

These instructions for Spectral-GEM correspond to the manuscript “Discovering Genetic Ancestry Using Spectral Graph Theory” (Lee, Luca, Klei, Devlin, Roeder). This is an improvement on GEM; see “On the Use of General Control Samples for Genome-Wide Association Studies: genetic matching highlights causal variants”, by Luca et al. *Amer J Human Genet.* 2008.

NOTE: Spectral-GEM is designed to find the ancestry vectors and match cases and controls for association analysis. However, it can be used for many other purposes. We list a few here to give insight into the flexibility of the program.

- Spectral-GEM allows for the identification of genetic outliers
- The eigenvectors produced by Spectral-GEM can be used for visualizing ancestry and to determine sub-populations in the data.
- The eigenvectors can be used as proxies for ancestry and be used to regress out ancestry effects in association analysis
- The clusters produced by Spectral-GEM can be used to account for in ancestry in association analysis using the MHC test
- The matches created by Spectral-GEM can be used in association analysis using the conditional logistic model.
- Another use of this output information is to determine populations for fine-mapping.
- Other applications that require a good grasp on ancestry might benefit from the output of Spectral-GEM as well.

HINTS

- Choosing the best number of dimensions in eigenanalysis is a notoriously challenging problem. Spectral-GEM chooses the number of significant dimensions based on an eigengap heuristic. This heuristic is far from perfect. Although it often works well, we have observed it to fail in practice for no apparent reason. When the cutoff fails to choose a sensible number of dimensions, the program usually hits the default maximum of 50 dimensions. To circumvent this problem it is necessary to trick the program into using a more desirable cutoff. The program chooses d , the number of non-trivial eigenvectors, by finding

$$d = \max\{i : |\lambda_i - \lambda_{i-1}| > -0.00016 + 2.7/n + 2.3/p\} - 1,$$

where n is the number of subjects and p is the number of SNPs. Because p is an input parameter in the program that is used solely for the purpose of finding this critical value, it can be manipulated. To obtain a lower dimensional eigenvector representation, input a smaller p . The eigenvalues are output as part of the analysis. If the program is defaulting to 50 dimensions, plot $|\lambda_i - \lambda_{i-1}|$ versus i to see what choice of p is required to obtain the desired number of dimensions.

WARNINGS. Eigenanalysis is highly sensitive to unexpected correlation among subjects and markers, as well as quality control issues. Problems with the data usually lead to a eigen-decomposition with 10 or more significant dimensions. If the problem remains stubbornly high dimensional even after removal of outliers and unmatched observations, quality control issues need to be re-examined. Here we provide several suggestions for what to look for if the data can not be simplified to a lower dimension.

1. Inclusion of close relatives can result in the discovery of numerous spurious dimensions of ancestry.
2. Inclusion of correlated SNPs can lead to discovery of spurious dimensions of ancestry. Just because GWA yields several hundred thousand SNPs is no reason to include them all in the Spectral-GEM analysis. We have found that the “effective number of independent” SNPs from a 500K genome scan is approximately 50,000. Including more SNPs often leads to the discovery of meaningless eivenvectors.
3. High missingness in your tag SNPs can result in the discovery of spurious dimensions of ancestry. We suggest that tag SNPs be limited to those SNPs with no more than 0.5% missingness.
4. Poor quality SNPs, such as those out of Hardy-Weinberg equilibrium, should be removed. Some of these SNPs are likely to be ancestry informative and would be useful to the analysis; however, many more are likely to be the products of genotyping errors. Inclusion of SNPs with substantial numbers of genotyping errors will have a detrimental impact on the eigen-analysis.
5. Tag SNPs, missingness and Hardy Weinberg violations may differ from ethnicity to ethnicity. QC should be done within groups.

1. Preprocessing:

- (a) Perform usual quality control checks for the data.
- (b) Remove any twins, parent-offspring pairs, full siblings or duplicates.
- (c) From the full set of SNPs measured, select p SNPs that are approximately independent. We use the H-clust software, posted at this site, to select tag SNPs that are nearly independent (threshold $r^2 = 0.04$). We suggest choosing tag SNPs based on the control subjects. For a GWA study p should be about 20,000-50,000. Note: SNPs with a high missing rate should not be used as tag SNPs.
- (d) Create an $n \times p$ matrix consisting of allele counts for n subjects at these p SNPs. Entries in the matrix will be 0, 1, 2 or missing.
- (e) Center and scale each column by subtracting the mean and dividing by the standard deviation. Call this matrix M .
- (f) Replace missing values in M by “0” entries, or impute them using any of the popular methods available on the web.

- (g) Create an $n \times n$ matrix by multiplying: MM^t and dividing by p . This matrix summarizes the data that go into the Spectral-GEM program. (Spectral-GEM computes the weight matrix described in Lee et al. internally.) A program that creates the matrix is available on the main software page.

http://wpicr.wpic.pitt.edu/WPICCompGen/MMp/MMp_page.htm.

2. Format of the matrix file:

- (a) This file can be space or comma delimited.
- (b) The first row gives n , the number of individuals in the analysis.
- (c) The second row gives p , the number of tag SNPs.
- (d) Starting with the third row, each row of the matrix corresponds to a subject. The first 4 columns of information are
 - i. Column 1: subject identifier (alpha-numeric).
 - ii. Column 2: Sex (numeric).
 - iii. Column 3: Affection status (case = 2, control = 1).
 - iv. Column 4: Group label (a user chosen numeric variable such as site of collection or disease subtype). This can be a copy of column 3.

The remaining columns of each row are the corresponding row of MM' . A fortran program to create this matrix will be provided upon request.

3. Input parameters:

- (a) line 1: run label (alpha-numerical code of exactly length 5 to use as an extension to label output). The fifth position will be updated at each stage of the analysis to differentiate output from each stage. It is natural to use a run label with a "1" in the fifth position initially to identify the output from stage 1.
- (b) line 2: directory where the data are stored, enclosed in double quotations
- (c) line 3: name of file containing data input, enclosed in double quotations
- (d) line 4: name of file containing excluded subjects, enclosed in double quotations. (This file will be expanded after running the program. It does not matter whether the file exists or not. So if initially no such file exist, just put a file name here and pretend it exists.)
- (e) line 5: maximum number of characters in the individuals identifier (maximum=20); i.e., the exact length of the longest identifier.
- (f) line 6: Cluster minimum size. Suggested value is 10. (See paper for details about the clustering process.)
- (g) line 7: amount of output created (0=limited, 2=lots); we suggest "0"

4. Output files

- (a) XXXXX is replaced by the user chosen file extension.
 - (b) log files: GEM_log_XXXXX.txt. These files record the output to screen for each run of the .C(“main”).
 - (c) distance files: distance_XXXXX.matrix. These files are used to examine the case-control distances and control-case distances and determine the outliers for the exclusion file.
 - (d) significant eigenvectors: significant_eigenvector_XXXXX.txt. These files are the significant eigenvectors. Plots of these eigenvectors reside in scatter_XXXXX.pdf
 - (e) clusters: clusters_XXXXX.txt. Plots of these clusters reside in clusters_XXXXX.pdf
 - (f) exclude file: this file expands as the program proceeds through the examination of the case-control and control-case distances. The file name is chosen by the user before the program starts. If there is already such a file then one uses that file name and the program excludes any individuals listed in the original file. If the file does not exist, chose an arbitrary file name.
5. Spectral-GEM moves iteratively between R and commands that the user cuts and pastes or enters in an R command window and a Fortran subroutine that does the bulk of the calculations. The R commands are found in the commandXXX.txt that is distributed with the program. There is a different command file for Windows and Linux. The Eigenvalue decomposition (EVD) portion of the data analysis is performed in multiple stages. The objective in the first stage is to determine outliers and/or unmatchables. The user can repeat this step if s/he desires to hunt for additional outliers. Keep in mind that the dimension and EVD of the problem changes after removal of individuals. Our experience with Spectral-GEM is that one round of outlier removal is sufficient for matching cases and controls. The last stage and, many times the second one, creates the files that are needed by the R matching algorithm fullmatch to match cases and controls based on optimal genetic distances.

Each time the Fortran subroutine is called (using the command .C(“main”)) it returns with one or more questions about which tasks should be performed. Comments in the command file indicate the typical answers, assuming no tasks are to be repeated. The program is flexible, and different responses permit a repeat of particular steps.

(a) Outliers/clustering step:

- i. Estimate D , the number of significant dimensions based on the eigengap heuristic.
- ii. Calculate the distance between each case and the nearest control, and vice versa, using the D dimensional eigen space.
- iii. You must choose whether to use simulations (S) or the eyeball (E) method to detect observations for which the case is too far from the nearest control

or vice versa. The simulations approach is better, but for a large data set it can be slow.

- iv. If you chose “E”, examine the pair of histograms depicting the case distances and the control distances. These distributions will be strongly skewed to the right, with some unmatchable observations appearing as extreme outliers. Based on these figures, input a cut off value that (i) removes the obvious outliers from both histograms and (ii) approximately equalizes the two distributions. We call these outliers “unmatchables”.
 - v. Repeat this stage until the pair of distributions look similar. After successfully trimming the distributions they will continue to be skewed, but the tails should become similar in shape and not have any big gaps between observations.
- (b) Matching step:
- i. Create matched strata using the optmatch algorithm.
 - ii. Output strata number in a file called fullmatch_XXXXX.txt.
6. Analyze the data using conditional logistic regression.
- (a) You will need to use the library ”survival” in R.
 - (b) To install a library in R under windows system, go to R menu and click on Packages— >Install package(s) and go from there.