt to systematize ideas of Iosif Lazaridis, using f_4 statistics in a regression context, but also incorporating methods from *qpWave* .

In [**?**, S6] we showed that if we took a set of a *left* populations U and a set of b *right* populations V and considered the matrix

$$X(u, v) = F_4(u_0, u; v_0, v)$$

where u_0, v_0 are some fixed populations of U and V, and l, r range over all choices of populations of U, V. We can assume that $u \neq u_0, v \neq v_0$, so that the matrix X is $(a-1) \times (b-1)$. We then showed that if X had rank r and there had been n waves of immigration from V to U with no back-migration from U to V, then:

$$r + 1 \leq n$$

In our initial application we used this to show that there must have been at least 3 waves of immigration into the (pre-Columbus) Americas.

2 Algorithmic details for *qpWave*

We describe our computational strategy in a little more detail. We compute $\hat{X}$, an estimate of X so that in the notation of [**?**]

$$\hat{X}(u, v) = f_4(u_0, u; v_0, v)$$

We can use the block jackknife [**?**] to compute V an estimate of the error covariance of X. To test if $\hat{X}$ has rank r we write

$$\hat{X} = A.B + E$$

where A is $(a-1) \times r$, B is $(e \times (b-1)$ and E is a matrix of residuals. The (log) likelihood for (A, B) and implicitly r is:

$$\mathcal{L}(A, B) = -\frac{1}{2} \sum_{i,j,k,l} V^{-1}_{i,j;k,l} E_{i,j} E_{k,l}$$

where the residual matrix E is defined by

$$E = \hat{X} - A.B$$

For each r we set A, B initially by an SVD analysis of X, and then iterate, minimizing $\mathcal{L}$ with respect to A, B in turn. For fixed A, $\mathcal{L}(A, B)$ is quadratic in B and can be minimized by solving linear equations. Since A, B only enter into the likelihood though a matrix product, we see that

$$A.B = (A.Q).(Q^{-1}B$$

for any non-singular $r \times r$ matrix Q. Thus the number of degrees of freedom is

$$d((r) = ((a-1) + (b-1))r - r * r = r(a + b - (r+2))$$

As a check, if r is the maximal rank $Min(a-1, b-1)$, then $d(r) = (a-1)(b-1)$ which is obviously correct. This is the *saturated* model, where we fit the data perfectly.

We compute statistics with a likelihood ratio test.

3 Parameters and output of *qpWave*

Here is a sample parameter file.

```
DIR:     /home/np29/broaddata/bl14
S1:              honjp
indivname:       DIR/S1.ind
snpname:         DIR/S1.snp
genotypename:    DIR/S1.geno
badsnpname:      ./cpgmf
popleft:  pleft
popright: pright
maxrank: 4
## not needed here
```

The top lines are parameters that will likely be familiar, for example they are the same in em convertf, qpDstat, qp3Pop. In this run I did not want to use CpG sites, which are removed by the badsnpname: line. pleft is a file of populations 1/line, pright also. We have
pleft:

```
WHG
LBKNeolithic
YamnayaEBA
```

while the right population list was:
pright:

```
Han
Eskimo
Mbuti
Karitiana
Kharia
Onge
Ulchi
```

(the right set of populations are chosen so that they are differently related to West Eurasia). Extracts from the output:

```
f4rank: 0 dof: 12 chisq: 330.440 tail:  1.86337038e-63
 dofdiff:  0 chisqdiff:  0.000 taildiff:               1
f4rank: 1 dof:  5 chisq:  46.979 tail:  5.73674279e-09
 dofdiff:  7 chisqdiff: 83.460 taildiff:   2.05163995e-57
f4rank: 2 dof:  0 chisq:   0.000 tail:              1
 dofdiff:  5 chisqdiff: 46.979 taildiff:   5.73674279e-09
```

For each line we rank a χ^2 statistic and tail area (chisq and tail) comparing with the saturated model, and also a chi-square statistic and tail for the model with one rank less. We see here that the rank 1 model has a $p-value$ of 5.7×10^{-9}, comparing with the saturated model and can be rejected. We have very strong evidence here that *WHG, LBKNeolithic, YamnayaEBA* are not the product of 2 waves from outside West Eurasia. o

The matrices A, B are published and may be useful.

```
B:
          scale     1.000     1.000
         Eskimo     1.323    -0.011
          Mbuti    -0.306     2.300
      Karitiana     1.995     0.262
         Kharia     0.190     0.592
           Onge    -0.206    -0.532
          Ulchi     0.308    -0.082
A:
          scale  1392.604  1651.101
   LBKNeolithic    -0.533     1.310
     YamnayaEBA     1.310     0.533
```

We show here A, B matrices for the saturated model. (Actually we show $transpos(B)$, with a scale factor for the columns. From the second column

Mbuti has a large coefficient, and *LBKNeolithic* also. From the first column we see Karitiana and YamnayaEBA. It is therefore not surprising to see from *qpDstat* output

```
                                      D        Z
WHG  LBKNeolithic  Han      Mbuti   0.0208   7.780
WHG    YamnayaEBA  Han  Karitiana   0.0239   7.627
WHG    YamnayaEBA  Han      Mbuti   0.0053   1.765
```

with the first 2 Z scores large, the last much smaller.

We note that the χ^2 statistics here, using the LRT are computed using a fixed covariance V. It would be formally more correct to reestimate V, simultaneously with A, B. This would greatly increase complexity, without adding much precision.

I strongly recommend attempting to keep the population lists here small. If a, b are large then the covariance V is a big matrix, and in practice will be estimated poorly. This can be expected to lead to trouble.

4 Finding mixture coefficients — *qpAdm*

We next describe a novel idea for finding admixture weights using f_4 statistics. This was motivated by work of Iosif Lazaridis, though the details here are quite distinct. Let T be a *target* population, $S = \{s_1, s_2, \ldots s_n\}$ a set of source populations. In the easiest case to consider, when T is an an admixture of populations of S we can write symbolically

$$T = \sum_{i=1}^{n} w_i s_i$$

It then follows that for any populations r_1, r_2

$$\begin{aligned} \sum_i w_i f_4(T, s_i, r_1, r_2) &= f_4(T, T, r_1, r_2) \\ &= 0 \end{aligned}$$

A little thought shows that this is true even if the populations s_i are descendents of the true source populations, provided that there has been no gen flow between the most recent ancestor of T, S on the one hand and ancestor of r_1, r_2 on the other. [In passing, we note that we used f_3 statistics in [?] to derive mixing coefficients for modern admixture. The methods there require samples of the actual source and admixed populations, but do not require outgroup populations as we do here with the r_i.]

Thus, if T is admixed, as above, pick a set of outgroup populations R, and

1. Check, using *qpWave* , setting left populations $L = S$, and right populations R that the matrix X has full rank $n - 1$.
2. Check, again using *qpWave* , that letting $L = \{T, S\}$ the re is no strong evidence that the rank of X increases.

We now will take T as the base population of $L = \{T, S\}$, which simplifies the algebra. We calculate matrices A, B as in *qpWave* , with the rank set to $n - 2$ (corank 1). So the recovered A is of dimensions $(n-1) \times (n-2)$. It then follows that estimates $\mathbf{w} = (\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_n})$ of the admixture weights can be found by solving the equations:

$$\begin{aligned} \mathbf{w.A} &= \mathbf{0} \\ \sum_{i=1}^{n} w_i = 1 \end{aligned}$$

We can use the block jackknife to compute a covariance matrix for the errors. [Formally, we should reestimate V as we delete blocks in the jackknife. This is not presently done, as it would add complexity and seems unlikely to make a material difference.]

4.1 All subsets regression

Suppose U is a proper) subset of S. It is interesting to require that $w_i = 0$ if $s_i \in U$. That is populations of U do not contribute to the admixture of T. This constrains the structure of the matrix A but optimization is still easy to carry through. It can be shown that if $|U| = f$, then the saturated model has $(b - a) + f + 1$ degrees of freedom. Since in practice n will be small, it is computationally reasonable to try all proper subsets of S; for each we can compute the best coefficients and a chi-square score using an LRT.

Here is a sample parameter file.

```
DIR:     /home/np29/broaddata/bl14
S1:              honjp
indivname:       DIR/S1.ind
snpname:         DIR/S1.snp
genotypename:    DIR/S1.geno
badsnpname:      ./cpgmf
popleft:  pleftx
popright: pright
maxrank: 4
## not needed here
```

The format of the parameter file is identical to that for *qpWave* . qpop1 is a file of populations 1/line, superpops also. We have
pleftx:

```
CordedWareNeolithic
WHG
LBKNeolithic
YamnayaEBA
```

while the right population list was the same as *pright* described in the section on parameters of *qpWave* .

BY convention the first population of the left list is the target. So here we are examining *CordedWareNeolithic* as a mixture of the other three populations.

Extracts from the output: We begin by testing using *qpWave* methodology whether a rank 2 matrix can be accepted. Here we get a p-value of 0.07 and proceed.

```
f4rank: 2 dof: 4 chisq: 8.647 tail: 0.0705644793
 dofdiff: 6 chisqdiff: -8.647 taildiff:             1
f4rank: 3 dof: 0 chisq: 0.000 tail:  i         1
 dofdiff: 4 chisqdiff:  8.647 taildiff:  0.0705644793
```

Next we give the mixture coefficients and standard errors, which are typically far from independent. Then an error covariance matrix, computed with the block jackknife.

```
best coefficients:     0.322     0.053     0.625
      std. errors:     0.177     0.114     0.099

error covariance (* 1000000)
     31255     -17297     -13957
    -17297      13044       4253
    -13957       4253       9704
```

We finally give an 'all subsets analysis' where the coefficient under a '1' is forced zero.

```
    fixed pat      dof     chisq       tail prob
          000 0      4     5.833        0.211948     0.322     0.053     0.625
          001 1      5    30.207               0     1.365    -0.365     0.000  infeasible
          010 1      5     6.101        0.296519     0.386     0.000     0.614
          100 1      5     9.903        0.078038     0.000     0.226     0.774
          011 2      6    37.560     1.36904e-06     1.000    -0.000     0.000
          101 2      6   158.115               0     0.000     1.000     0.000
          110 2      6    22.327      0.00105618     0.000    -0.000     1.000
```

Here we see that, at least in this analysis there are reasonable models with CordedWareNeolithic is a mix of either WHG or LBKNeolithic and YamnayaEBA.

This is unsurprising, given the standard errors above. The point of this note is not to give a serious phylogenetic analysis but the results here certainly support a major Steppe contribution to the Corded Ware population, which is entirely concordant with the archaeology [**?**].