How To Use Factor Clustering Algorithms

Erich Kummerfeld

September 27, 2016

Abstract

This document is a tutorial on how to use the *FindOneFactorClusters* (FOFC) algorithm and similar *Factor Clustering* (FC) algorithms. It covers their acquisition via the free online software, how to use that software to run these algorithms, how to prepare your data so that the software can read it properly, various important features of the algorithms that you should understand and consider when using them, and how to interpret the algorithms' output. This document focuses on the FOFC algorithm, but most of the information applies to other FC algorithms as well.

1 Assumptions I Make About You

Before moving on to the real content of this tutorial document, I will clarify what it is that I am assuming about you, the reader.

I assume that:

- You have access to a working computer and internet connection.
- Your computer has the latest version of Java installed.
- You know the difference between measured variables and unmeasured, or latent, variables.
- You have data to apply a Factor Clustering algorithm to.

2 Preparing Your Data

FOFC can be run on tabular data or on a covariance matrix. If your data is already in a covariance matrix, then there shouldn't be any additional preparation that is required. The rest of this section considers the case where your data is not already stored in a covariance matrix.

When preparing your data for this analysis, there are two main criteria that have to be met. First, the data must be in a format that Tetrad's Data Wrapper can understand. Second, the data must of a type that FOFC works on. I will cover each of these points in this section.

The easiest way to make sure that Tetrad's Data Wrapper can load your data is to format your data in the following way:

- Format your data into a single flat table, with columns corresponding to variables and rows corresponding to samples.
- Make the top row of this table contain variable names. This is not necessary, but often helpful later.
- Make sure all other rows contain ONLY the data samples. Do not include extra rows with more information about the variables.
- There should be no other columns! This means that your tabular data should not include, e.g., a column enumerating or otherwise identifying your samples.

To ensure that FOFC works on your data, confirm that:

- ALL your measured variables can be treated as continuous variables. Binary variables, real-valued variables, variables with a 1-5 scale, will all work. Categorical variables with 3 or more values will not work.
- Your data set does not include more than a thousand variables. Current implementations of FOFC are too slow for data sets with thousands of variables. Unless you are using a supercomputer, make sure your data set has at most several hundred variables.

3 Acquisition

FOFC is included as part of a larger software package called *Tetrad*. The simplest way to get Tetrad is to go to http://www.phil.cmu.edu/tetrad/current.html and download the Jar.

4 Using the Software

Navigate to the file you downloaded, and run it. This will open Tetrad's Graphical User Interface (GUI). This interface is not friendly for new users, and I will not attempt to explain everything that you are seeing, but I will guide you through the specific steps that you can use to run FOFC on your data. If you are interested in learning how to use Tetrad for other purposes, some video tutorials are available at the Tetrad website.

First, left-click on the button on the left of the screen labeled "Data". Then left-click somewhere in the empty white space occupying most of the middle of the screen. You will see a rectangular box appear, labeled Data1. Double-click on Data1, note that the dropdown menu this brings up is set by default to "Data Wrapper" and left-click on OK. This will bring up a window with an empty table.

Inside this new window, in the top left, left-click on "File" and then "Load Data". This will open another window, which you can use to navigate to the data you want to analyze. Navigate to your data, select it, and left-click on "Open" in the bottom right. This will bring up another window, with a lot of options. You will need to select the appropriate options depending on how your data is formatted. In most cases, this means simply selecting whether your data is stored in a table of samples or a covariance matrix, and whether elements in each row of your data are separated by whitespaces, tabs, or commas. Once you've done all of that, left-click on the "Load" button in the bottom right. If there is a problem trying to load the data, the loading log on the right side of the window should indicate something about what the problem is. After your data is loaded, left-click on "Save" at the bottom of the window.

The table that was previously empty in Data1 should now be populated with your data set. If you don't see your data set here, something is wrong. If you ran into problems while trying to load your data, you may need to modify the way your data is formatted before loading it into the data wrapper: read section 2 to check if your data is formatted appropriately, and then return to this section.

Now that your data is loaded into the Data Wrapper, left-click “Save” at the bottom on that window. On the left panel, left-click on the button labeled “Search”. Then left-click in an empty part of the white space. Left-click on the button labeled with a pair of arrows on the left panel (it’s one of the buttons near the bottom), then left-click-and-drag from the box labeled Data1 to the box labeled Search1. This should make an arrow pointing from the Data Wrapper box to the Search box.

Double-click on the Search box. This will bring up a long list of options. Choose the option called FOFC, which should be located about 3/4 of the way down the list, under “Multiple Indicator Model searches”. Then left-click OK. This will open a new window, where you can modify some of the FOFC parameters, run the algorithm by left-clicking on “Execute*”, and view the output.

5 Interpreting the Results

After you have successfully run FOFC on your data set, you will see variable names in boxes moved into a few different groups. At the top will be the largest group, showing all the variables that did not make it into any output clusters. By placing the variables into this group, FOFC is not saying anything except that it does not know what to do with those variables. There are several ways a variable could end up in that box, but I won’t go into detail here what those ways are, as it would require a deeper understanding of how FOFC works. There is really only one reason to keep track of which variables end up unclustered: if, after using a variety of different parameter values, particular variables are always left unclustered, you might consider flagging those variables as “not useful for this analysis”, removing those variables from the data set entirely, and trying the analysis again without them.

Variables that are clustered will end up in one of the groups shown below the unclustered group. Each of these groups will have a label above it such as “_L1” or “_L12”. These are temporary names given by FOFC for the latent variables that it hypothesizes correspond to each cluster of variables. The clusters output by FOFC collectively make the following hypotheses:

- There are at least as many important latent variables as there are clusters.

- Latent variables from different clusters are distinct from one another.
- A measure shares a direct causal relationship *only* with its cluster's latent variable.
- The value of a given latent variable can be safely estimated from the value of the measures in its cluster. Attempting to estimate its value from any other measures *may* be inaccurate.

As for interpreting what these latent variables mean, that is a problem for the user to handle, as it will depend heavily on the variables in the data. The important thing to remember is that these latent variables are *causal*, meaning that they can be intervened upon (at least *in principle*), and other events might occur as an indirect consequence. They are *not* descriptions of the data, but rather, they are hypothesized entities that exist in some meaningful way such that they cause, and can be caused, by other things.

6 Things You Should Know About FOFC

The most important thing to know about Tetrad's implementation of FOFC is that the output you get depends on some additional parameters. The gritty details of FOFC include a variety of places where one could make tweaks, but it is best to leave worrying about those things to the algorithm developers. Most users should only concern themselves with four parameters. Three of these parameters are explicit and directly changeable in the GUI, while the other is implicit and can only be changed indirectly either inside or outside of the GUI. What I mean here by explicit vs. implicit is simply that for the explicit parameters there is a visual display in the GUI alerting you to their existence, while for the implicit parameter there is nothing in the GUI to inform you that it has an impact on the algorithm.

First I will cover the explicit parameters. The first two are labeled in the top left: "Test" and "Algorithm". Personally I recommend setting Test to TETRAD_WISHART and Algorithm to GAP. There's nothing wrong with using the other values of these parameters, but you may want to learn more about how FOFC works and what these parameters mean first.

The third explicit parameter is Alpha. Alpha is an important parameter, with a potentially large impact on the algorithm's performance, so you should not ignore it. Without going into the details, FOFC runs a very large number of statistical tests during its runtime, and this parameter is used for those tests to determine if the null hypothesis is accepted or rejected each time. The default value that it takes in the GUI is typically a poor choice. For a starting value, I recommend $1/n$, where n is your sample size, but you should experiment with a variety of different values to see how it affects the output.

The implicit parameter is *the order of your variables*. It may seem a bit strange, but FOFC's output can change when you change the order that your variables are in. Here, by the order of your variables, I am referring to the order that the variables appear in, from left to right, in the data set. FOFC will, for instance, tend to cluster together variables that are closer to each other in the variable ordering than variables which are far away from each other. The variable ordering can be changed in a variety of ways in the GUI, such as by using a Data Manipulation box prior to feeding the data into the Search box. You can also change the variable order in your tabular data file prior to loading it into Tetrad in the first place: the Data Wrapper inherits the variable order from the original data file. Either way, the main thing to remember is that variable ordering *does* matter, and you may want to experiment with different variable orderings, or even try random variable orderings.