Version 1.1, May 1997

Spinne User's Guide

A program for visualizing the distortions due to dimensional reduction.

Introduction

Projection methods such as principal components analysis (PCA), nonlinear mapping (NLM) or self-organizing maps (SOM, Kohonen's map) are valuable algorithms for visualizing high-dimensional data on two-dimensional (2D) plane . Unfortunately, the reduction of the dimensionality involves distortions. To graphically localize the distortions of the projected data, we have developed a program "Spinne", which superimposes colored graphs onto 2D plots. The color of the edges of these graphs encodes the original high-dimensional distances between the connected points.

Spinne is written in the C programming language. It runs on most UNIX and DOS (32 bit mode) computers. Since the program does not have a graphical user interface, it can easily be ported to new platforms. The main graphics output file format is Encapsulated Postscript (EPS), which can be used by many programs. Spinne works better on NeXTSTEP and OPENSTEP computers.

Note that the aim of this user guide is not to explain the application of Spinne to multidimensional data analysis, but to describes briefly the functionalities of the program. Further information, examples and references can be found at the WWW address http://schiele.organik.uni-erlangen.de/Bruno_Bienfait/Spinne/ .

Example run

The distribution of Spinne contains input file examples, which can be found in the "samples" subdirectory. Change to this directory, and enter the command:

```
spinne -nD 4d.vec -2D 2d.vec
```

If everything goes right, three new EPS files are created in the current working directory. These files can be viewed with a PostScript previewer. File **`4d.vec`** contains 37 four-dimensional vectors, whereas **`2d.vec`** contains 37 two-dimensional vectors obtained by taking the two first principal components (PCA method) of **`4d.vec`**.

Usage

The command line options of Spinne are:

```
spinne [-h | -help] [-nD FileND.vec | -inter File.inter] [-2D File2D.vec] [-3D File3D.vec]
[-mst] [-mstlow] [-dist aNumber] [-labels File.lab] [-select File.sel]
```

- **-h** or **-help** : prints some help.

- **-nD** : the argument is the name of the file containing the coordinates in the original n-D space.
- **-inter** : the argument is the name of the file containing the interpoint distances. The **-nD** and **-inter** options are mutually exclusive.

- **-2D** : the argument is the name of the file containing the coordinates in the projected 2-D space.
- **-3D** : the argument is the name of the file containing the coordinates in the projected 3-D space.

- **-mst** : minimal spanning tree (MST) calculated in the original n-D space.
- **-mstlow** : minimal spanning tree (MST) calculated in the 2 (or 3) -D space (the edges are colored according to the original space).
- **-select** : the argument is the name of the file containing pairs of numbers for the manual selection of the edges (one pair of integers per line)

- **-labels** : the argument is the name of the file containing a list of characters strings, one per line, to label each point of the plot.

Depending on the command line options, Spinne creates encapsulated PostScript files named `mst.eps`, `mstlow.eps`, `select.eps`, `dist.eps` , or *Mathematica* files `distrs3D.m`, `mst3D.m` if 3-D plots are requested. If such files already exist in the current working directory, they will be overwritten. It is up to the user to rename or copy the output files.

Input file formats

Spinne uses four different kinds of input file formats. They are all ASCII text formats. The most important ones are those related to the multidimensional vectors.

Multidimensional vectors (options -nD, -2D and -3D)

This is an example of of a 3-D data file containing four entries:

```
3 4
0.5 0.98 1.54
4.8 8.0 0.54
3.14156 10e-3 4.5
-2.8 10.5 4.2
```

The first line is composed of two numbers m, n separated by one or more blank characters. The first number m specifies the dimension of the vectors. The second number n specifies the number of entries in the file, e.g., the number of vectors. Optionally, a third number o can be present in the first line. It tells the program to ignore the o (considered as output or y values) last columns. The n next lines consists of then $n \times m$ floating point numbers.

Labels (options -labels)

Labels are used to mark each point of the plot. By default, Spinne uses numbers, 1,2,3,4, ... , m where m is the number of points to be displayed, i.e the number of vectors in the input files. A label file consists of m lines. For example, if m=4, a label file could be:

```
first
second
third
fourth
```

Selection of lines to be displayed (options -select)

Spinne provides minimal spanning tree and largest distortion graph to select lines between points. One may wish to use a custom graph, which would draw colored lines between some selected points. The simplest way to achieve this is to use the -select option to import a file containing pairs of numbers (integers) *i*, *j*. For example, a selection file could be:

```
3 2
1 4
4 3
```

Note that the numbers *i*, *j* may not be larger than m, the number of points (or vectors).

Interpoint distances (options `-inter`)

After reading a multidimensional data file with the **`-nD`** option, Spinne calculates all interpoint distances. The only distance metric available in Spinne is Euclidean. To overcome this limitation, Spinne can import from a file all the *m(m-1)/2* interpoint distances. For example, if *m*=4, the number of interpoint distances is 6, and a file for the **`-inter`** option could contain:

```
4
3.221
1.478
4.30
0.20
32.56
19.435454
```

The first line of the file indicates *m*, the number of points. The subsequent lines correspond to the distance between points 1 and 2, 1 and 3, etc. The order series for *m* =4 is thus:

```
1 2
1 3
2 3
2 4
3 4
```

Output file formats

Encapsulated PostScript (EPS)

PostScript is a computer language designed for high quality graphics printing. Spinne creates graphics files in the format called Encapsulated PostScript. These files can be viewed on the screen with the help of PostScript viewer, for example Ghostview, which is freely available for many platforms. On X windows systems equipped with Display PostScript, one can use also the program xpsview. On NeXTSTEP/OPENSTEP computers, the EPS file can simply be opened by a double click, printed, and dragged and dropped to graphics applications.

The EPS output file does not contain a preview picture as it is usually the case on Macintoshes and PCs. However, it can be imported by desktop publishing and word processing programs such as MS Word, Framemaker, and probably many others.

You can send it directly to a postScript printer, but probably, nothing will be printed because the PostScript "showpage" instruction is absent from the file. This command can be easily added with the help of text editor.

Mathematica

With the -3D option, Spinne generates a 3-D plot file, which can be viewed with *Mathematica*.

Problems and limitations

The computation time and the memory needed by Spinne increase with the square of the number of points.