Visualising Relationships between Multi-Species Measures of Biodiversity and the Environment

Ritei Shibata Keio University, Yokohama, Japan

Outline

- Data Visualisation
- What is Textile Plot?
- GBR data
- Exploratory analysis through Textile Plot
- Grouping taxa through Textile Plot

The aim of data visualisation

- Illustrate current status (sensor data)
 - Plant status
 - Network status
- Illustrate result
- Explore original data
 - High dimensional
 - Large data (records)
 - View the data as it is
 - Mixed data types
 - Numeric, Logical or Categorical
 - Help understanding of data

Parallel Coordinate Plot







Parallel Coordinate Plot

Edsall, 2002, Unwin et al., 2003, Tory et al., 2004,

- ✓ High dimensional
- ✓ Large data (records)
- ✓ View the data as it is
- Mixed data types
 - Numeric, Logical or Categorical Proper choice of coordinates
- Help understanding of data
 - Proper choice of scale and location for each axis

LD triangular display with squared correlation coefficients for Control



Categorical 4 Variables: HLA-C, HLA-B, HLA-DRB1, HLA-DPB1

Easier to grasp whole picture of data More details if necessary



species peral with sepal Length



Iris Flower

PetalLength

Horizontalisation Criterion

 Choose location and scale of each axis so that connected line segments become as horizontal as possible

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \beta_j \mathbf{x}_j$$
: coordinates on each axis $j = 1, 2, ..., p$

$$\sum_{j=1}^{p} \|\mathbf{y}_{j} - \boldsymbol{\xi}\|^{2} \xrightarrow[\alpha_{j},\beta_{j},j=1,...,p,\boldsymbol{\xi}]{\min}$$

n

Kumasaka and Shibata, High-dimensional data visualisation: The textile plot. 2007, Computational Statistics and Data Analysis

$$\sum_{j=1}^{p} \|\mathbf{y}_{j} - \boldsymbol{\xi}\|^{2} = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} (y_{ij} - \boldsymbol{\xi}_{i})^{2} \right) \to \min$$

 ξ_i : target horizotal level of the *i* th record

$$\sum_{j=1}^{p} (y_{ij} - \xi_i)^2 : \text{ squared deviance of the } i \text{ th record}$$

from the level ξ_i

Mixed Data Type Case

If \mathbf{x}_{j} is categorical, apply a contrast to get a data matrix X_{j}

 $\mathbf{y}_{j} = \alpha_{j} \mathbf{1} + \boldsymbol{\beta}_{j} X_{j}$: coordinates on the *j* th axis

Horizontalisation criterion $\sum_{j=1}^{p} \|\mathbf{y}_{j} - \boldsymbol{\xi}\|^{2} \xrightarrow[\alpha_{j}, \beta_{j}, j=1,..., p, \boldsymbol{\xi}]{\min}$ determines a *unique* location of the levels of each category.

Independent of the choice of contrast



Proportional to the multiplicity of the value

GBR data





Most frequent 50 taxa



taxon1 ... taxon50 : logical covariate1 ... covariate34: numeric

Piers K. Dunstan, Scott D. Foster and Ross Darnell, Model based group of species across environmental gradients Ecological Modelling, 2010

Textile Plot (34 covariates and 50 taxa)



Hierarchical Clustering of Axes (Variables)



√- 127.0.0.1

Order of axes and knots

- Order of axes
 - Order appeared in the data table
 - Variance of each axis
 - Clustering of axes
 - Distance of two axes=Sum of squares of slopes
 - Ordered single end-linkage clustering algorithm(Hurley, 2004)
- Variables with Knots
 - Orthogonal to other variables

Covariates Orthogonal to Existence of Taxa

Standard deviations and other covariates (20 covariates)

Remove

[1] "GBR_ASPECT" "GBR_SLOPE" "GBR_TS_BSTRESS" "GMCS_STRESS_TMN"
[5] "GMCS_STRESS_IQR" "NO3_SD" "PO4_SD" "O2_AV"
[9] "O2_SD" "S_SD" "T_SD" "SI_SD"
[13] "CHLA_AV" "CHLA_SD" "K490_SD" "SST_SD"
[17] "BIR_AV" "BIR_SD" "TRWL_EFF_I" "TOPO_CODE"

14 Covariates



14 Covariates



Group1: Sand appetite 9 Taxa



√ 127.0.0.1

Remaining 41 Taxa: Knots: NO3_AV, PO4_AV and K490_AV



127.0.0.1

After 3 knots removed



Group2: Most frequent 3 taxa



127.0.0.1

Remaining 38 taxa



Remaining 38 taxa



Group3: Unique taxon



Group3: Unique taxon



Group4:22 taxa



Group5: 15 taxa



Grouping taxa through TextilePlot

- Knot covariates removed step by step
- Examine Hierarchical Cluster Tree of Axes (Variables)



Textile Plot Home Page

http://www.stat.math.keio.ac.jp/TextilePlot/index.html

Textile Plot Public Alpha The world's first data browser. Now on Windows+Mac+Linux.

Textile Plot is the first versatile data browser in the world. With its simple, familiar interface, Textile Plot improves to interpret given data with various attributes being displayed as well as missing value information.

Download Now

What's Textile Plot?

The tectle pot (Kumasaka and Shibata 2007 In press) is a parallelcoordinate plot In which the ontering, locations and scales of the axes are similareously chosen so that the connecting lines, each of which represents a case, are aligned as horizontally as possible. Pibls of this type can accommodate numerical data se well as ordered or unordered categorical data, or a mixture of these different data types. The leadle pible data set, with various attributes of the total byte nata set, with various attributes of the data byte nata set, with various attributes of the data byte nata set, with various attributes of the data byte nata set, with a sing value information. Knots and parallel wefts within the testile piot also ad in the visual interpretation of the data. Several practice examples are presented which illustrate the potential usefulness of the testile piot as an ald to the visualitation of multivariate data.

Design Principle

As Cleaking states in the elements of graphing data ', a graphical method is successful only if the decoding process from the giken graphic by the viewer is effective. Thus, our aim in designing the textile pict was not only to graphically represent the data points themselves but also to assist the user in their interpretation of the data. With this aim in mind, it would appear reasonable to display any other information that might be heipful to the user in the textile pict together with the data. Also the ordering of the axes needs to be carefully determined.

How to Use

Appearance of Testile Pick is very similar to ordinary web provver. Use it and learn how to use it. The difference is that the target object is not at HTML file but related within is described by XML along with DanD DTD. We call the XML document DanD (Data and Description) instance. You can specify any instance at the URL field, you can see the whole picture of the data in the DanDD instance. An easiest way to create a DandD instance at the URL field, you can see the whole picture of the data in the DanDD instance. An easiest way to create a DandD instance throm a CSV the is to instail CSV2DAD (EKL, AR) you may find any other softwares related to DandD Instance in DandD Project Home Page. You need an internet connection since Testile Pixt is a client. Software of the DandD Server-Client System. You need to Instail a DandD Server on your local computer is no intermet connection is available.

Visual Operation

User can explore data or Textile Plot through various interactions like zooming, highlighting and so on. Visual instructions given by user aria text in Textile Plot according to a reference model proposed in Kumasaka and enitbata (2006 in Japanee). The reference model to design an lossile indivoment for working with data through the textile plot consists of a sequence of four opjeds, the data, the parallel coordinate the visual analogue and the textile plot objects. A data object is transformed into a parallel coordinate object which is a set of coordinate vectors. The visual analogue is an abstract representation of the textile plot opploced. The textile plot object is a textile plot but constructed without any restriction in the real world like size of display or fits resolution. User can vector this object through various interfaces like zooming or resizing. Visual instructions given by user are, therefore, sent to one of the objects according to its own nature.

Publication

Kurrasaka N. et al. (2008) High-dimensional data visualisation: The textile plot. Computational Statistics & Data Analysis, 52(7):3616-44

Talks

Kumasaka N. (2012) A Haplotype Visualisation of Multi-Allelic Genetic Markers. IBC2012 Kobe, invited Session 9: Data Visualization: Optimization and Applications

Shibata R. (2012) Visaulising Relationships between Multi-Species Measures of Biodiversity and the Environment, IBC2012 Kobe, Invited Session 9: Data Visualization: Optimization and Applications 合大口