



COMPARISON OF MULTIVARIATE DATA REPRESENTATIONS: THREE EYES ARE BETTER THAN ONE

Antony Unwin
Augsburg University

Natsuhiko Kumasaka
Keio University

MULTIVARIATE DATA REPRESENTATION

○ Numerical data

- Biplot (Gabriel 1971)
- Scatter plot matrix (Cleveland 1984)
- GGobi (Cook, D. et.al 2007)
- Glyphs (Anderson 1957, Chernoff 1973, Fienberg 1979)
- Parallel coordinate plot (Inselberg 1985, Wegman 1990)
- Matrix visualization (Chen 2002)

○ Categorical data

- Mosaic plot (Hartigan 1981)
- MANET (Unwin et.al. 1996)

○ Numerical/Categorical data

- Trellis/Lattice (Chambers 1992, Cleveland 1993)
- Textile plot (Kumasaka and Shibata 2008)

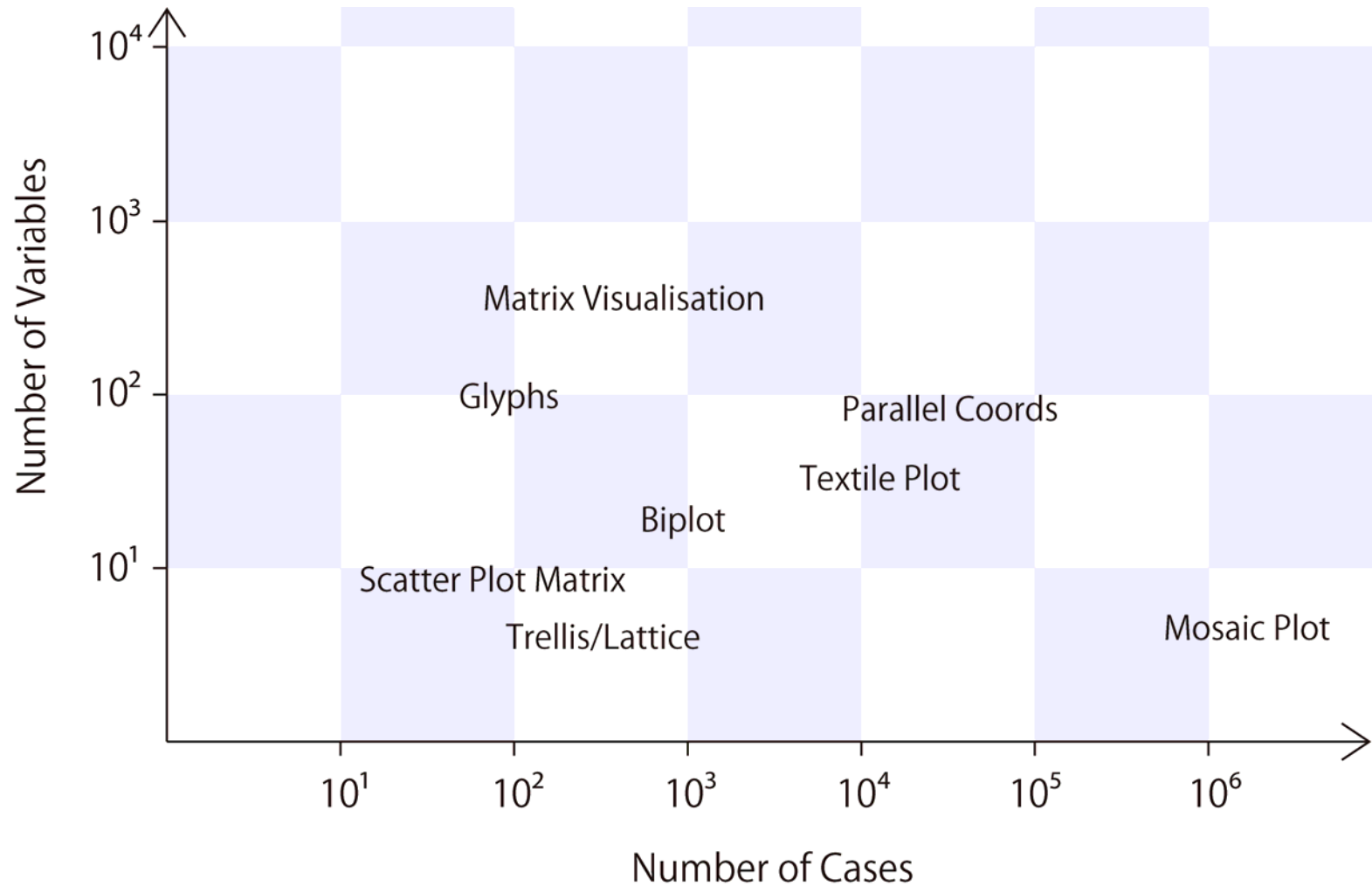
○ General

- Graphics of Large Datasets (Unwin et.al. 2006)
- Handbook of Data Visualization (Chen, et.al. 2008)

CONTENT

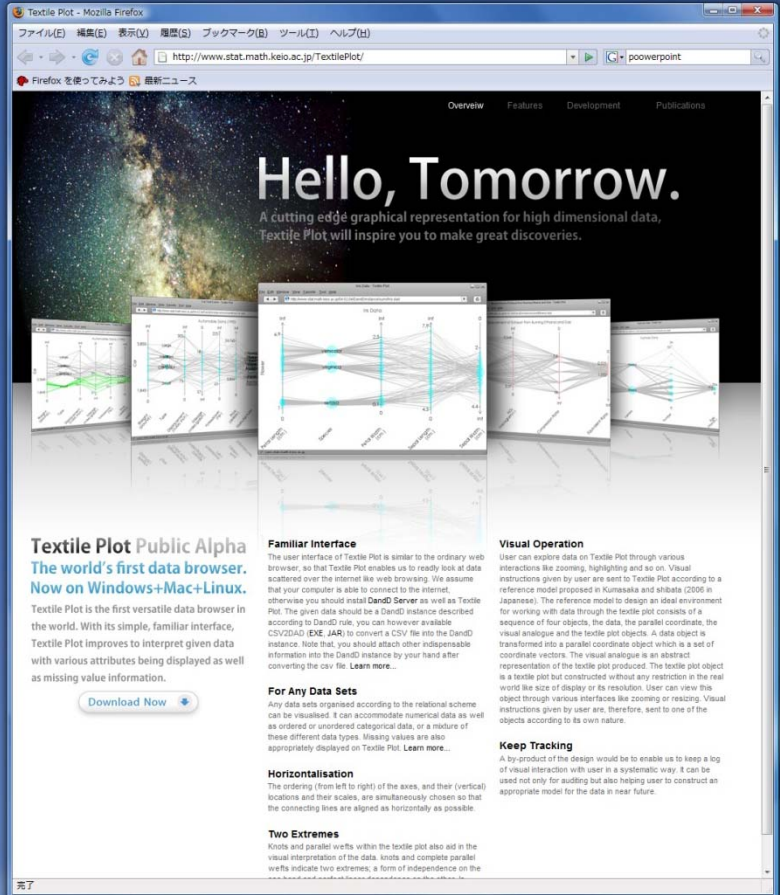
- 3 multivariate data representations
 - Parallel coordinate plot
 - Mosaic plot
 - Textile plot
- Visual data analysis
 - Decathlon data
 - Wine data
 - Animal data
 - Titanic data

CAPABILITY OF THE NUMBER OF CASES AND VARIABLES ON A DISPLAY



TEXTILE PLOT (Kumaska and Shibata 2006, 2007, 2008)

- Parallel coordinate system
- Horizontalisation criterion
- Any type of data
- Order of Axes



Hello, Tomorrow.
A cutting edge graphical representation for high dimensional data,
Textile Plot will inspire you to make great discoveries.

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](#)

Familiar Interface
The user interface of Textile Plot is similar to the ordinary web browser, so that Textile Plot enables us to readily look at data scattered over the internet like web browsing. We assume that your computer is able to connect to the internet, otherwise you should install **DandD Server** as well as Textile Plot. The given data should be a DandD instance described according to DandD rule, you can however available CSV2DAD (EXE, JAR) to convert a CSV file into the DandD instance. Note that, you should attach other indispensable information into the DandD instance by your hand after converting the csv file. [Learn more...](#)

For Any Data Sets
Any data sets organised according to the relational scheme can be visualized. It can accommodate numerical data as well as ordered or unordered categorical data, or a mixture of these different data types. Missing values are also appropriately displayed on Textile Plot. [Learn more...](#)

Horizontalisation
The ordering (from left to right) of the axes, and their (vertical) locations and their scales, are simultaneously chosen so that the connecting lines are aligned as horizontally as possible.

Two Extremes
Knots and parallel wefts within the textile plot also aid in the visual interpretation of the data. knots and complete parallel wefts indicate two extremes, a form of independence on the

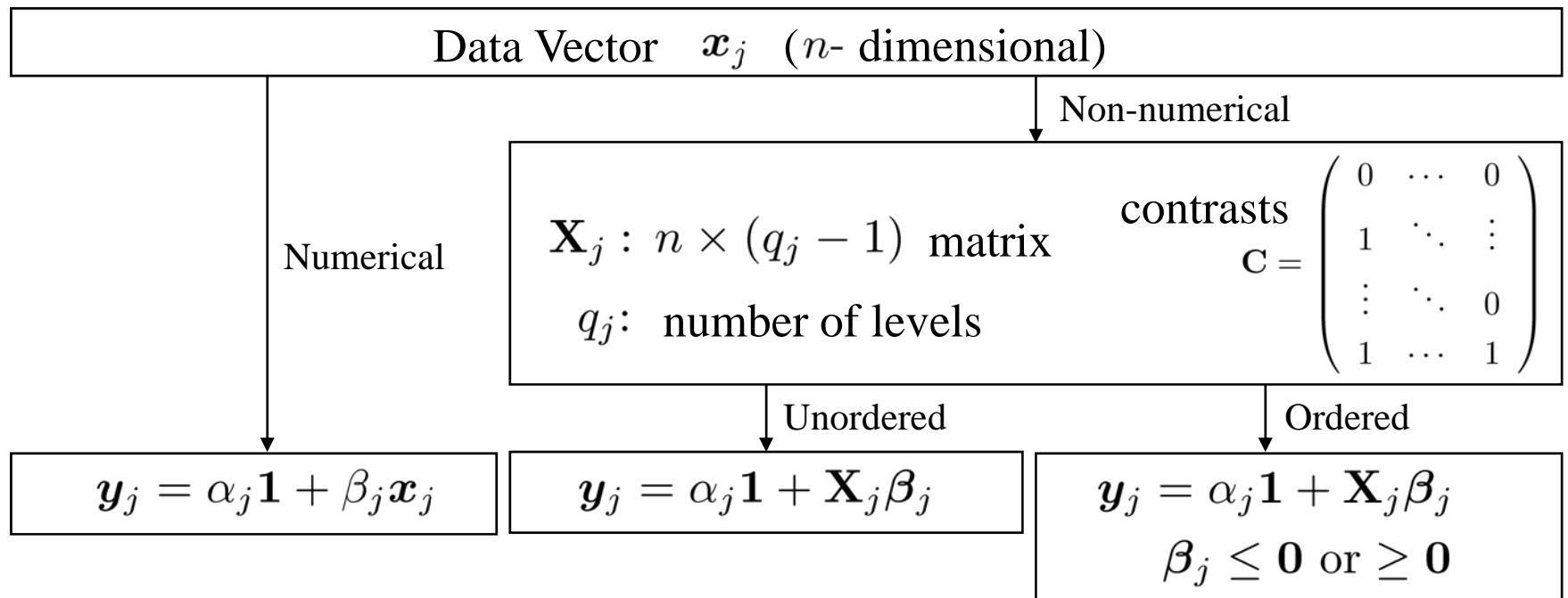
Visual Operation
User can explore data on Textile Plot through various interactions like zooming, highlighting and so on. Visual instructions given by user are sent to Textile Plot according to a reference model proposed in Kumaska and Shibata (2006 in Japanese). The reference model to design an ideal environment for working with data through the textile plot consists of a sequence of four objects, 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 produced. The textile plot object is a textile plot but constructed without any restriction in the real world like size of display or its resolution. User can view 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.

Keep Tracking
A by-product of the design would be to enable us to keep a log of visual interaction with user in a systematic way. It can be used not only for auditing but also helping user to construct an appropriate model for the data in near future.

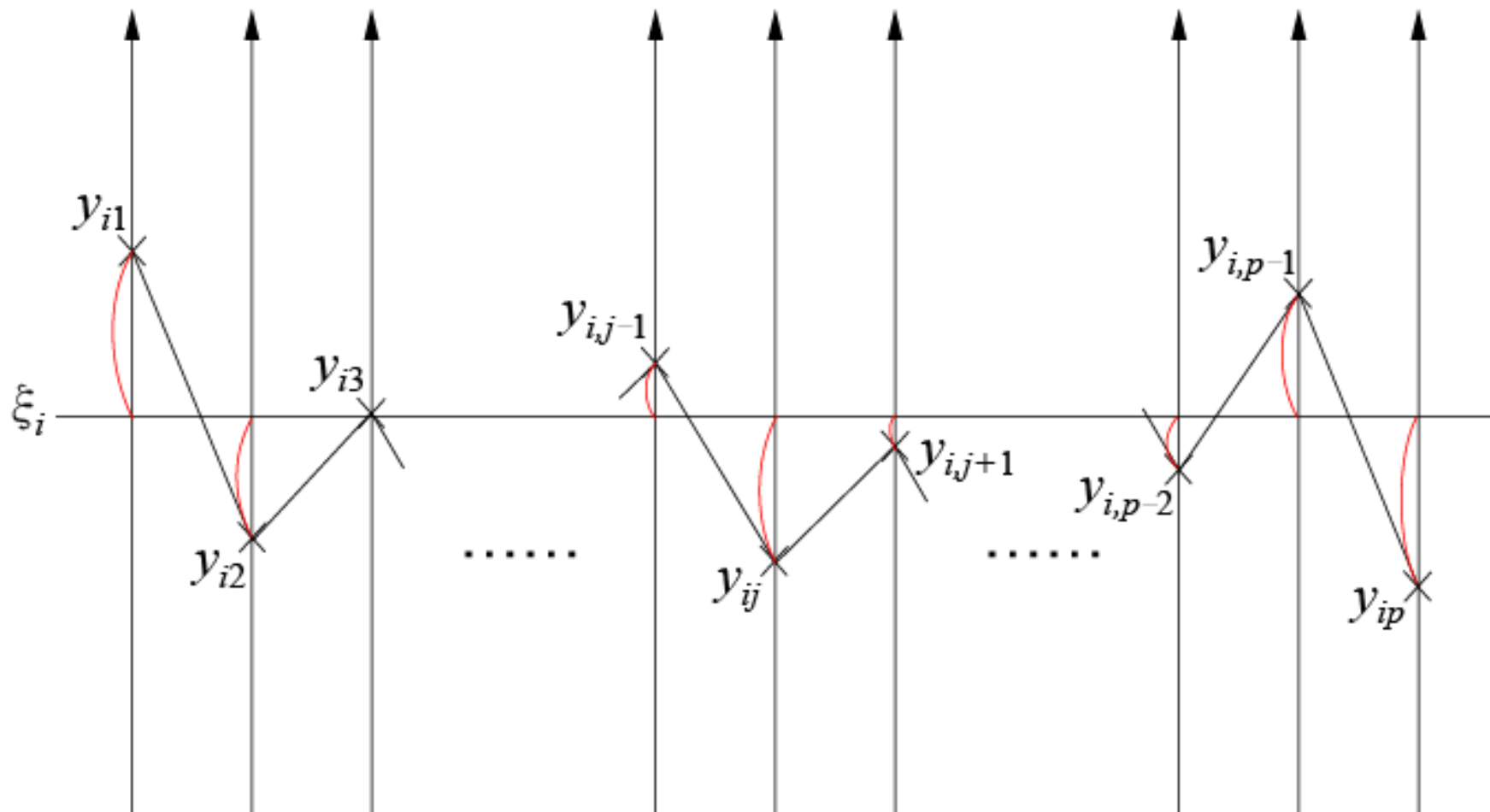
<http://stat.math.keio.ac.jp/TextilePlot/>

TRANSFORM DATAVECTOR INTO COORDINATE VECTOR

$$\begin{pmatrix} x_{ij} \\ n \times p \end{pmatrix} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \Rightarrow \begin{pmatrix} y_{ij} \\ n \times p \end{pmatrix} = (\mathbf{y}_1, \dots, \mathbf{y}_p)$$



HORIZONTALISATION CRITERION

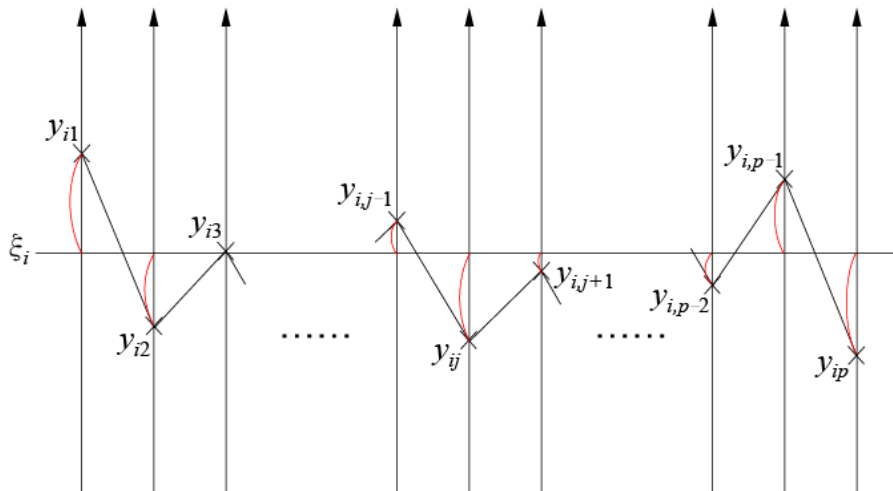


$$\sum_{j=1}^p (y_{ij} - \xi_i)^2$$

OPTIMISATION PROBLEM

Minimise
$$\sum_{i=1}^n \sum_{j=1}^p (y_{ij} - \xi_i)^2 = \sum_{j=1}^p \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 \xrightarrow{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}} \min$$

Subject to
$$\sum_{j=1}^p \|\mathbf{y}_j - \bar{y}_{\cdot j} \mathbf{1}\|^2 = np$$



Location Parameter Vector

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$$

Scale Parameter Vector

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$$

Ideal Coordinate Vector

$$\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$$

SOLUTION ($\mathbf{1}^T \mathbf{x}_j = 0$ and $\|\mathbf{x}_j\| = 1$; $j = 1, \dots, p$)

$$\sum_{j=1}^p \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 = \sum_{j=1}^p \|\mathbf{y}_j - \mathbf{m}\|^2 + p\|\mathbf{m} - \boldsymbol{\xi}\|^2$$

$$\Rightarrow \hat{\boldsymbol{\xi}} = \mathbf{m} = \frac{1}{p} \sum_{j=1}^p \mathbf{y}_j \quad (\mathbf{m} : \text{mean vector})$$

$$\sum_{j=1}^p \|\mathbf{y}_j - \mathbf{m}\|^2 = \sum_{j=1}^p \|\mathbf{y}_j - \bar{y}_{\cdot j} \mathbf{1}\|^2 - p\|\mathbf{m} - \bar{y}_{\cdot\cdot} \mathbf{1}\|^2 + \sum_{j=1}^p \|\bar{y}_{\cdot j} \mathbf{1} - \bar{y}_{\cdot\cdot} \mathbf{1}\|^2$$

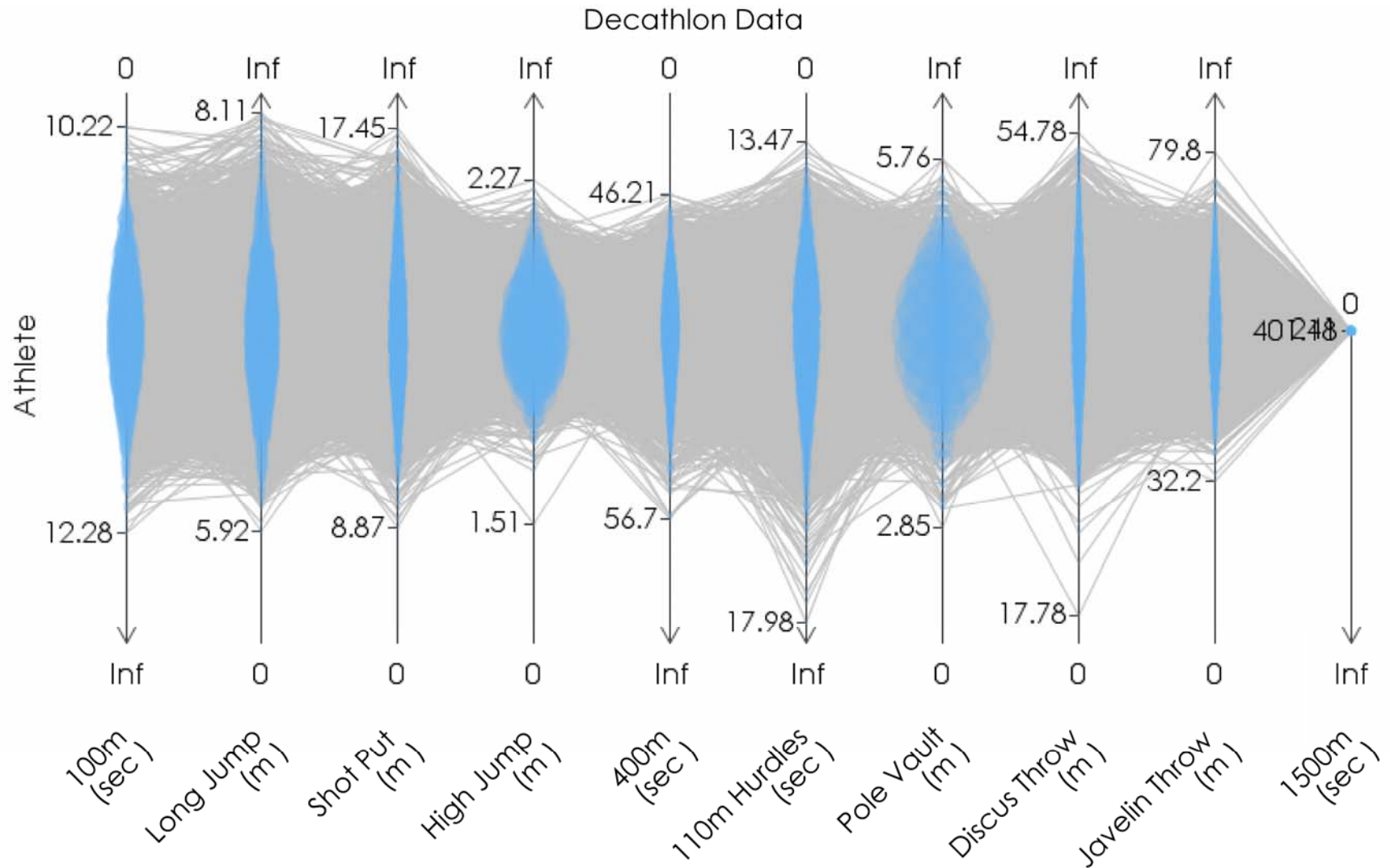
$$\Rightarrow \hat{\alpha}_j = \alpha_0; \quad j = 1, \dots, p \quad (\bar{y}_{\cdot\cdot} = \sum_{j=1}^p \bar{y}_{\cdot j} / p)$$

$$\sum_{j=1}^p \|\mathbf{y}_j - \bar{y}_{\cdot j} \mathbf{1}\|^2 - p\|\mathbf{m} - \bar{y}_{\cdot\cdot} \mathbf{1}\|^2 = \|\boldsymbol{\beta}\|^2 - \frac{1}{p} \boldsymbol{\beta}^T \mathbf{R} \boldsymbol{\beta}$$

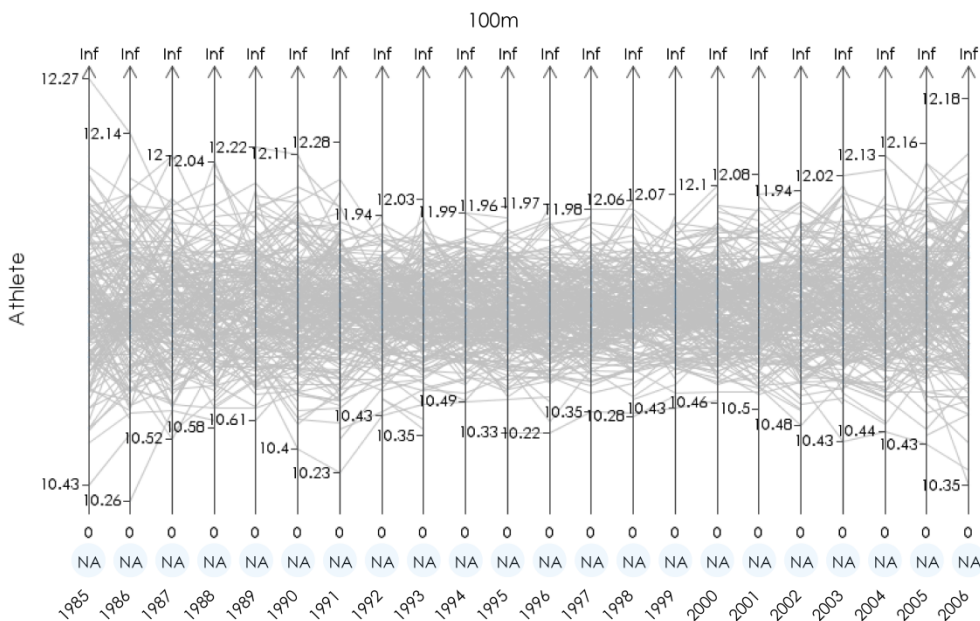
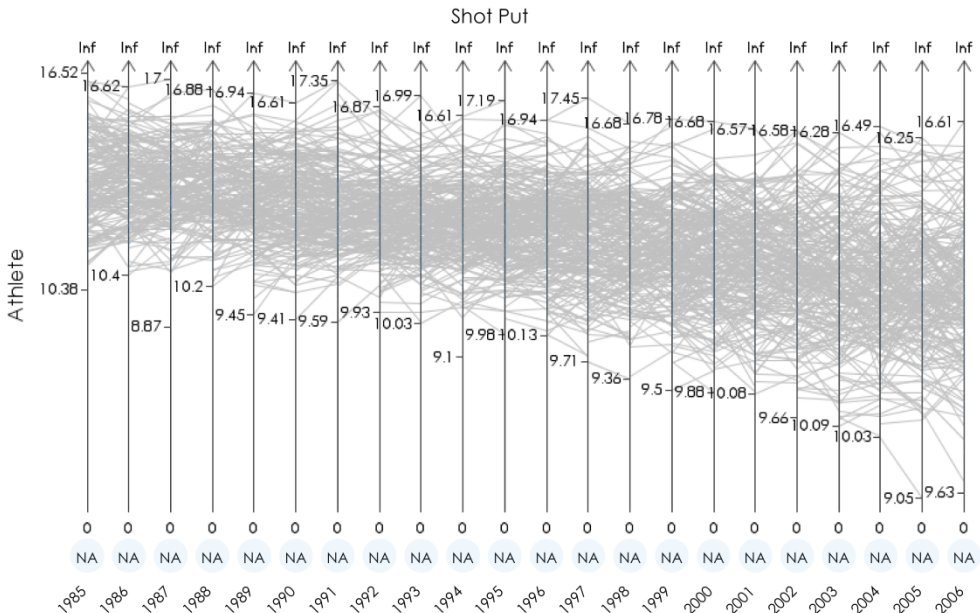
$\Rightarrow \hat{\boldsymbol{\beta}}$: Eigenvector of sample correlation matrix \mathbf{R} with the largest eigenvalue, such that $\|\hat{\boldsymbol{\beta}}\|^2 = np$.

$$(\sum_{j=1}^p \|\mathbf{y}_j - \bar{y}_{\cdot j} \mathbf{1}\|^2 = \|\boldsymbol{\beta}\|^2 = np)$$

TEXTILE PLOT OF DECATHLON DATASET (PERFORMANCES NOT POINTS)

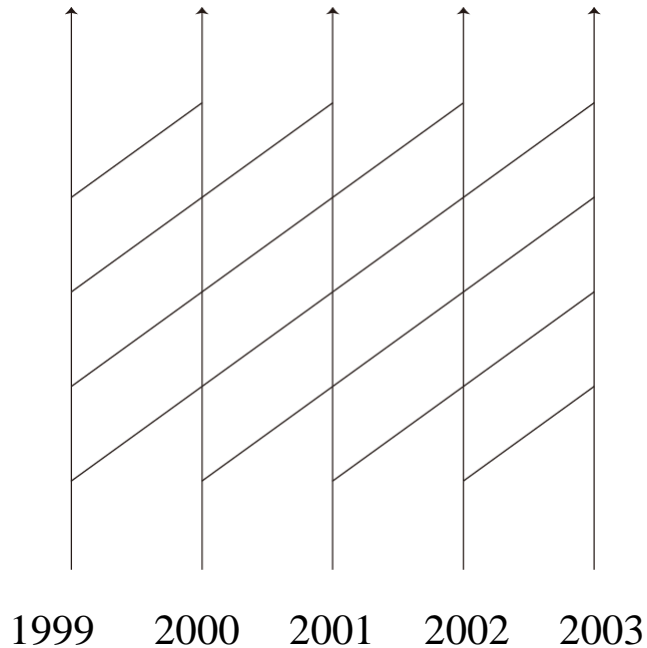


RE-ORGANISED DECATHLON DATASET



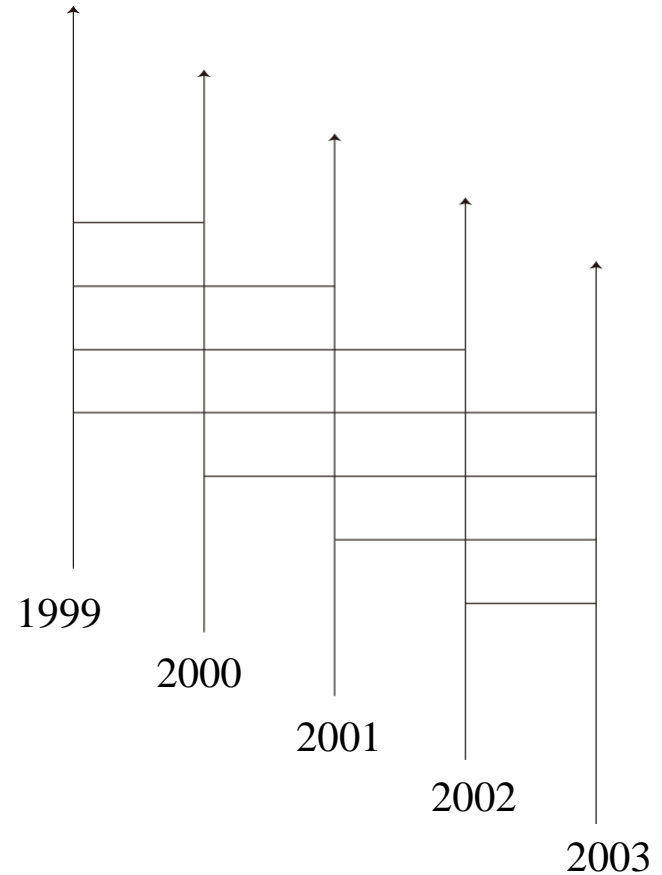
TREND ON TEXTILE PLOT

Assumption: All athletes improve their performances year by year, and stop their careers at their peak.



Parallel Coordinate Plot
with Common Scaling

Horizontalisation

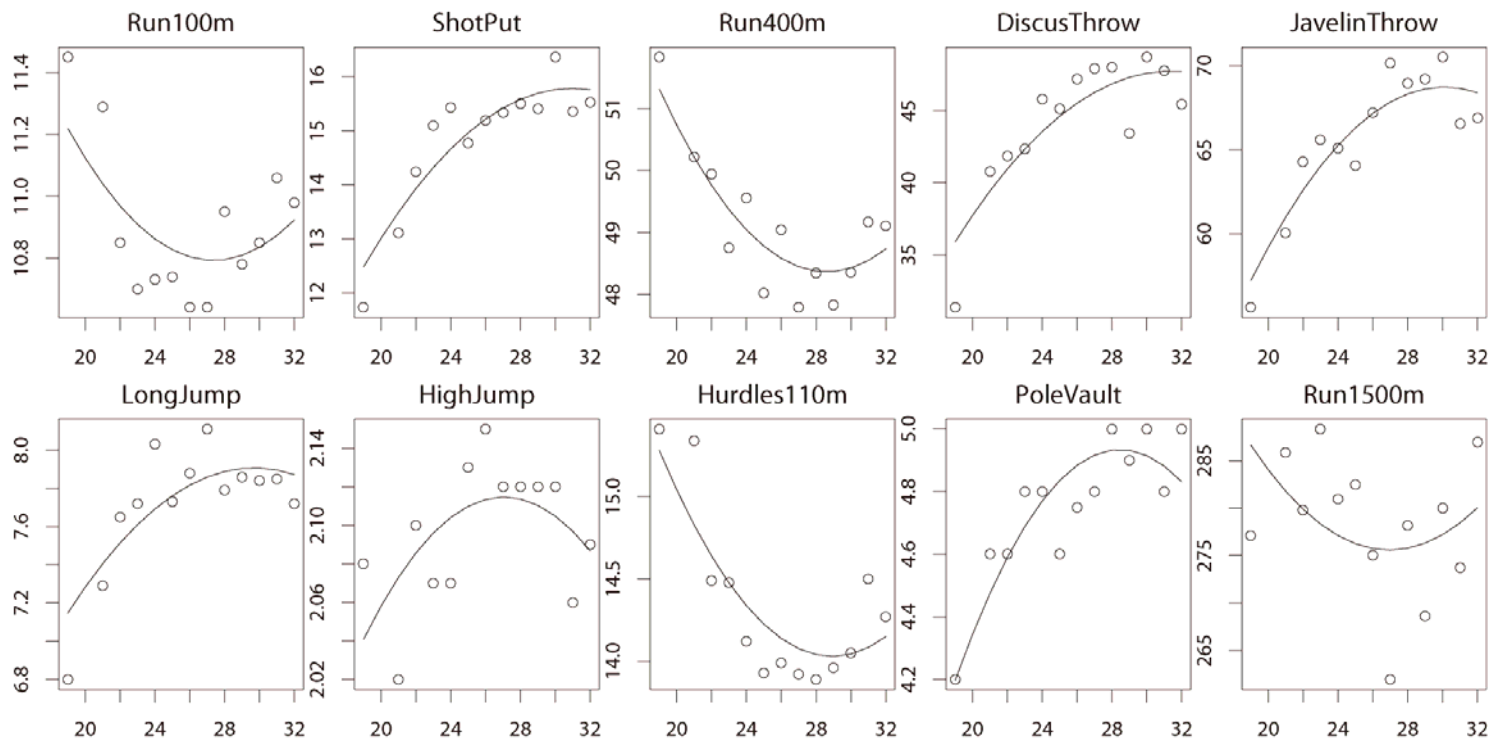


Textile Plot

FURTHER CONFIRMATORY DATA ANALYSIS



- Performances of Mr. Roman Sebrle (best record holder)



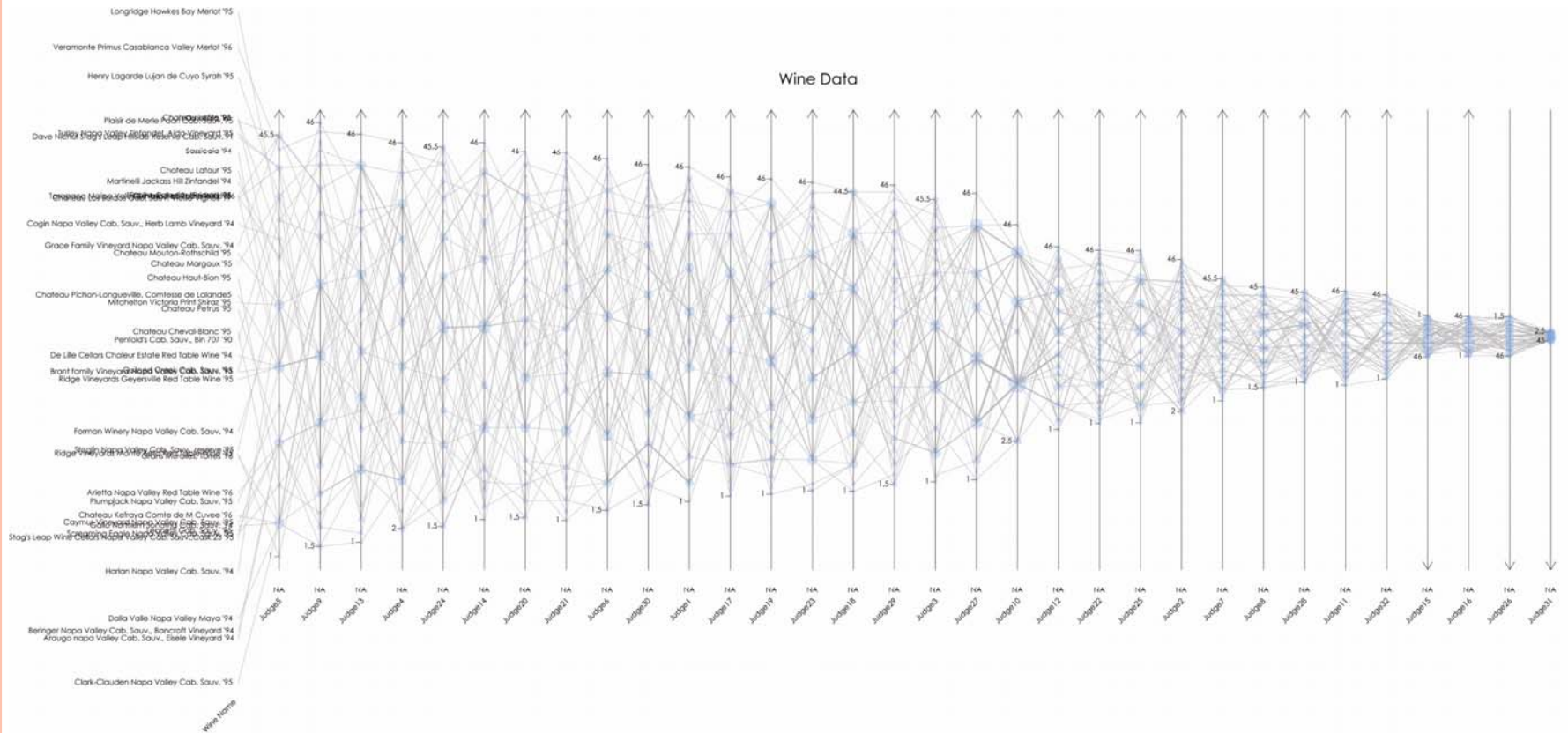
$$Performance_{ik} = \alpha_{ik} + \beta_k (Age_i - \gamma_{ik})^2 + \varepsilon_{ik},$$

i : athlete, k : event

WINE DATASET (LIQUID ASSETS: WWW.LIQUIDASSET.COM)

- Cabernet challenge 1999
- Case
 - 47 (only 46 rated) Cabernet Sauvignons: 34 US, 9 French, 2 Italian, 2 others
 - Vintages from 1994 to 1996
- Variable
 - 33 judges (Californian) ranked the wines
- Analysis goals:
 - Which wines were rated best?
 - Is the ranking of wines clear-cut?
 - Do the judges have similar opinions?
 - Are there clusters of judges?

TEXTILE PLOT OF WINE DATASET

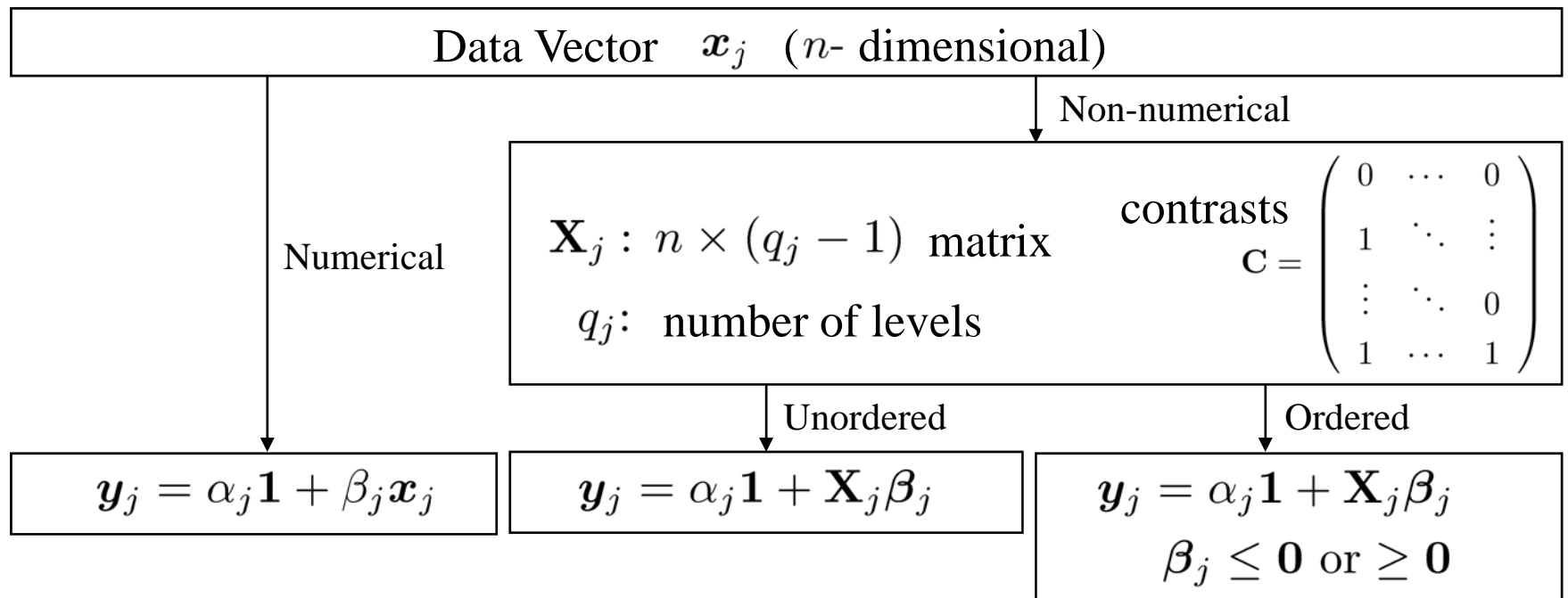


GRAPHICS FOR MULTIVARIATE NUMERICAL DATA

- Textile plot or parallel coordinate plot gives an overview of the data
- Re-organisation of data is always useful to know another aspects of the data
- Textile plots suggest potential avenues for subsequent further confirmatory data analysis

TRANSFORM DATAVECTOR INTO COORDINATE VECTOR

$$\begin{pmatrix} x_{ij} \\ n \times p \end{pmatrix} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \Rightarrow \begin{pmatrix} y_{ij} \\ n \times p \end{pmatrix} = (\mathbf{y}_1, \dots, \mathbf{y}_p)$$



CATEGORICAL DATAVECTOR

Coordinate vector transformation

- Determine optimised default position for each level
- Introduction of a set of contrasts

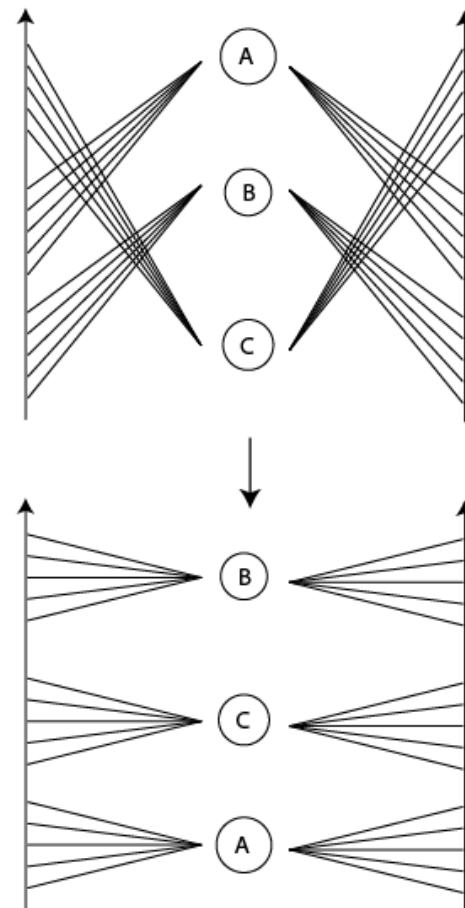
Ex: Using a *treatment* contrast

A	(0	0)
B	(1	0)
C	(0	1)

$$\mathbf{x} = \begin{pmatrix} A \\ A \\ B \\ C \\ C \end{pmatrix} \longrightarrow \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Coordinate vector:

$$\mathbf{y} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_2 \\ \alpha + \beta_2 \end{pmatrix}$$



ORDERED CATEGORICAL DATAVECTOR

Order of levels are retained on the axis

- Introduction of the specific contrasts
- Additional constraints

Ex:

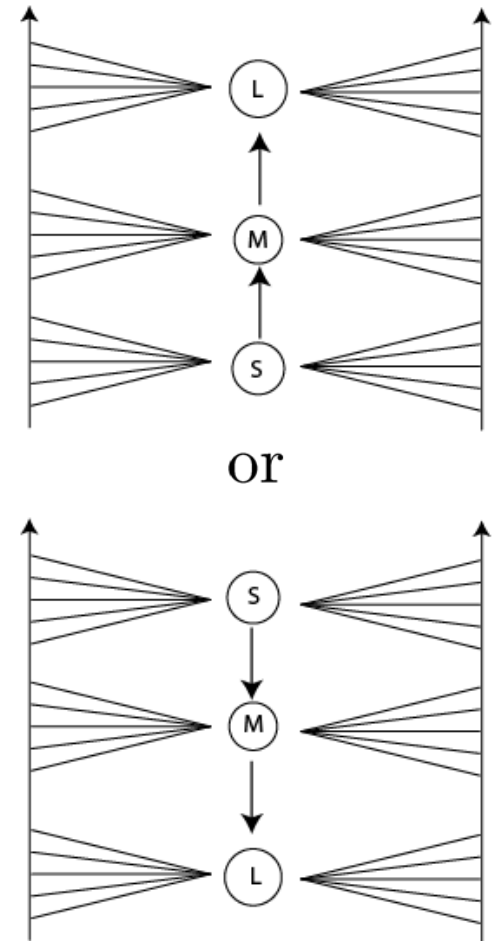
$$\begin{matrix} \textit{Small} \\ \textit{Medium} \\ \textit{Large} \end{matrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} \textit{Small} \\ \textit{Small} \\ \textit{Medium} \\ \textit{Large} \\ \textit{Large} \end{pmatrix} \longrightarrow \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Coordinate vector:

$$\mathbf{y} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_1 + \beta_2 \\ \alpha + \beta_1 + \beta_2 \end{pmatrix}$$

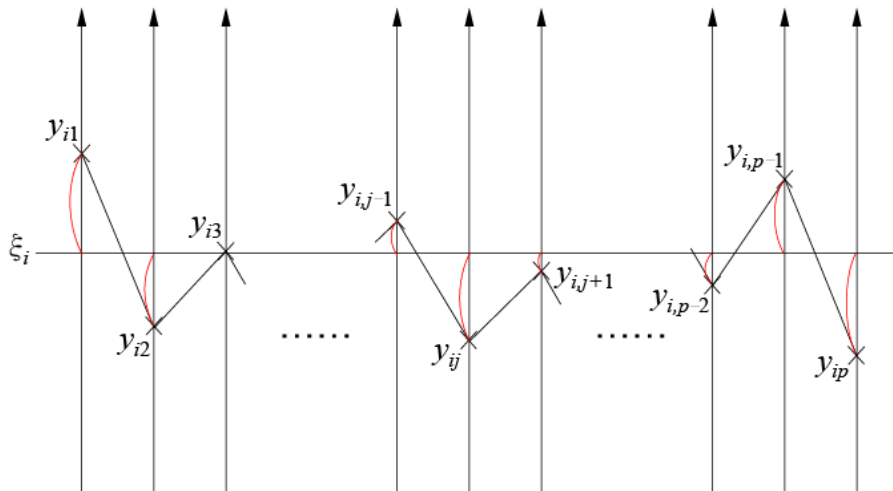
$$\beta_1, \beta_2 \geq 0 \quad \text{or} \quad \beta_1, \beta_2 \leq 0$$



OPTIMISATION PROBLEM

Minimise
$$\sum_{i=1}^n \sum_{j=1}^p (y_{ij} - \xi_i)^2 = \sum_{j=1}^p \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 \rightarrow_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}} \min$$

Subject to
$$\sum_{j=1}^p \|\mathbf{y}_j - \bar{y}_{\cdot j} \mathbf{1}\|^2 = np \quad (\boldsymbol{\beta}_j \geq \mathbf{0} \text{ or } \boldsymbol{\beta}_j \leq \mathbf{0})$$



Location Parameter Vector

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$$

Scale Parameter Vector

$$\boldsymbol{\beta} = (\beta_1^T, \dots, \beta_p^T)^T$$

Ideal Coordinate Vector

$$\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$$

ANIMAL DATA (UCI MACHINE LEARNING GROUP 2008)

○ Case

- 101 animals
- Invalid cases
 - Two frogs
 - Girl?

○ Response

- Animal type
 - Mammal
 - Reptile (爬虫類)
 - Amphibian (両生類)
 - Fish
 - Insect
 - Bird
 - Invertebrate (無脊椎動物)

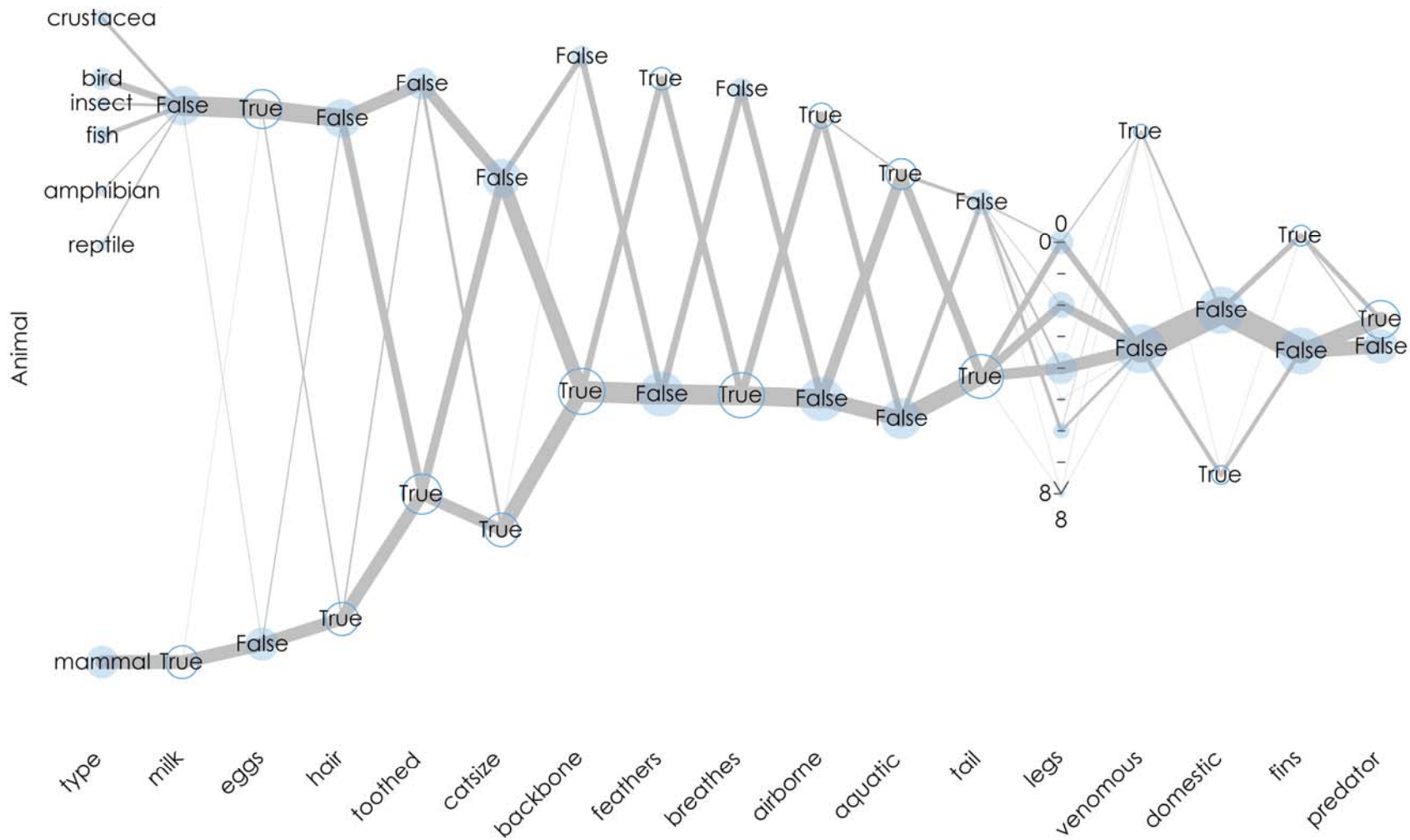
○ 16 Covariates (binary)

- Hair/Feathers/Eggs/Milk/Airborne/Aquatic/Predator/Toothed/Backbone/Breathes/Venomous/Fins/Legs/Tail/Domestic/Cat-size

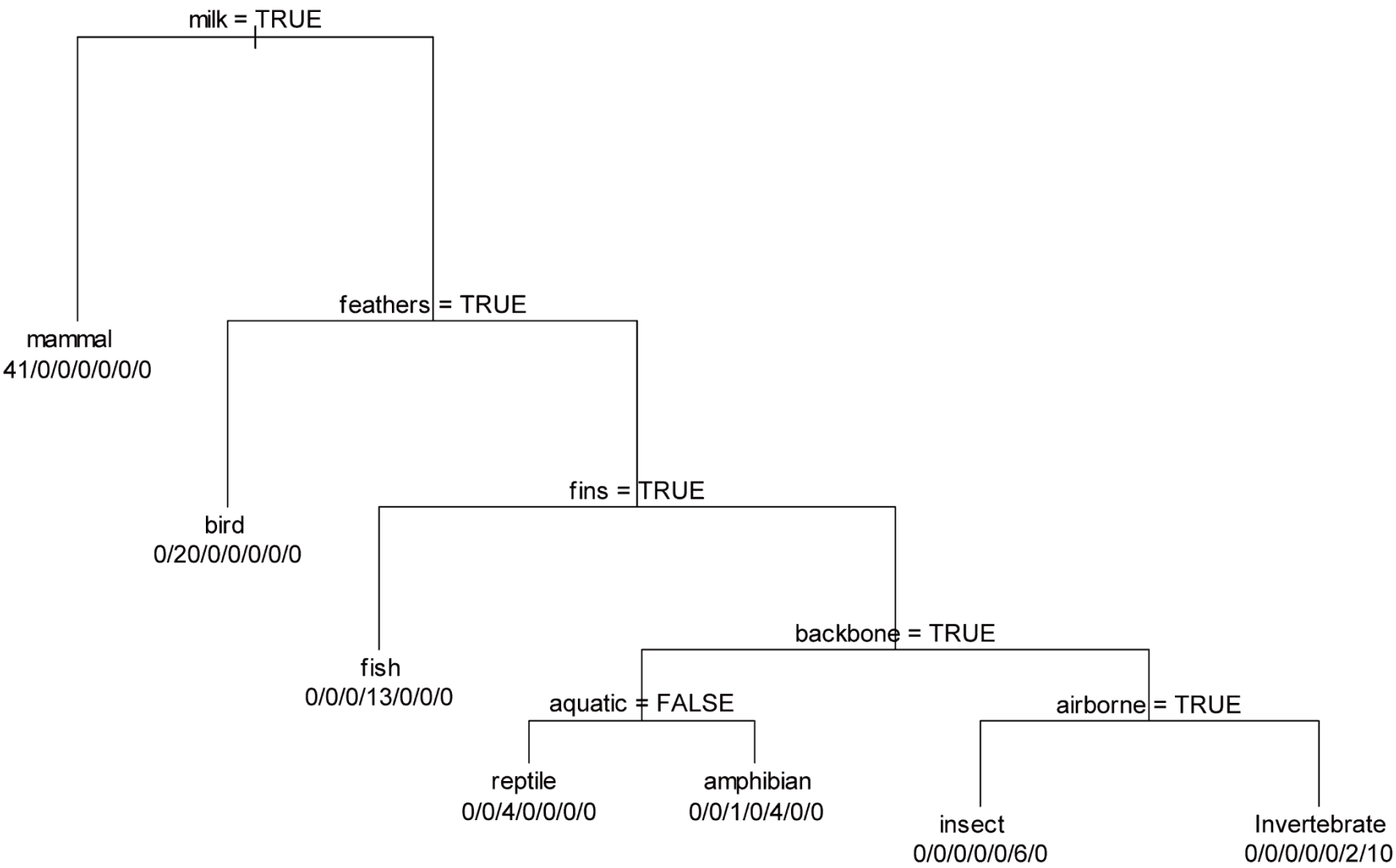
○ Analysis goals:

- What features best classify animals by type?
- How are the features related?



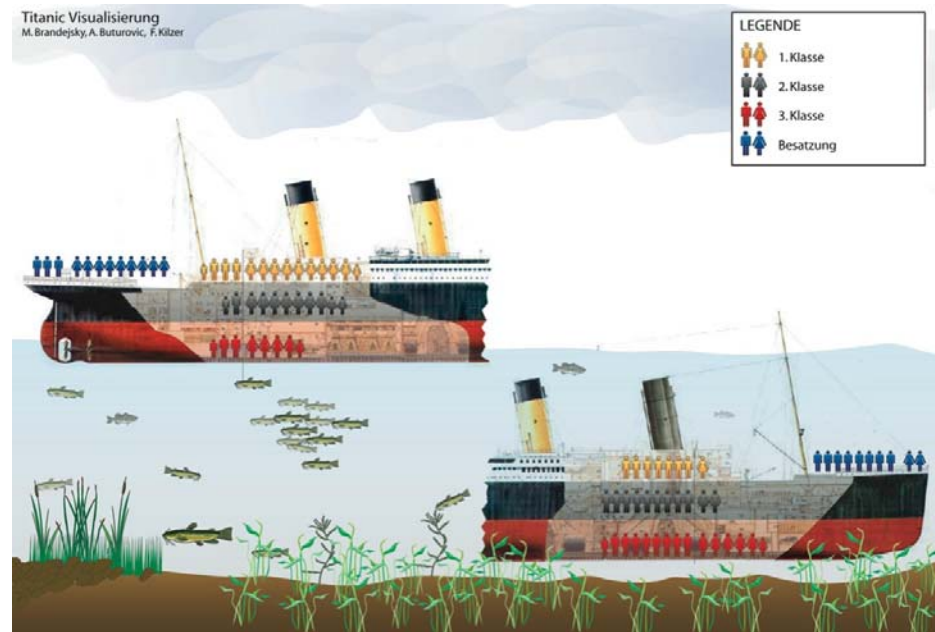


RELATION TO PARTITIONING TREE

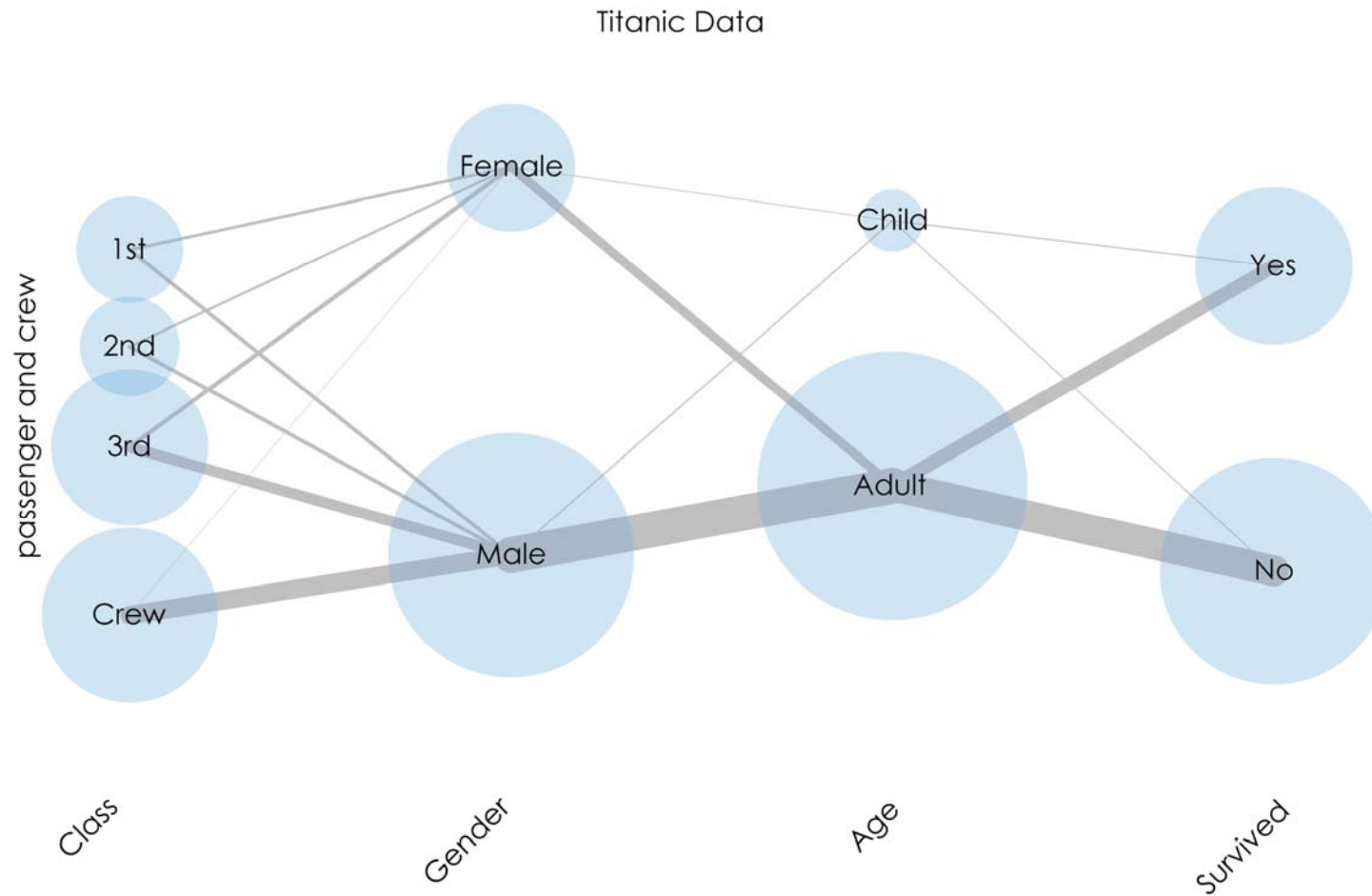


TITANIC DATASET (BRITISH BOARD OF TRADE 1990)

- Case
 - 2201 passengers and crew
- Variable
 - Class (First, Second, Third, Crew)
 - Age (Young, Old)
 - Gender (Male, Female)
 - Survived (Yes, No)



TEXTILE PLOT OF TITANIC DATASET

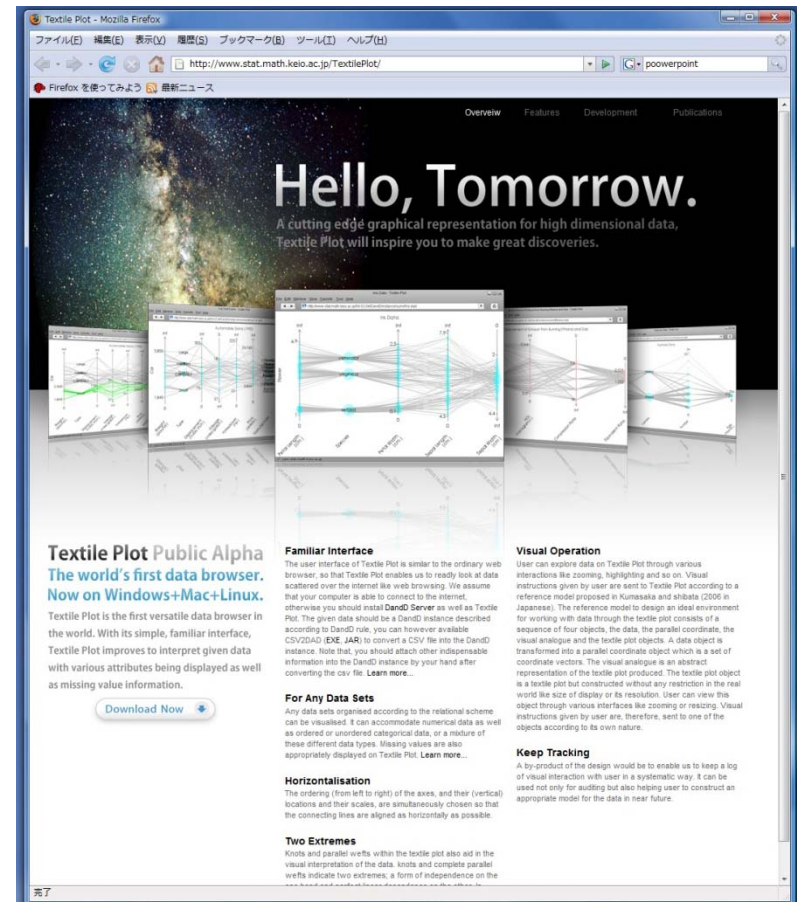


SUMMARY: TEXTILE PLOT FOR MULTIVARIATE CATEGORICAL DATASETS

- Textile plots provide rough idea of classifying cases
- Textile plots of multivariate categorical data emphasise absolute numbers
- Detailed conditional probability is difficult to interpret

SOFTWARE: TEXTILE PLOT ENVIRONMENT

- Network ready
 - Based on DandD Client Server System (Yokouchi and Shibata 2004)
- Cross-platform
 - JAVA JRE 1.5
- Interactive user interfaces
 - Reference model (Kumasaka and Shibata, 2007)



<http://stat.math.keio.ac.jp/TextilePlot/>

CONCLUSIONS

○ Textile plot

- Show an overview of the given data in an optimal way
- Accommodate numerical and categorical data
- Suggest several avenues for further exploratory or confirmatory data analysis

○ Three eyes are better than one

- Further investigations should be carried out with other graphical representations

BIBLIOGRAPHY

- Anderson, E. (1957). A semigraphical method for the analysis of complex problems. *Proceedings of the National Academy of Sciences* **13** 923-927.
- British Board of Trade (1990), *Report on the Loss of the 'Titanic' (S.S.)*. British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing.
- Chambers, J.M., and Hastie, T.J. (1992) *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove CA.
- Chen, C. H. (2002) Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica* **12** pp. 7--29.
- Chen, C. H., haerdle, W. and Unwin, A.R. (2008) *Handbook of Data Visualization*, Springer, Berlin.
- Chernoff, H. (1973) The use of faces to represent points in k-dimensional space graphically, *Journal of American Statistical Association* **68** 361-368.
- Cleveland, W.S. and McGill, R. (1984). The Many Faces of a Scatterplot, *Journal of the American Statistical Association* **79** 807-822.
- Cook, D. and Swayne, D. Interactive and dynamic graphics for Data Analysis, Springer New York.
- DandD Project (2007) Home Page, <http://www.stat.math.keio.ac.jp/DandD/>.
- Decathlon 2000 (2008) Decathlon 2000 Home Page, www.decathlon2000.ee.
- Fienberg, S.E. (1979) Graphical Methods in Statistics, *The American Statistician* **33** 165-178.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics* , **58** (3) pp. 453-467.
- Hartigan, J. A. and Kleiner, B. (1981) Mosaics for Contingency Tables, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer-Verlag, pp.268-273.

BIBLIOGRAPHY

- Hurley, C. B. (2004) Clustering Visualizations of Multidimensional Data, *Journal of Computational and Graphical Statistics* **13** 788-806.
- Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer* **1** 69-91.
- Kumasaka, N., Shibata, R. (2007) Textile Plot Environment, 統計数理特集号「統計データの可視化」, **55**, 47-68.
- Kumasaka, N. and Shibata, R. (2008) High Dimesional Data Visualisation: the Textile Plot, *Computational Statistics & Data Analysis*, Submitted, **52**, 3616-3644.
- Kumasaka, N. ANOVA on Textile Plot, *Proceedings in COMPSTAT 2008*, submitted.
- Liquid Assets (2008) Liquid Assets Home Page, <http://www.liquidasset.com/>.
- Unwin, A.R., Hawkins, G., Hofmann, H., and Siegl, B. (1996) *Interactive Graphics for Data Sets with Missing Values* – MANET, *Journal of Comp and graph Stat*, **5** 113-122.
- Wegman, E. (1990) Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association* **85** 664-675.
- Yokouchi, D. and Shibata, R. (2004), DandD: Client Server System, *Proceedings in COMPSTAT 2004*, Physica-Verlag.