A method for approximate PCA projection when there is much missing data

Nick Patterson

May 11, 2013

A problem that sometime arises is that we calculate principal components in the usual way using samples with nearly complete genotype data. We than have an additional sample $\mathbf{s}$ (or many such) which we wish to 'project' onto the PCA axes.

One situation in which this occurs is that $\mathbf{s}$ is admixed and we wish to analyze only SNPs where we believe that the sample is homozygous for one ancestry. For instance in [1] sections of European ancestry in Native Americans were masked out as the primary interest was in indigenous genetics.

Another situation is analysis of ancient DNA where poor preservation may mean that coverage is poor, with many SNPs having missing data.

In the default mode of *smartpca* we construct eigenvectors $\mathbf{e_1}, \mathbf{e_2}, \ldots \mathbf{e_n}$ which form an orthonormal set, and then calculate coordinates $c_1, c_2, \ldots c_n$ of s by filling in all missing data of $\mathbf{s}$ and simply computing

$$c_i = \mathbf{s}.\mathbf{e_i} \tag{1}$$

If $\mathbf{s}$ has a great deal of missing data, the default algorithm will perform poorly, as the missing data is filled in with an average of the allele frequencies in the base populations used for the PCA. This will not usually be sensible, Skoglund [2] found a solution using Procrustes analysis (see also [3]. We implemented a different simpler solution. Note that the c_i of equation (1) are the solution that minimizes

$$||\mathbf{s} - \sum_{\mathbf{i}} \mathbf{c_i}\mathbf{e_i}||^2$$

It is natural to consider the same problem, where we only examine genotypes where $\mathbf{s}$ is not missing. Thus we set c_i to minimize

$$\sum_{j \in X} s_j - \sum_i c_i e_{i,j}{}^2$$

where X is the set of SNPs where $\mathbf{s}$ has valid data. This is a simple least squares problem, reducing to the default if $\mathbf{s}$ has no missing data. To use this feature code

```
lsqproject: YES
```

in the *smartpca* parameter file.

References

[1] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, et al. Reconstructing Native American population history. *Nature*, 488(7411):370–374, Aug 2012.

[2] P. Skoglund, H. Malmstrom, M. Raghavan, J. Stora, P. Hall, E. Willerslev, M. T. Gilbert, A. Gotherstrom, and M. Jakobsson. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, 336(6080):466–469, Apr 2012.

[3] C. Wang, Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol*, 9:Article 13, 2010.