Ranked data cluster analyzation

Brendan Mikolajczyk, Eric Mustee, Dan Skrodzki



User’s Guide

Table Of Contents

Introduction ..............................................................................1

Startup …………………………………………………………….…2

Loading a File …………………………………………..…..2

Formatting Guidelines …………………………………...2

Opening a Saved Session ………………………….……….3

Running the Program..……………… …………………….………….…..……3

Cluster Analyzer……… …….……………………………..4

Additional Settings 4

Data Description ………………...…………………………4

Random Data Generation 5

Interpreting the Results ……………………………………………#

Saving/Exporting #

Glossary …………………………………………………………....#

Introduction

What is a ranking?

Rankings are a relationship between a set of objects such that one object is better, or ranked higher than another. For instance, a list of one's favorite ice cream cones is a ranking.

What does this software do?

This program clusters rankings. This means finding rankings that are similar enough, and grouping them together in a way that makes the most sense. The program computes what are called cluster centers from a set of partial rankings.

Startup

Loading a File

Before running a cluster analysis, a ranked data file must first be loaded in. To load in a file, navigate to the menu bar and select File>Import Rankings> and then browse your computer for an acceptable file type. Only **.txt** files and **.csv** files can be loaded in.

Formatting Guidelines

Acceptable ranked data files must adhere to the following rules:

* File must not contain any non-numeric characters
* Rankings cannot include the number 0 (nonzero integers only)
* Each line represents a pi vector, so no line may contain more than one of the same number
* Each line must stick to one style of appropriate delimiters
  + Appropriate delimiters include:
  + Spaces
  + Commas not followed by spaces
  + Commas followed by spaces

Example of an appropriately formatted file:

1, 2, 3, 4, 5

5,3,-4,1,2

-2 -4 5 3 1

3 5 4

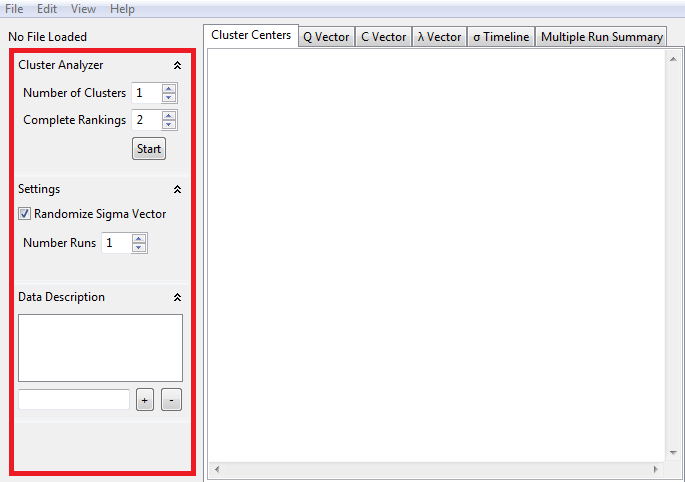
-1 5 4 2

Opening a Saved Session

This program also allows you to save current settings and results into a separate **.rnkr** file. To restore a previously saved session, navigate to the menu bar and select File>Open Session> and then browse your computer for the desired **.rnkr** save file. Once this file is loaded in, all settings and results will be restored from that session.

Running the Program

Before analyzing the input, there are various settings which can be adjusted in the left-hand sidebar of the program window.



Cluster Analyzer

There are two settings which can be adjusted under the Cluster Analyzer settings:

**Number of clusters** – adjust the number of expected clusters the program should analyze with. In turn, this also changes the size of the sigma vector.

**Complete rankings** – this setting is automatically adjusted when a ranked data file is loaded. This can be adjusted to represent the expected number of elements in a complete ranking; however, it cannot be decreased lower than the default value that was already automatically set.

**Start** – click this button once a pi vector has been loaded in and all settings have been appropriately adjusted to begin the analysis of the ranked data.

Additional Settings

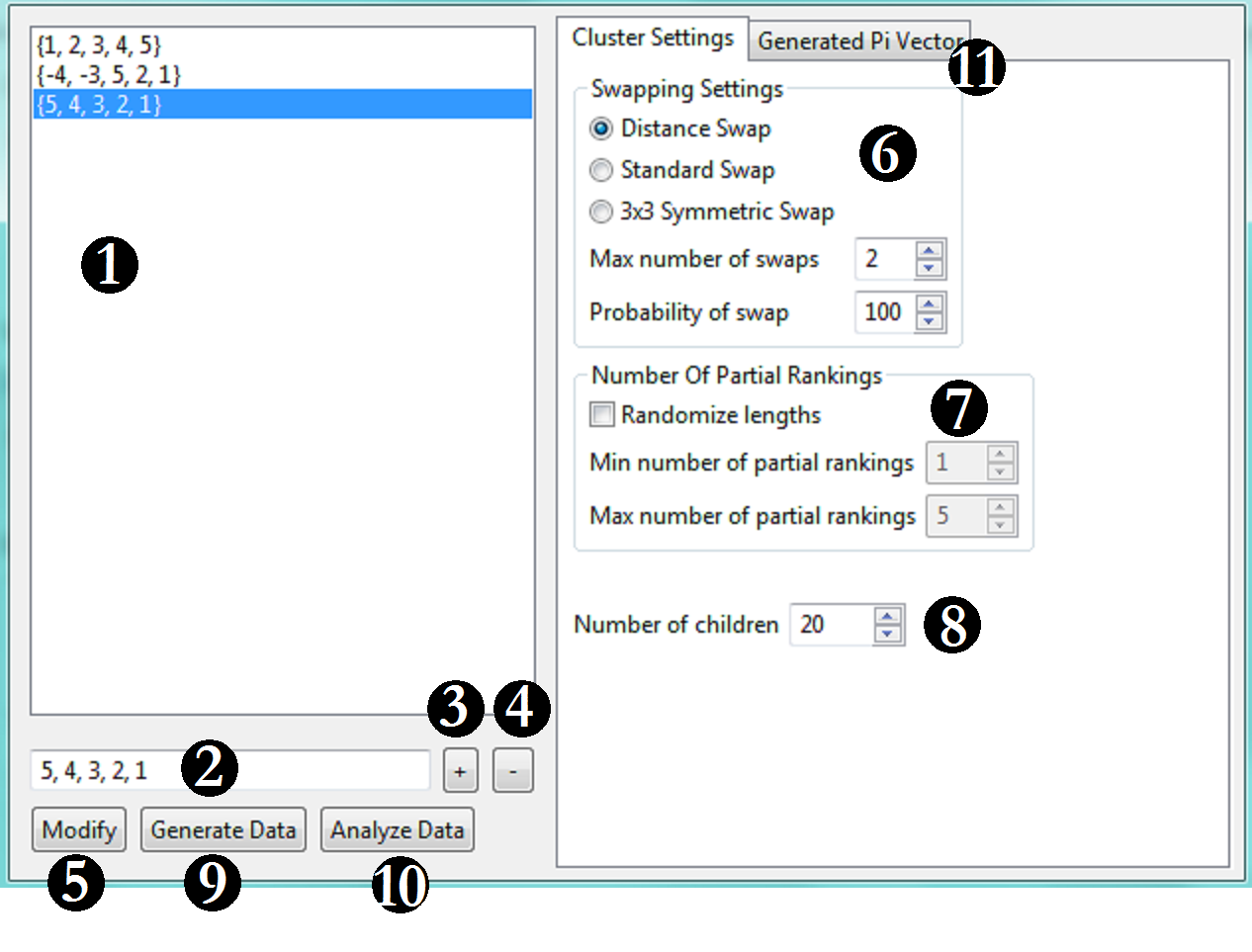
Another setting that can be adjusted is the randomization of the sigma vector. Checking the randomize sigma vector checkbox will generate random sigmas in the sigma vector, thus the analysis will start with completely random starting locations for the initial cluster centers. If the checkbox is not checked, then the initial sigma values will only be arbitrarily chosen if there is two or less cluster centers. Their values will be {1, 2, 3, …, n} for the first cluster and {n, …, 3, 2, 1} for the second cluster.   
  
Additionally, the number of runs that the analyzer performs in one analysis can be adjusted. The number of runs can be increased to get a better idea for what the ideal cluster centers are. All data from each run can be viewed in the Multiple Run Summary Tab.

Data Description

Another feature is the option to give concrete meaning to the numbers in each ranking. In the text field under Data Description, you can type in a word and click the “**+**” button to add it to the existing word list. Click the “**–**” button to remove a word from the list. The order of the words represents which number is replaced by which word, so the first word replaces all 1’s and -1’s, the second word replaces all 2’s and -2’s, etc. An example of a word list could be favorite ice cream flavors: chocolate, vanilla, and strawberry. So the ranking {2, 1, 3} would be replaced by {vanilla, chocolate, strawberry}, meaning that vanilla is the highest ranked ice cream flavor.

Random Data Generation

If you don’t have a ranked data file to load in, you have the option to create a randomly generated ranked data file that can also be customized to fit certain needs. To access the random data generator, navigate to the menu bar and select View>Random Data Generation> and then from there, a window will pop up with the random data generation settings.



Above, we can see options for randomly generating data. The left side of the screen shows each individual cluster center, and the right side of the screen shows data for scattering the members for a single selected cluster center.

To add a cluster center, type the ranking into the textbox located at (**2**), and then press the + button (**3**), To remove a ranking, select the ranking in the listbox (**1**), and press the – (**4**). To change a ranking, select the ranking in the listbox (**1**), type the new desired ranking into the textbox (**2**), and press the modify button (**5**). To change the settings of a given cluster center, select the cluster center in the listbox at (**1**), and select the desired settings on the right half of the screen.

The settings that may be modified are swapping settings (**6**), partial ranking settings (**7**), and the number of children (**8**).

In the swapping settings (**6**), the max number of swaps determines up to how many times a swap can occur. The probability of swap determines how likely a swap will occur on a scale of 0 (not at all) to 100 (will happen every time guaranteed). A distance swap will choose one of the elements in the ranking, and move it by 1 each time a swap occurs. If the last number is swapped, it is made negative. The standard swap chooses two elements to be swapped each time at random, and swaps them each time a swap occurs. The 3x3 symmetric swap works a bit differently than the two previous swaps in that it doesn’t use a max number of swaps or probability of swap. The 3x3 symmetric swap requires a ranking with 9 elements to work. It will generate 9 sets of unique rankings: 4 rankings will be two steps away from the original cluster, 4 sets of two cloned rankings will be one away from the cluster, and the original cluster will appear 8 times. This leads to a total of 20 rankings.

In the partial ranking settings (**7**), one can generate incomplete rankings. To use this feature, check the box Randomize lengths. Then, select the smallest partial ranking you want to appear, and select the largest partial ranking you want to appear. These numbers affect how many elements are in a ranking before it gets cut off.

The final option, number of children (**8**), determines how many mock rankings will be associated with a given cluster center. When using the 3x3 cluster swap, the option number of children functions slightly differently. Since this method of swapping requires 20 children, modifying number of children changes the multiple of how many children occur. If number of children is set to 3 for a 3x3 symmetric swap, 60 children will be generated.

To generate the data, select the Generate Data button (**9**). This will switch the tab over the Generated Pi Vector tab, and you will see all the rankings that were generated. If you are satisfied with the randomly generated ranking, select the Analyze Data button (**10**). Otherwise, regenerate the data (**9**) or change the tab back to Cluster Settings (**11**) to modify the settings.

Interpreting the Results

Once the analysis has finished, all results can be viewed in the six tabs located near at the top of main text field.

**Cluster Centers** – displays the cluster center for each cluster of ranked data, starting with σ0 and going up from there.

**Q Vector** – displays the *q* values for every ranking in the π vector. Each ranking, displayed as π*n*­, will have a *q* value for every cluster in the analysis, representing how closely the ranking is associated with each cluster.

**C Vector** – displays the *c* values (cluster weights) for each cluster, starting with c0 and going up from there.

**λ Vector** – displays the dispersion parameter value for each cluster, starting with λ0 and going up from there.

**σ Timeline** – displays how each cluster center, σ*n*, has changed throughout each step in the analysis.

**Multiple Run Summary** – allows the user to see the summary of each run. The recorded data for each run includes the σ timeline for each run and the distance between the starting cluster centers and the supposed cluster centers possibly specified in the Random Data Generator, and the final clusters and the supposed clusters.

Saving/Exporting

**Saving.**  
To save all results and settings from the current analysis, navigate to the menu bar and select File>Save Session> and then select a location on your computer to save your file. Once saved, your work will be saved into a **.rnkr** file. This file can later be accessed to open and restore that session with the same settings and results.

**Exporting Results.**To export all results from each tab in the current analysis, navigate to the menu bar and select File>Export Results> and then select a location on your computer to save your file. Exporting your results will create a **.txt** file containing all information you would normally see in each of the six tabs.

Glossary

**cluster** – group of ranked data that is closely associated with a particular center (σ).

**cluster center (σ)** – a single ranked data element that represents the center of a cluster, which may or may not be an actual element of that cluster.

**complete ranked data** – an ordered set of numbers from 1 to *n*, that represent a particular person’s preferences, where the order in which these numbers determine the ranking of these preferences (first element is top ranked and last element is lowest ranked). These numbers can also be negated to represent an opposition of a particular preference.

**c vector** – list containing cluster weights for the current clusters in the sigma vector.

**data description** – words replacing numbers in ranked data sets to provide concrete meaning to the rankings.

**dispersion parameter** **(λ)** – represents the standard deviation of a particular cluster; in other words, the average distance of every ranked data element to the cluster center.

**partially ranked data** – an incomplete ranked data set; a complete ranked data set that is missing at least one element, where the missing elements represent a neutral preference.

**pi vector** – list containing all the ranked data elements for basis of analysis.

**q vector** – two-dimensional assignment probability vector that contains probabilities (0 – 1.0) of how closely any one particular ranked data element is associated with a particular cluster.

**sigma vector** – list containing current existing cluster centers (σ).