PCFAM Package Example (Version 1.0)

Yi-Hui Zhou

August 26, 2024

1 The problem

The issue of robustness to family relationships in computing genotype ancestry scores such as eigenvector projections has received increased attention in genetic association, and is particularly challenging when sets of both unrelated individuals and closely-related family members are included.

We consider two main novel strategies: (i) matrix substitution based on decomposition of a target family-orthogonalized covariance matrix, and (ii) using family-averaged data to obtain loadings. We illustrate the performance via simulations, including resampling from 1000 Genomes Project data, and analysis of a cystic fibrosis dataset. The matrix substitution approach has similar performance to the current standard, but is simple and uses only a genotype covariance matrix, while the family-average method shows superior performance. Our approaches are accompanied by novel ancillary approaches that provide considerable insight, including individual-specific eigenvalue scree plots.

2 Example

In this document, we give an example to run `PCFAM` package in R. The first step is to load the example genotype data by the following R statements. Users can download the example genotype data from `http://www4.ncsu.edu/~yzhou19/`

```
originalX=geno.combined.sibs
```

The second step is to row scale and residualize the original genotype data.

```
X=rowscale(originalX)
Xresid=residualize(X)
#Xresid is the residualized version of X,
# after removing effects of larger-scale ancestry.
```

Then we calculate a correlation matrix of residualized X, used as input for findfamilies

```
corXresid=cor(Xresid)
```

The next intermediate step is important to robustly indentify family members, but is not otherwise used in the ancestry score calculation. `myfam` is the family IDs identified. User can use output from other software such as KING, alternately. Singletons are labeled family=0, so if F is the number of families, we expect `F+1` unique `myfam` IDs.

```
myfam=findfamilies(corXresid,0.1)
length(unique(myfam))
#Number of ancestry scores to use.
K=6
```

Now we can apply matrix substitution approach and family average approach.

```
myms.pca=ms.pca(X,corXresid,0.1,K)
myeigenvectors=myms.pca$eigenvector
#Ancestry scores from the family average approach
familyave.result=familyave(X,myfam,top=K)
```