morderstats.py Documentation

March 23, 2017

This document describes how to use the program *morderstats.py* to generate multivariate prediction regions from forecast data. The name *morderstats* is short for *multivariate order statistics* and indicates the capability of this program to produce multivariate information using different methods and then order them based on certain metrics. Currently the only methods of producing prediction regions are by using the Mahalanobis distance, Direct Convex Hull Peeling, and Halfspace Convex Hull Peeling.

1 Required Modules

morderstats.py is dependent on the entire scipy stack. This is most easily obtainable by downloading and installing the most recent Anaconda distribution and can be found at `https://www.continuum.io/downloads`. Alternatively, one can install the modules *numpy*, *scipy*, *matplotlib*, and *pandas* by using the command `pip install <module>` and replacing `<module>` with the appropriate module name. To achieve complete functionality of the program, you will need to have installed *scipy* 0.19 or later.

There is an alternate implementation of the program which is dependent on the module *pyhull*. This can be installed using the command `pip install pyhull`.

2 Formatting the Input Data

To properly use this program, the input file should simply contain the raw data. Each row should represent an individual point in multiple dimensions, and each column should represent a dimension. The following text demonstrates how to properly format the file:

```
101,101
102,103
102,107
107,110
```

3 Configuring Options

To actually execute the program, the user must create an options file which specifies the various options required for running the program. For clarity purposes, we will refer to this file as `morderstats_exec.txt` but note that this file can be given any name.

This file must start with the following line verbatim:

```
command/exec morderstats.py
```

This specifies that the options file is for executing *morderstats.py*

There are four options that are mandatory and one option that is optional. The first option that must be set is `--sources-file` which specifies the data file with which you wish to compute multivariate order statistics. The second option that must be set is `--method` which designates which method you wish to use for producing the prediction regions. If set to *mahal*, then regions will be computed using the mahalanobis distance. If set to *halfspace*, then regions will be computed using the halfspace peeling algorithm. If it is set to *direct*, then regions will be computed using the direct convex hull peeling algorithm.

The option `--write-file` must be set and designates the name of the text file which will contain information about the prediction regions computed. The option `--write-directory` must also be set and designates the name of the directory to contain all files produced by *morderstats*.

The option `--alpha` is optional and designates the desired proportion of points to be outside the prediction region. If unspecified and `--method` is set to *mahal*, then regions will be computed for $\alpha = 0.01, 0.02, \ldots, 0.99$. If it is unspecified and the `--method` is set to either *halfspace* or *direct*, then the entire sequence of convex hulls will be computed by the respective method.

The following figure demonstrates an example for how the options file could be written:

```
command/exec morderstats.py
--sources-file test_files/random_data_2d.csv
--method direct
--alpha 0.1
--write-file direct_regions.json
--write-directory direct_regions
```

Once the options file is written, *morderstats.py* can be executed with the following command:

```
python runner.py morderstats_exec.txt
```

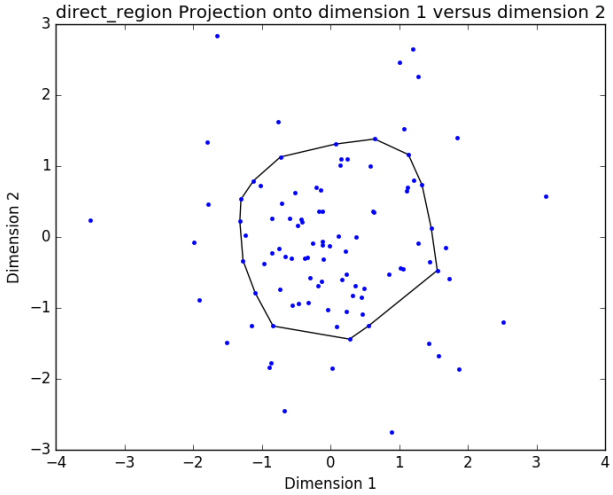
If you gave your options file a different name than *morderstats_exec.txt*, then you can change the command to use the appropriate name.

4 Output Files

After running, the program will output data files and plots in the directory specified in the options file. The first will be a text file named `regions.csv` which contains the points for the defining polygon of the convex hull for the specified alpha level and method of producing the prediction region. It will also contain the proportion of points outside the region as well as the volume of the region. Here is an example of what this file may look like:

```
Alpha: 0.4
Volume: 3.001125099478032
Points:
-0.6402172620000001,0.70008383
-0.830046849,0.479766109
-0.919373599,0.252441087
-0.975288009,-0.01288018
-0.9706044090000001,-0.048589188
-0.937123695,-0.228741992
```

The second file produced will be a plot of the region or regions. The following figure is an example plot that was produced by *morderstats*



5 Testing

This program comes with an associated test suite for testing if *morderstats* or the associated modules in *pint* function properly. To run this tester, simply execute the following command in the terminal in the *pint* directory:

```
python tester.py
```