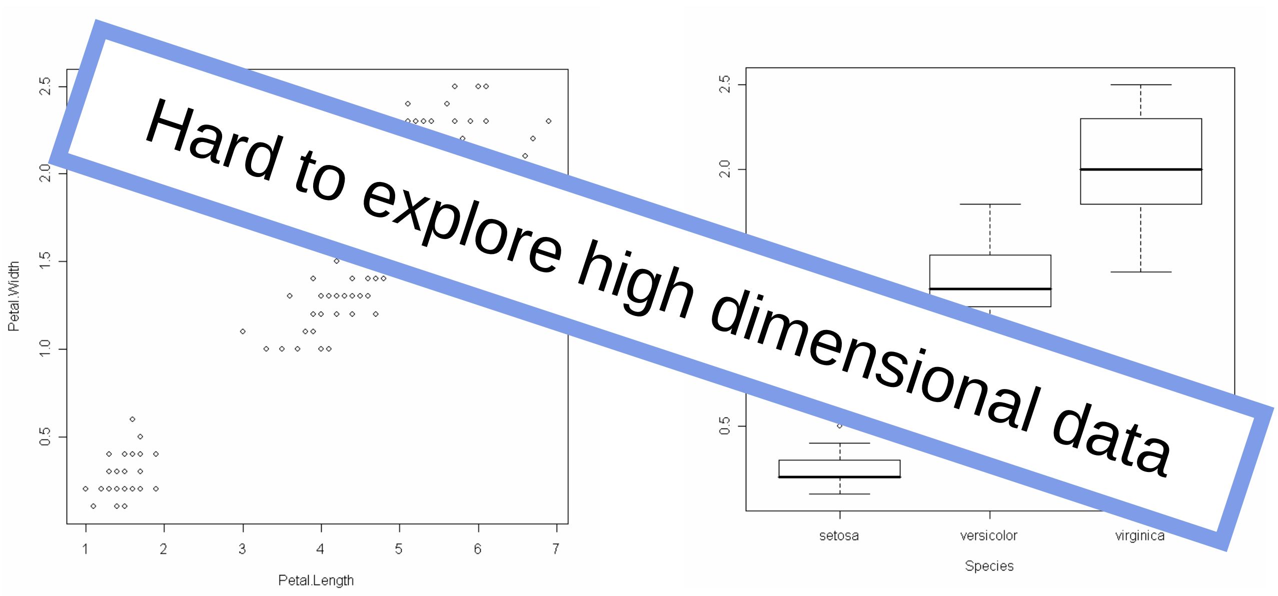
High Dimensional Data Visualisation: the Textile Plot

Natsuhiko KUMASAKA
PhD Student
School of Fundamental Science and Technology
Keio University
kumasaka@stat.math.keio.ac.jp

Ritei SHIBATA
Department of Mathematics
shibata@math.keio.ac.jp

Building good models from data

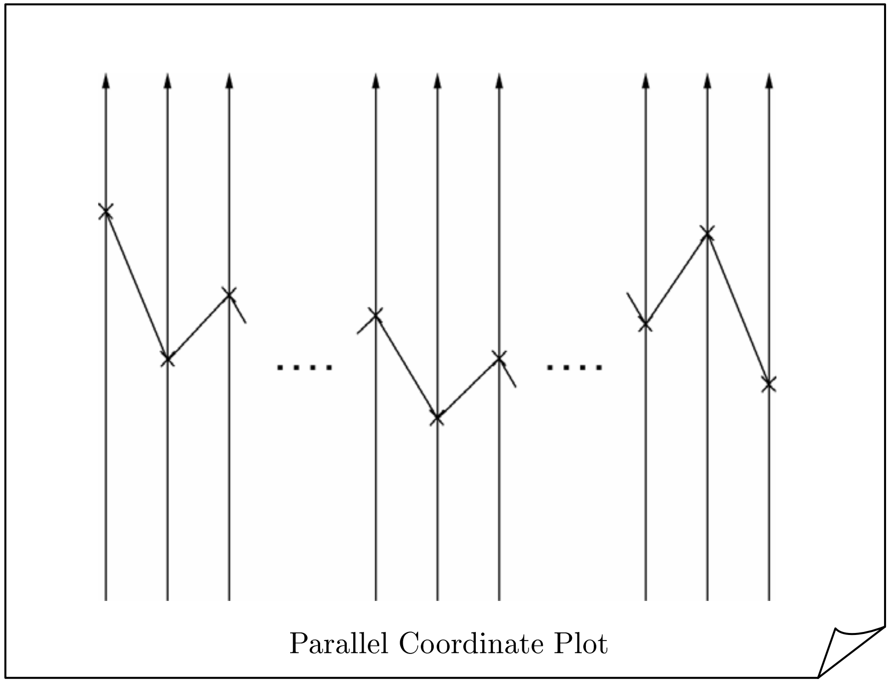
- Exploring data through visualisation
  - Finding outliers
  - Clustering observations
  - Investigating relationships between variables



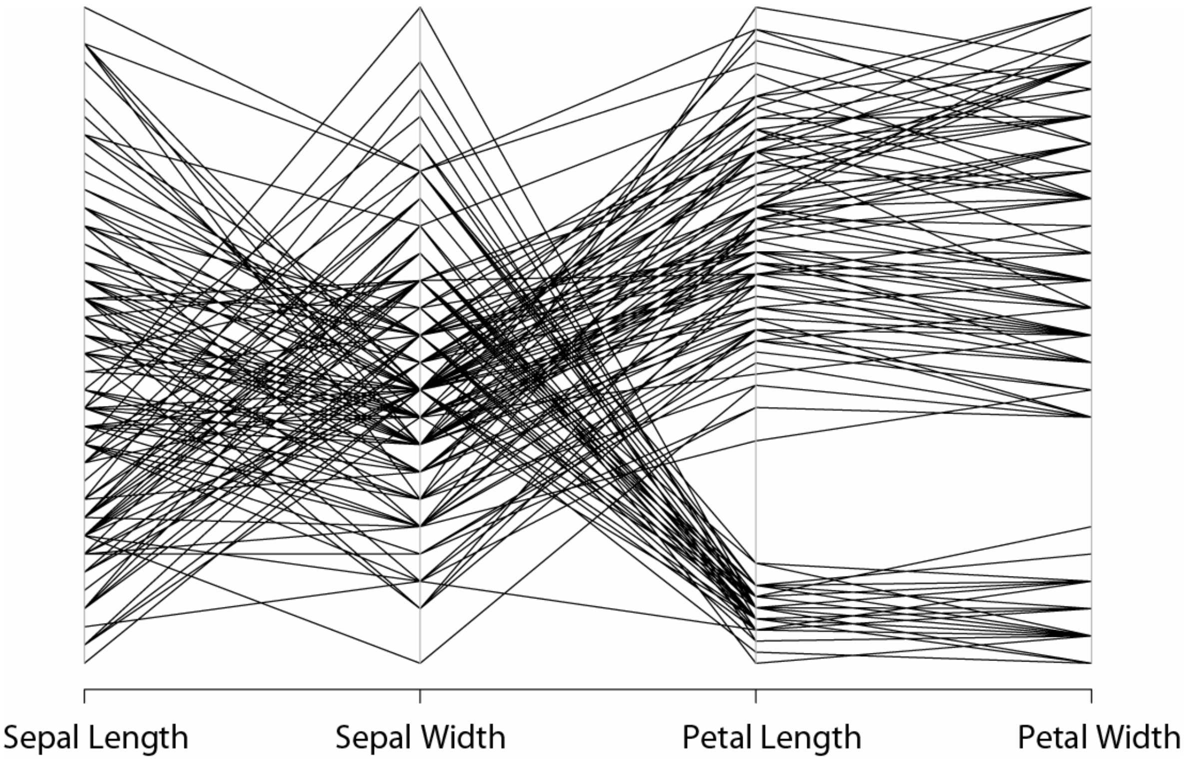
Parallel Coordinate Plots

(Inselberg 1985, Wegman 1990)

- Visualising a set of points in high dimensional space
  - Axes are placed in parallel (not right angle)
  - Coordinates of each point are connected by segments

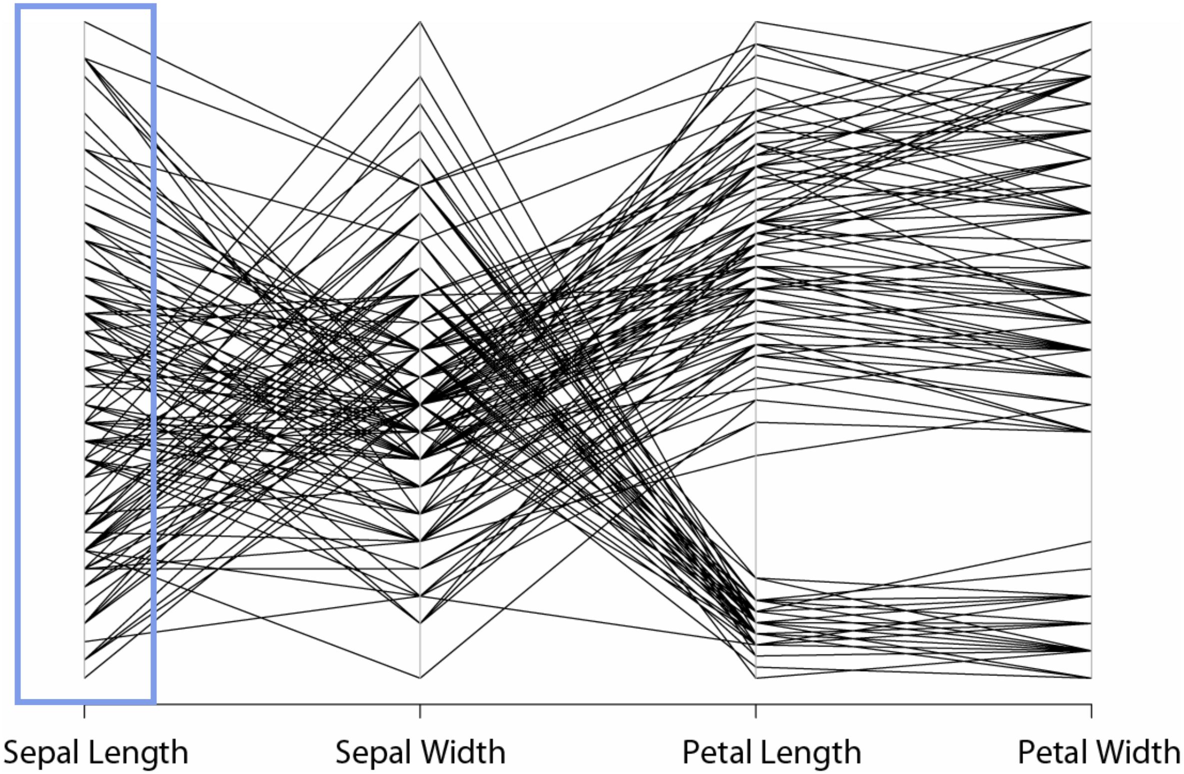


Example: Parallel coordinate plot of Iris data



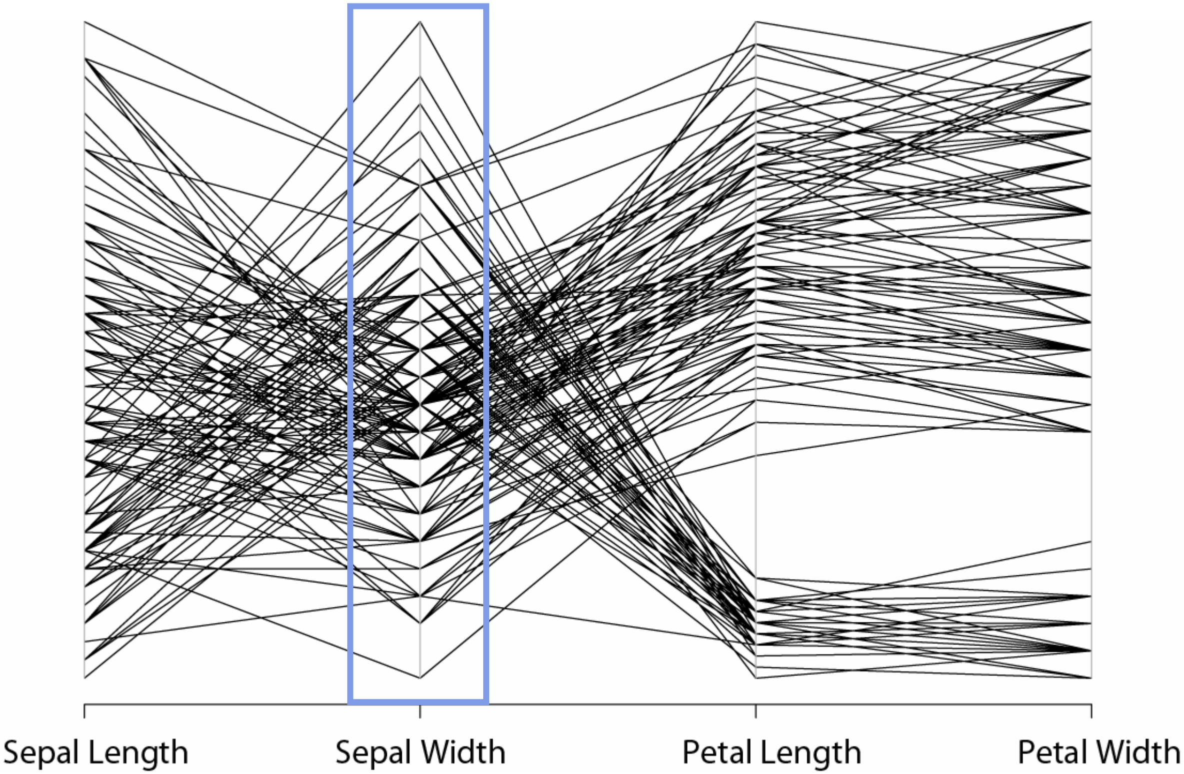
Iris

Example: Parallel coordinate plot of Iris data



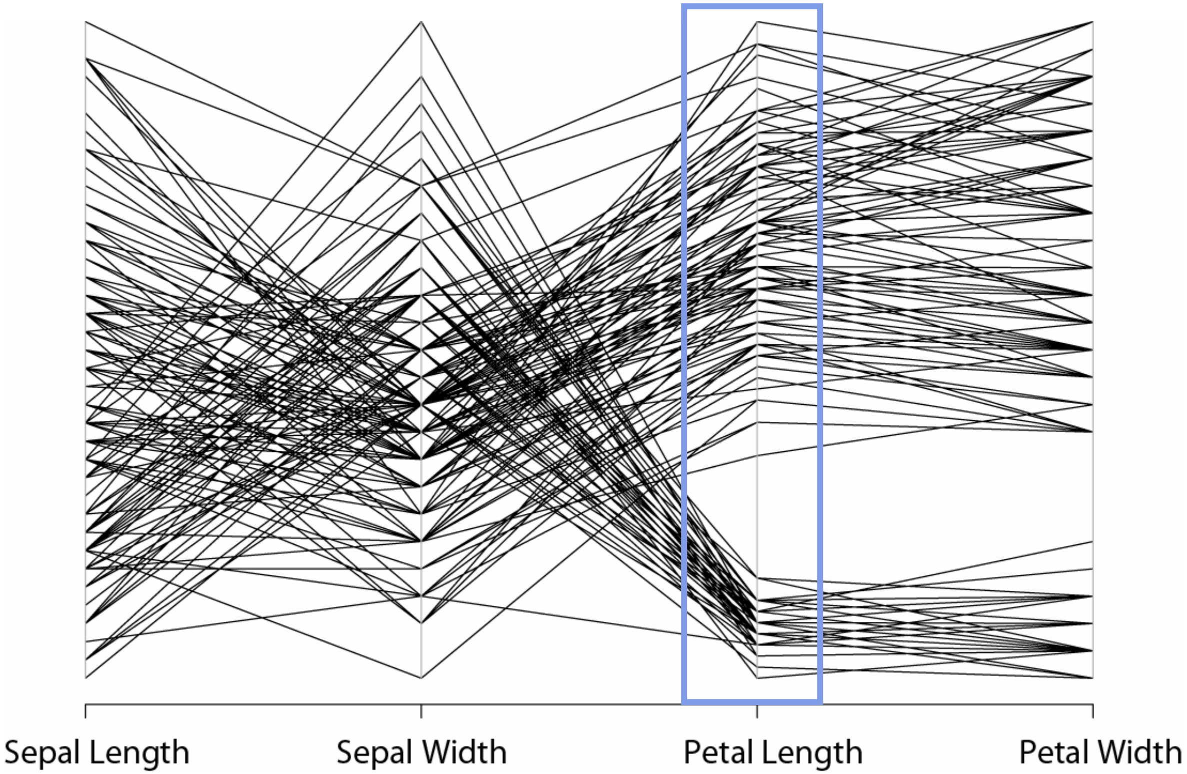
Iris

Example: Parallel coordinate plot of Iris data



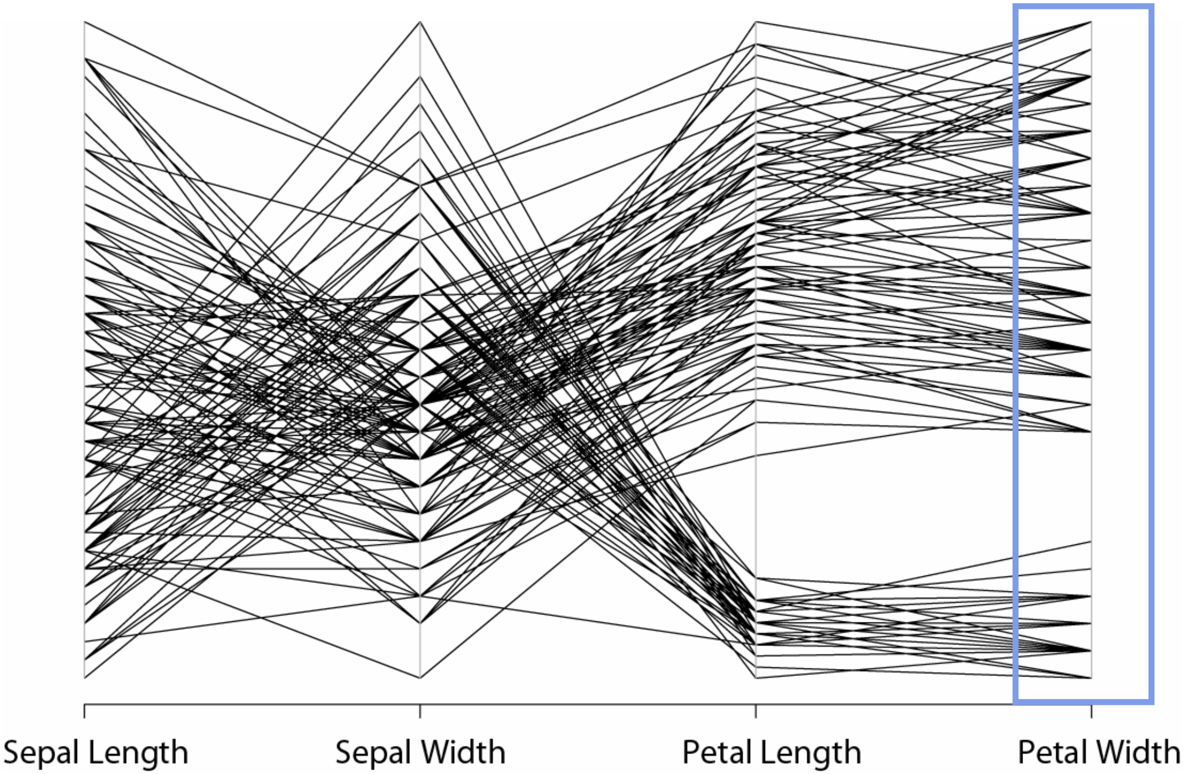
Iris

Example: Parallel coordinate plot of Iris data



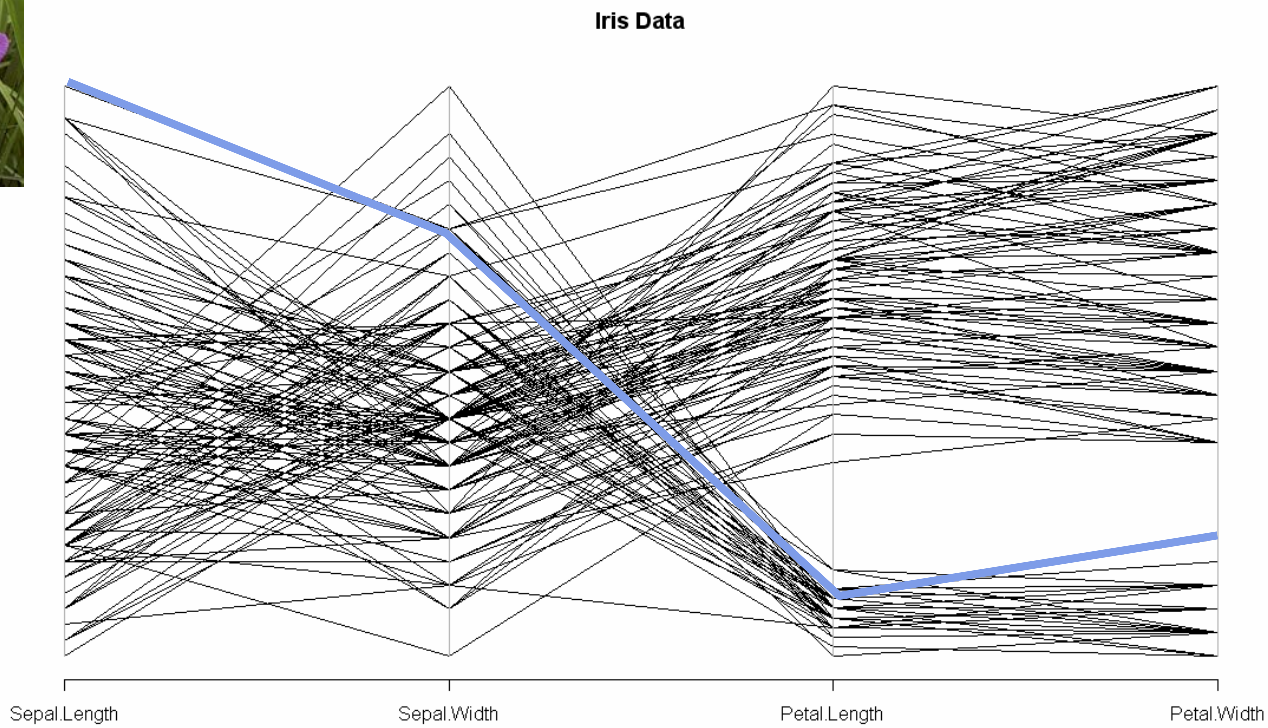
Iris

Example: Parallel coordinate plot of Iris data

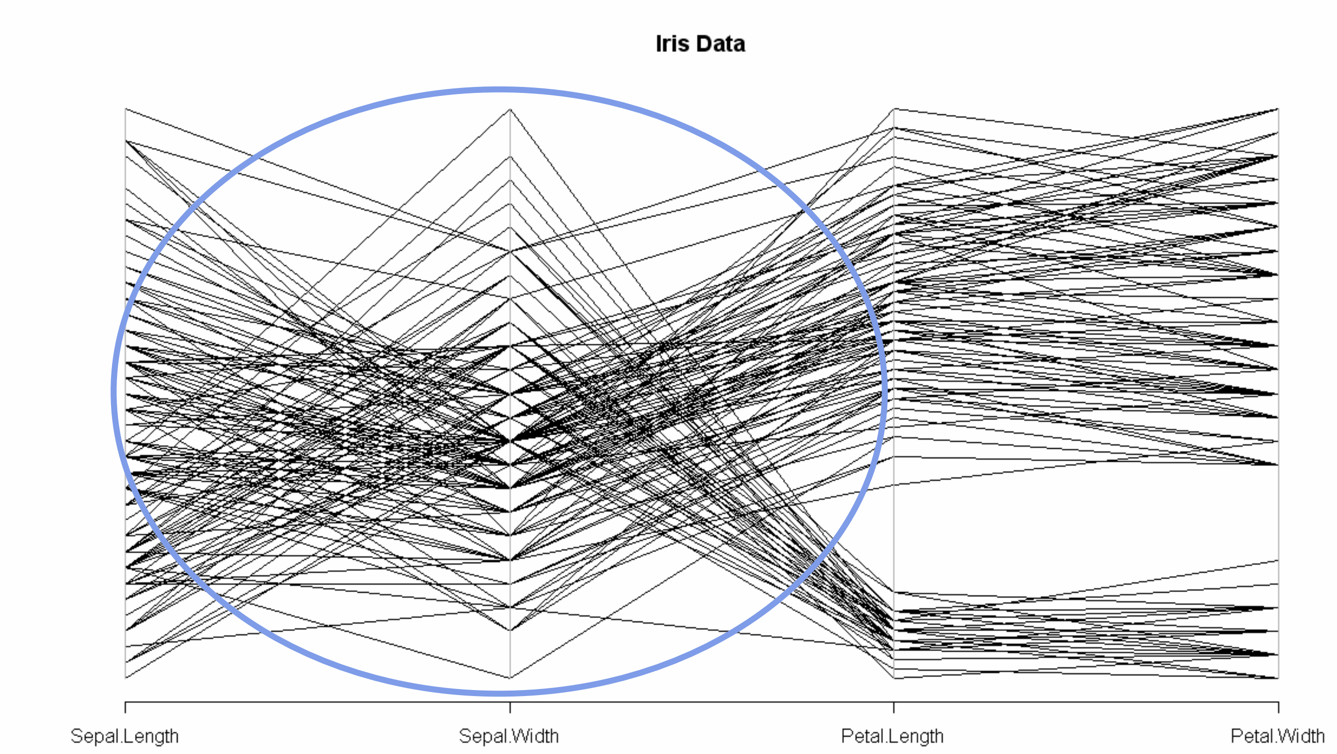


Iris

One polygonal line indicates one observation

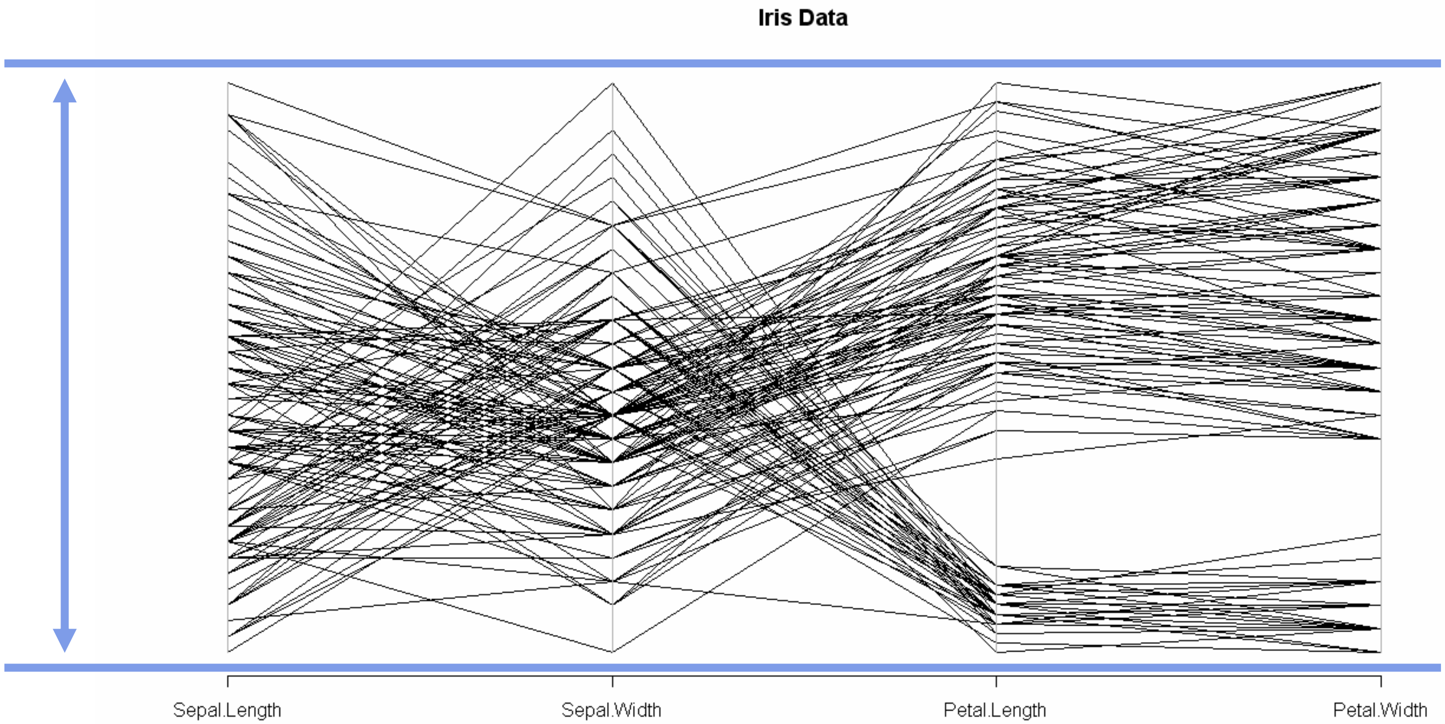


Difficult to understand any mechanism behind the data



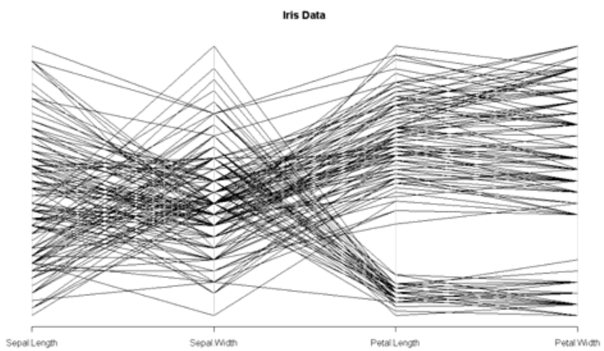
The number of the intersections increases

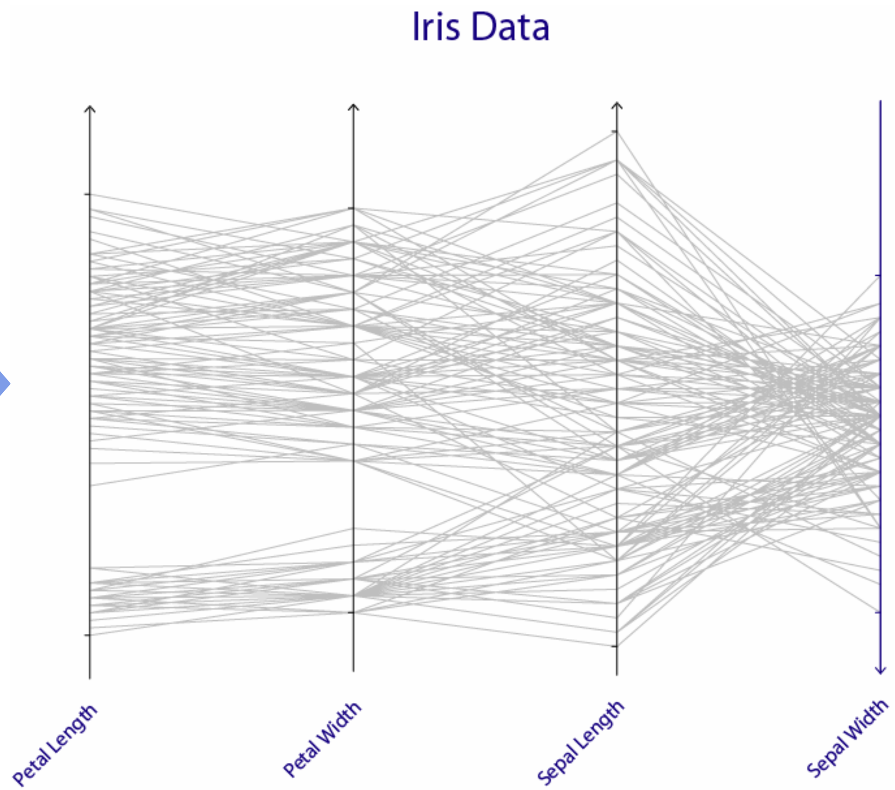
Location and scale of each axis are independently chosen



All coordinate points fill up the range of the axis.

Choosing appropriate locations and scales and the order of the axes

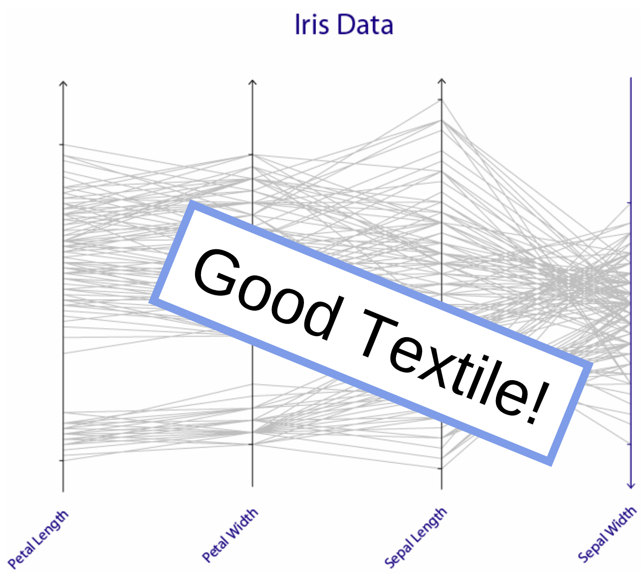




Textile plot

(Kumasaka and Shibata, submitted)

- A parallel coordinate plot
  - **Locations and scales** are simultaneously chosen
    - All polygonal lines are aligned as horizontally as possible
  - **Order of axes** is carefully chosen
    - To provide a clear image of the data to the user
  - **Any kind of data** can be displayed
    - Numerical data
    - Unordered categorical data
    - Ordered categorical data
    - Missing values
- Named by analogy to a fabric
  - Warp and Weft



Choice of locations and scales for numerical data

- Data (*p*-dimensional *n* observations)

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$$

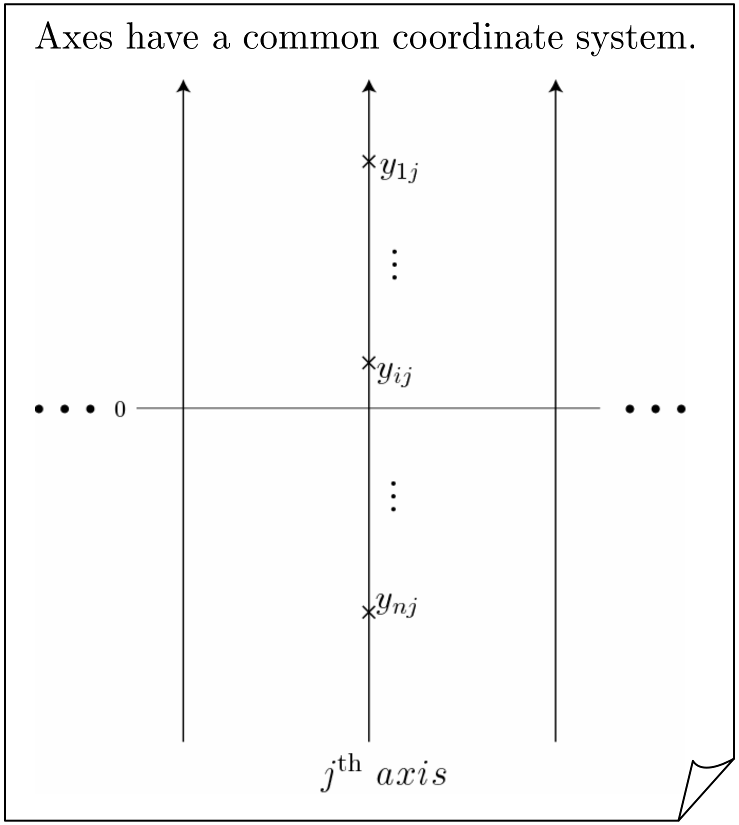
- Data vector

$$\boldsymbol{x}_j,\ j = 1, \ldots, p$$

$$(\mathbf{1}^T \boldsymbol{x}_j = 0,\ \|\boldsymbol{x}_j\| = 1)$$

- Coordinate vector (for numerical data)

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j \quad (j = 1, \ldots, p)$$



Choice of locations and scales for numerical data

- Data (p-dimensional n observations)

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$$

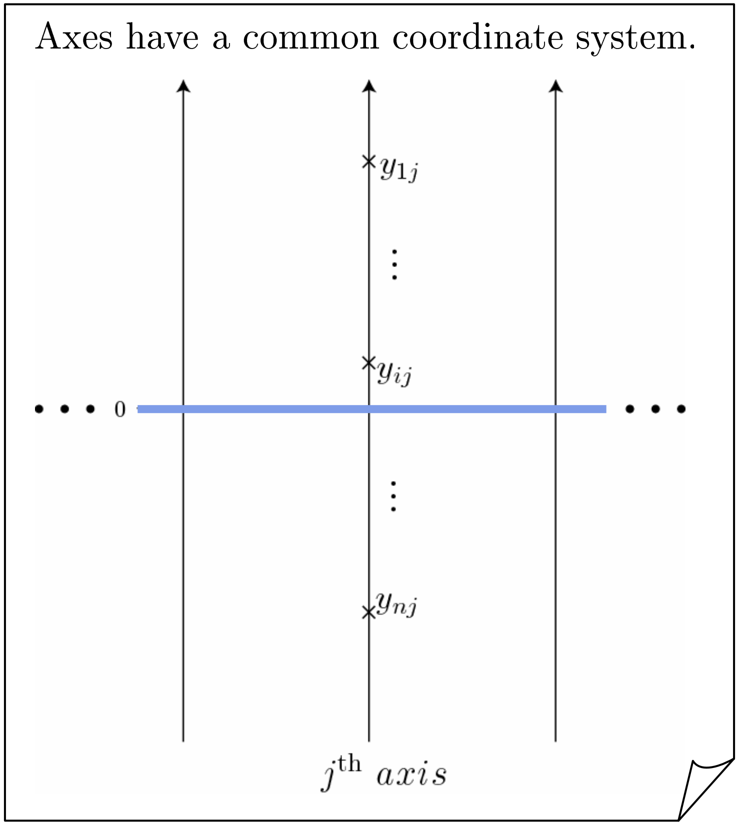
- Data vector

$$\boldsymbol{x}_j, \ j = 1, \ldots, p$$

$$(\mathbf{1}^T \boldsymbol{x}_j = 0, \ \|\boldsymbol{x}_j\| = 1)$$

- Coordinate vector (for numerical data)

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j \quad (j = 1, \ldots, p)$$



Choice of locations and scales for numerical data

- Data (p-dimensional n observations)

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$$

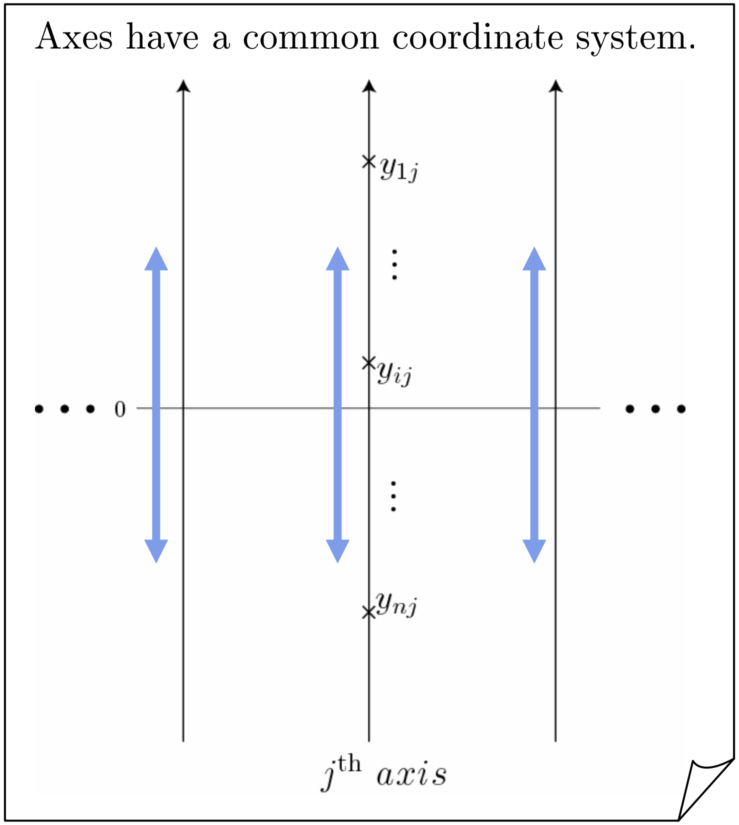
- Data vector

$$\boldsymbol{x}_j, \ j = 1, \ldots, p$$

$$(\mathbf{1}^T \boldsymbol{x}_j = 0, \ \|\boldsymbol{x}_j\| = 1)$$

- Coordinate vector (for numerical data)

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j \quad (j = 1, \ldots, p)$$



Choice of locations and scales for numerical data

- Data (*p*-dimensional *n* observations)

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$$

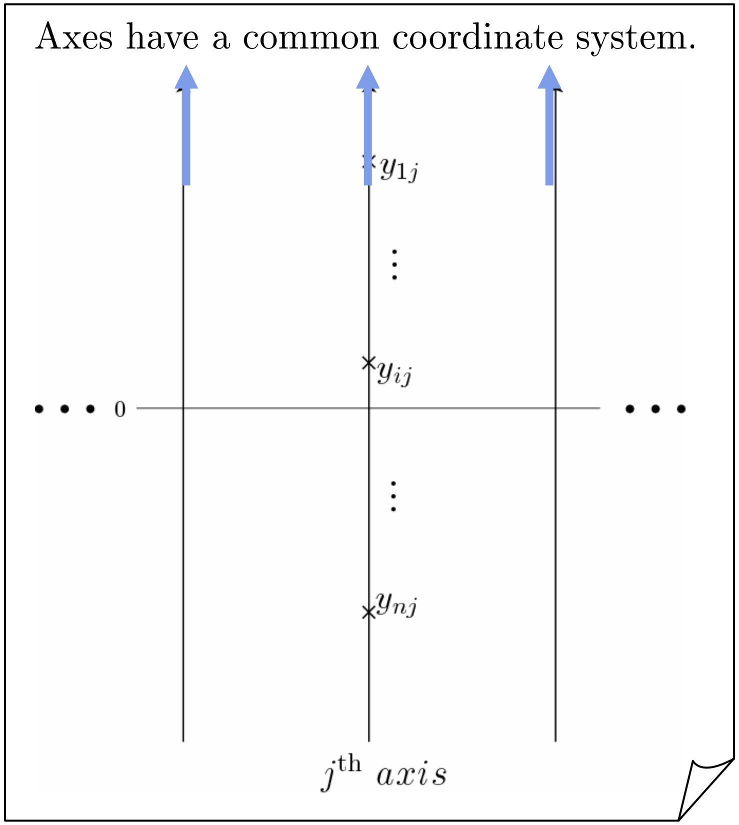
- Data vector

$$\boldsymbol{x}_j, \ j = 1, \ldots, p$$

$$(\mathbf{1}^T \boldsymbol{x}_j = 0, \ \|\boldsymbol{x}_j\| = 1)$$

- Coordinate vector (for numerical data)

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j \quad (j = 1, \ldots, p)$$



Criterion

- Coordinate vector

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j \quad (j = 1, \ldots, p)$$

- Location parameter vector

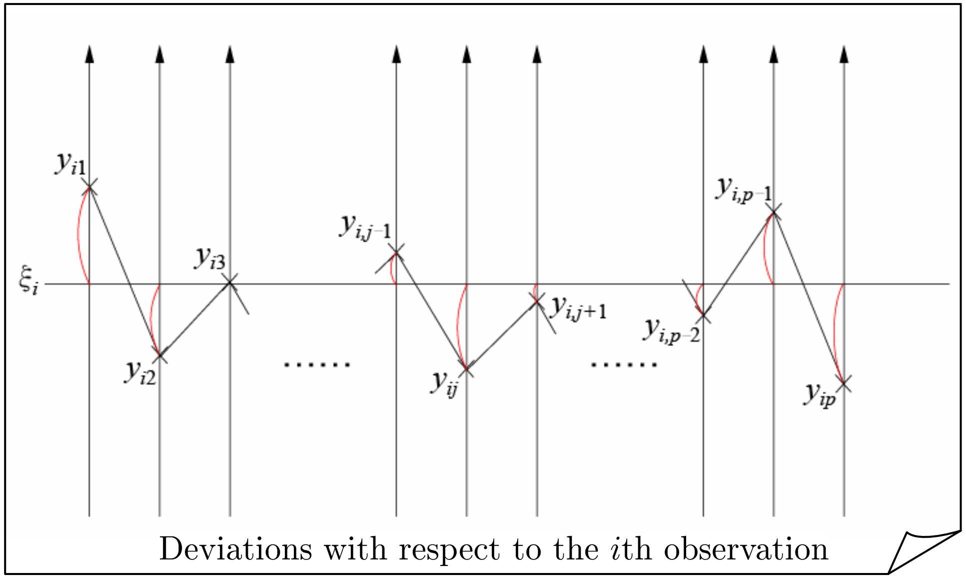
$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$$

- Scale parameter vector

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

- Ideal coordinate vector

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



- The sum of squared deviations is minimised

$$S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p} \left\| \boldsymbol{y}_j - \boldsymbol{\xi} \right\|^2 \rightarrow \min$$

Solution of the ideal coordinate vector

By introducing the mean vector

$$\boldsymbol{m} = \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{y}_j,$$

we can decompose S^2 into

$$\begin{aligned} S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) &= \sum_{j=1}^{p} \left\| \boldsymbol{y}_j - \boldsymbol{\xi} \right\|^2 \\ &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2 + p\|\boldsymbol{m} - \boldsymbol{\xi}\|^2. \end{aligned}$$

$$\downarrow$$

$$\hat{\boldsymbol{\xi}} = \boldsymbol{m}$$

Constraint

We need a constraint for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to avoid trivial solutions like

$$\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0} \quad \Rightarrow \quad S^2(\mathbf{0}, \mathbf{0}, \mathbf{0}) = 0.$$

The constraint would be that

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

where $\bar{y}_{\cdot j} = \sum_{i=1}^{n} y_{ij}/n$.

Solution of location parameter

By introducing the mean $\bar{y}_{\cdot\cdot} = \sum_{i,j} y_{ij}/np$, we can decompose

$$\begin{aligned} S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{m}) &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2 \\ &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 - p\|\boldsymbol{m} - \bar{y}_{\cdot\cdot}\mathbf{1}\|^2 + \sum_{j=1}^{p} \|\bar{y}_{\cdot j}\mathbf{1} - \bar{y}_{\cdot\cdot}\mathbf{1}\|^2 \end{aligned}$$

Since the first two terms on the right hand side of the equation are independnet of $\boldsymbol{\alpha}$, minimisation with respect to $\boldsymbol{\alpha}$ yields the solution $\hat{\boldsymbol{\alpha}}$, such that

$$\hat{\alpha}_j = \alpha_0, \ j = 1, \ldots, p,$$

for an arbitrary constant α_0.

we can assume $\alpha_0 = 0$

Solution of scale parameter

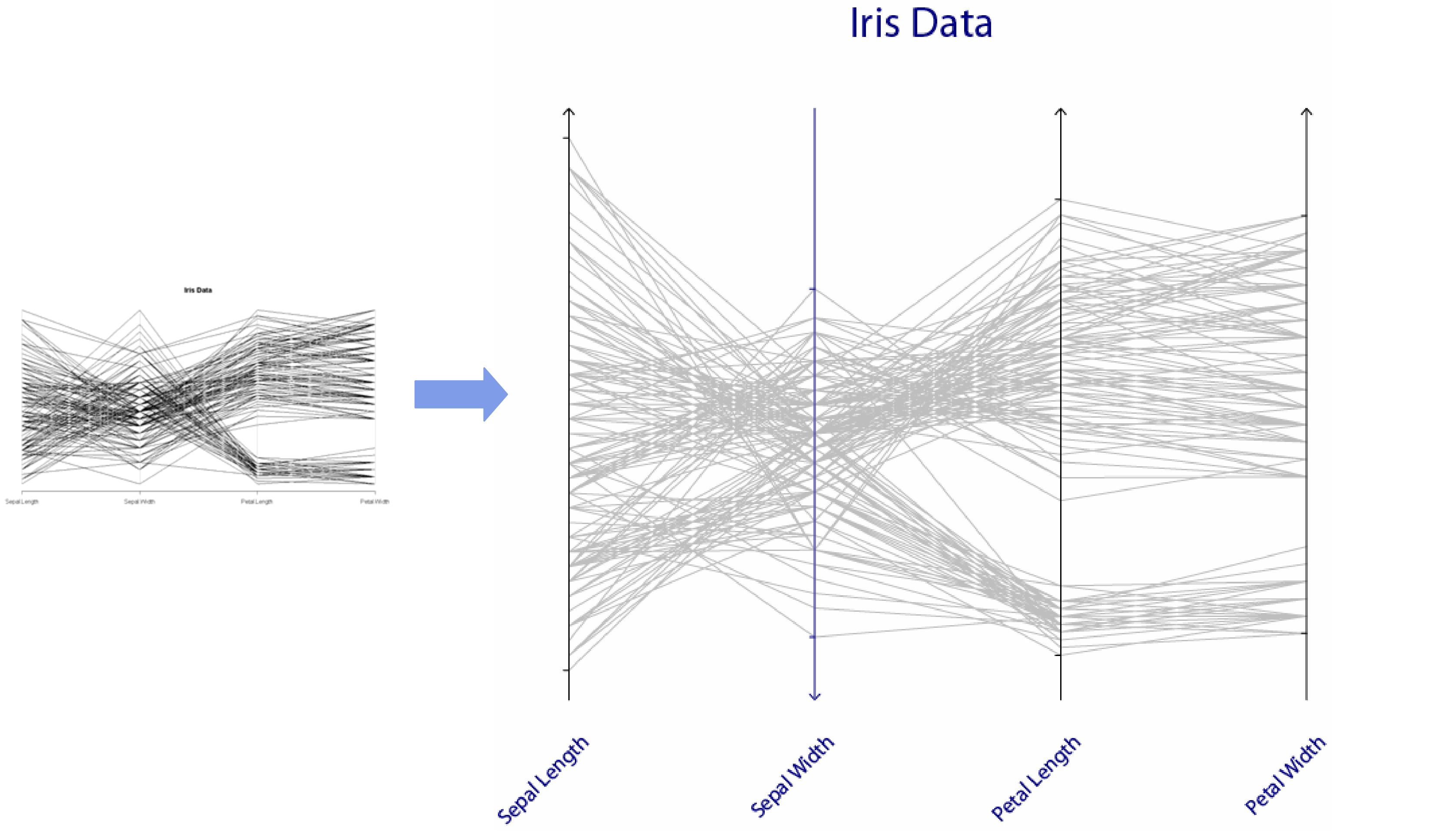
$$\begin{aligned} S^2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{m}) &= \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\bullet j}\mathbf{1}\|^2 - p\|\boldsymbol{m} - \bar{y}_{\bullet\bullet}\mathbf{1}\|^2 \\ &= \|\boldsymbol{\beta}\|^2 - \frac{1}{p}\boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta} \end{aligned}$$

($\mathbf{R}$: sample correlation matrix of the $\boldsymbol{x}_j$)

$$\begin{cases} \text{maximise} & \boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta} \\ \text{subject to} & \|\boldsymbol{\beta}\|^2 = np \end{cases}$$

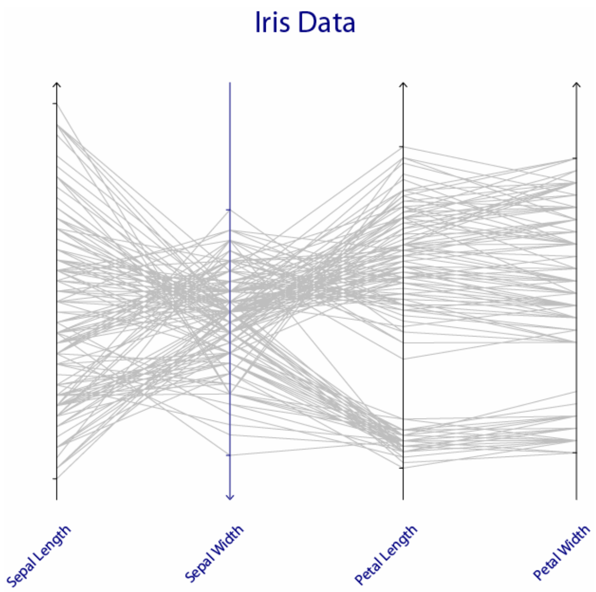
The solution $\hat{\boldsymbol{\beta}}$ is the eignevector of $\mathbf{R}$ with the largest eigenvalue, such that $\|\hat{\boldsymbol{\beta}}\|^2 = np$.

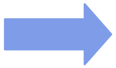
Optimal choice of locations and scales

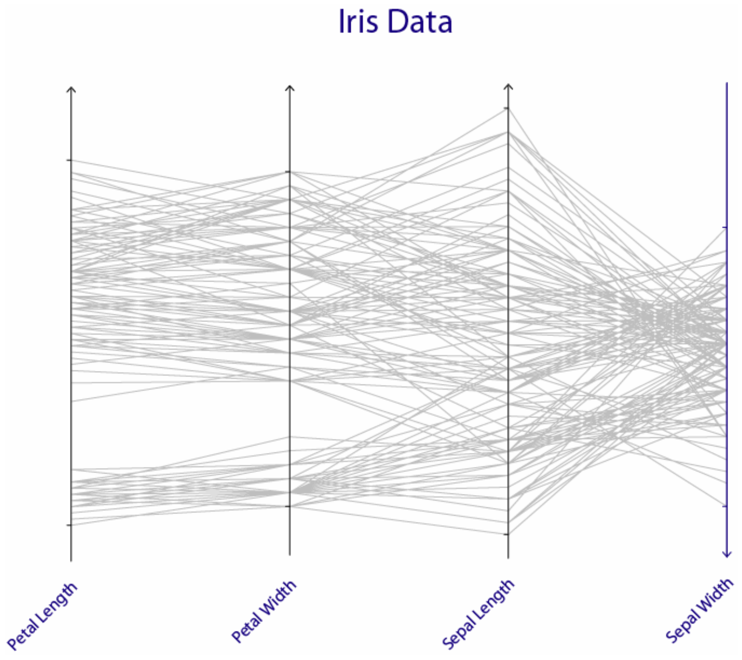


Order of axes

- According to the squared distance $\|\boldsymbol{y}_j - \boldsymbol{m}\|^2$
  - The further left axis is closer to the mean vector







$$S^2(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \boldsymbol{m}) = \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2$$

	Petal Length	Petal Width	Sepal Length	Sepal Width
$\|\boldsymbol{y}_j - \boldsymbol{m}\|^2$	16.62	21.53	34.63	89.45

Categorical data vector

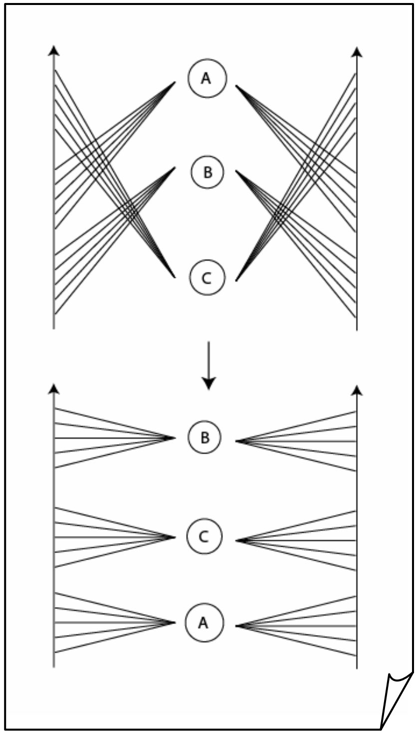
- **To determine a coordinate of each level**
  - Encoding the categorical data vector $\boldsymbol{x}$ by a set of contrasts

Example Using a *treatment* contrast

$$\boldsymbol{x} = \begin{pmatrix} A \\ A \\ B \\ C \\ C \end{pmatrix}$$

$$\begin{array}{c} A \\ B \\ C \end{array} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Rightarrow \quad \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$



Coordinate vector

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

$$\text{columns of } \mathbf{X} \text{ are linearly independnet to } \mathbf{1}$$

Choice of locations and scales for numerical and categorical data

- Data Matrix $\mathbf{X}_j \in \mathbb{R}^{n\times(q_j-1)}$ $(j = 1, \ldots, p)$
  - Encoded matrix for a categorical data vector $\boldsymbol{x}_j$ with q_j levels
  - Original data vector $\boldsymbol{x}_j$ for a numerical data vector $(q_j = 2)$
- Coordinate vector

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j \quad (j = 1, \ldots, p)$$

- Location parameter vector

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

- Scale parameter vector

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T \qquad (\boldsymbol{\beta}_j \in \mathbb{R}^{q_j-1})$$

- Sum of squared deviations is minimised

$$S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p} \left\|\boldsymbol{y}_j - \boldsymbol{\xi}\right\|^2 \to \min \qquad \left(\text{under } \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 = np\right)$$

Solution of location and scale

By introducing the following matrix notations

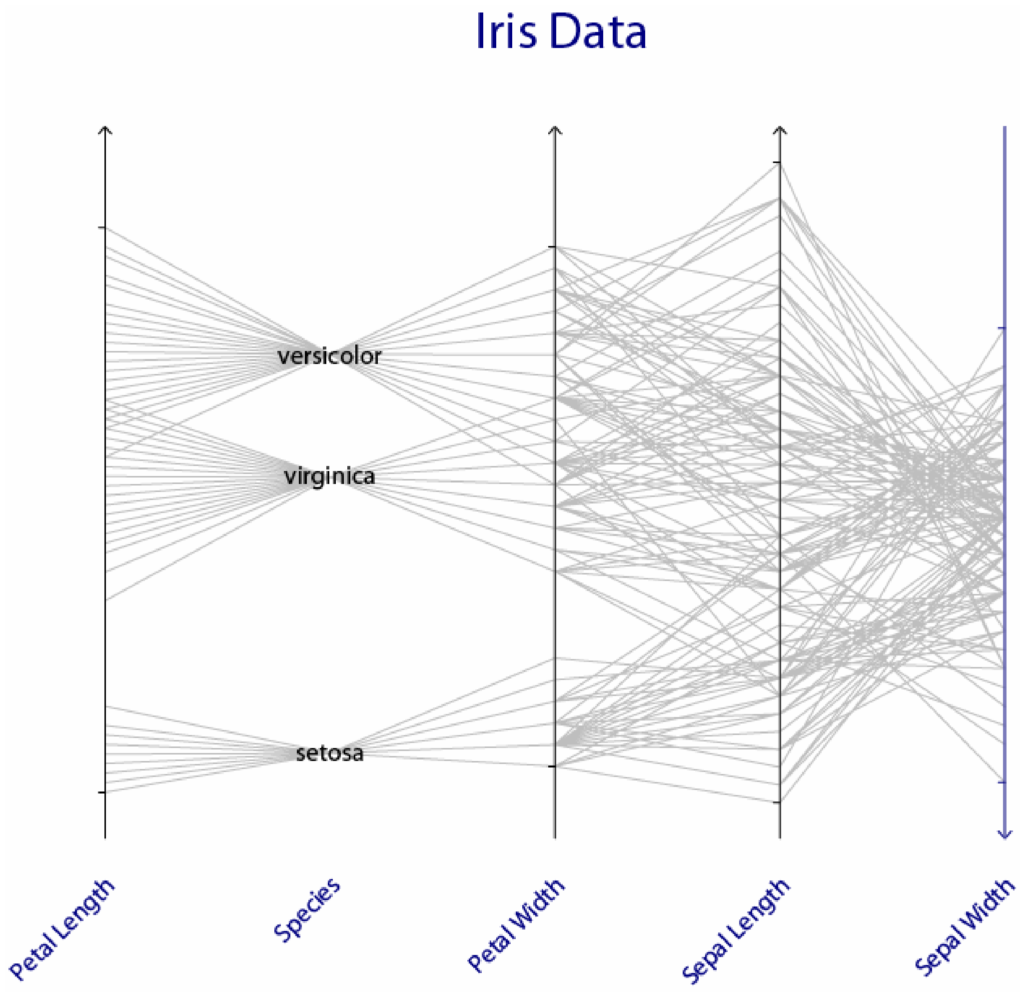
$$\mathbf{A} = \frac{1}{p}\left(\mathbf{X}_j^T\mathbf{X}_k - \frac{1}{n}\mathbf{X}_j^T\mathbf{1}\mathbf{1}^T\mathbf{X}_k;\ 1 \le j,k \le p\right)$$
$$\mathbf{B} = \operatorname{diag}\left(\mathbf{X}_j^T\mathbf{X}_j^T - \frac{1}{n}\mathbf{X}_j^T\mathbf{1}\mathbf{1}^T\mathbf{X}_j;\ 1 \le j \le p\right)$$

The optimal choice for the locations is given by

$$\hat{\alpha}_j = \alpha_0 - \bar{\boldsymbol{x}}_{\cdot j}^T\hat{\boldsymbol{\beta}}_j,\ \ j = 1, \ldots, p$$

for an arbitrary constant α_0, where $\bar{\boldsymbol{x}}_{\cdot j}^T = \mathbf{1}^T\mathbf{X}_j/n$. That of the scales is given by $\hat{\beta}$ which is the eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$ with the largest eigenvalue, such that $\hat{\beta}^T\mathbf{B}\hat{\beta} = np$.

Categorical data on parallel coordinate plot



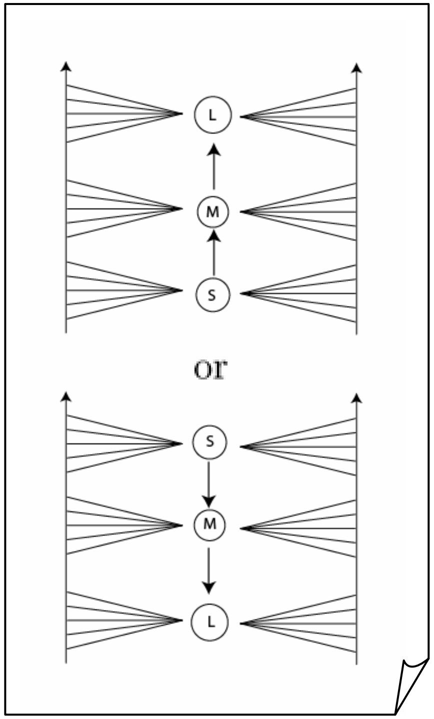
Ordered categorical data

- Using the specific contrast matrix

$$\mathbf{C} = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ 1 & \cdots & 1 \end{pmatrix}$$

- Additional constraints

$$\boldsymbol{\beta}_j \geq \mathbf{0} \quad \text{or} \quad \boldsymbol{\beta}_j \leq \mathbf{0}$$



Example

$$\boldsymbol{x} = \begin{pmatrix} Small \\ Small \\ Medium \\ Large \\ Large \end{pmatrix} \rightarrow \boldsymbol{y} = \alpha \mathbf{1} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_1 + \beta_2 \\ \alpha + \beta_1 + \beta_2 \end{pmatrix}$$

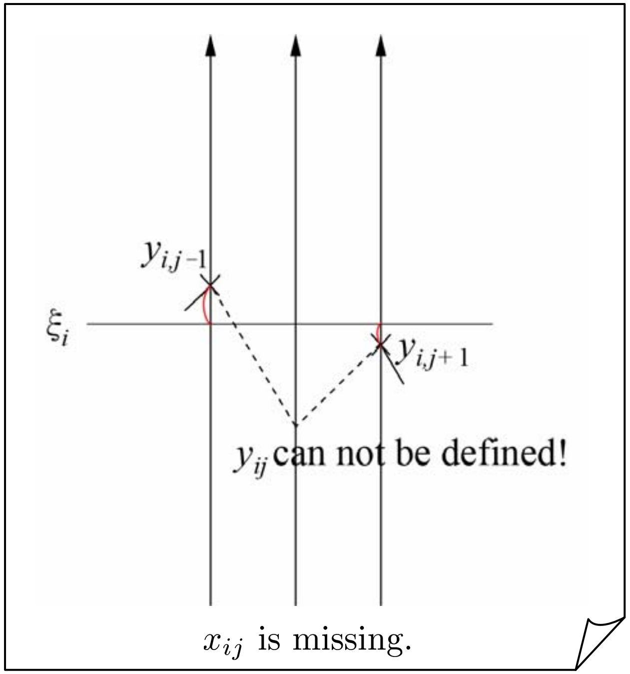
The order of levels is retained if $\beta_1, \beta_2 \geq 0$ or $\beta_1, \beta_2 \leq 0$

Missing values

- Indicator matrix reflecting missing information

$$\begin{pmatrix} w_{11} & \cdots & w_{1p} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{np} \end{pmatrix} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)$$

$$\text{where } w_{ij} = \begin{cases} 0 & \text{if } x_{ij} \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases}$$



- Sum of squared deviations

$$S^2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p} \| \boldsymbol{y}_j - \boldsymbol{\xi} \|^2_{\boldsymbol{w}_j}$$

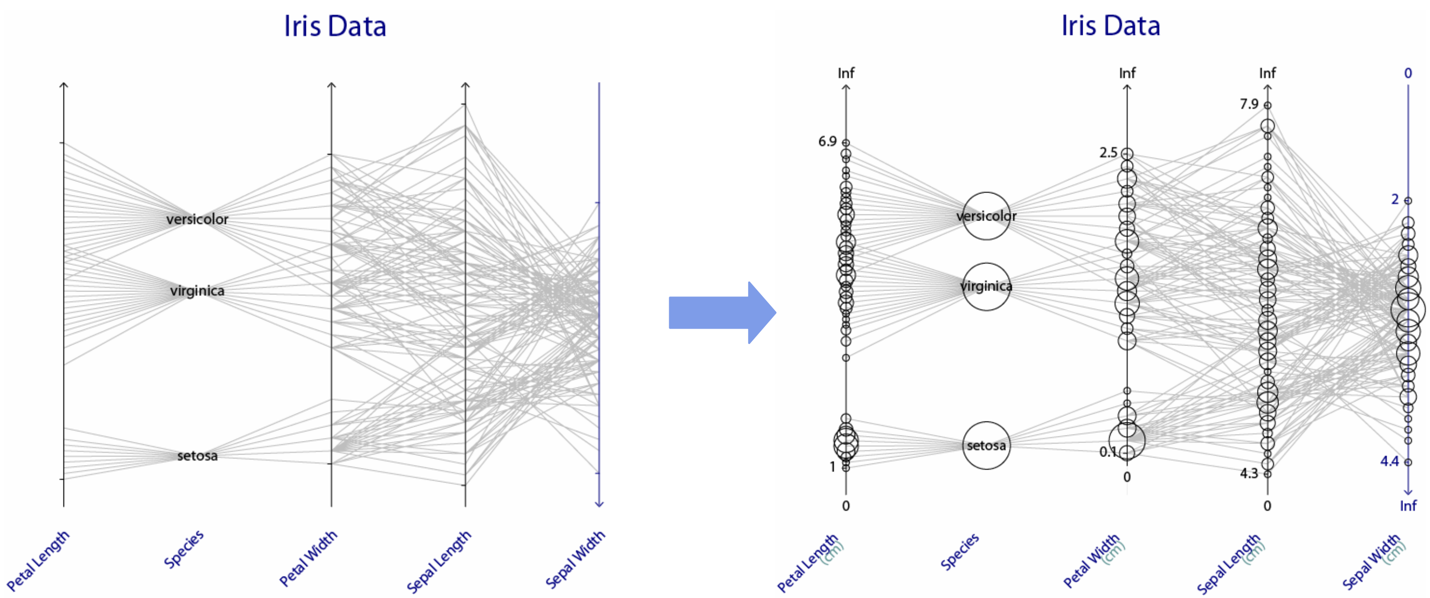
- Constraint

$$\sum_{j=1}^{p} \| \boldsymbol{y}_j - \bar{y}_{\cdot j} \|^2_{\boldsymbol{w}_j} = \sum_{i,j} w_{ij}$$

$$\left(\|\boldsymbol{x}\|^2_{\boldsymbol{v}} = \sum_{i=1}^{n} v_i x_i^2 : \ \text{weighted norm} \right)$$

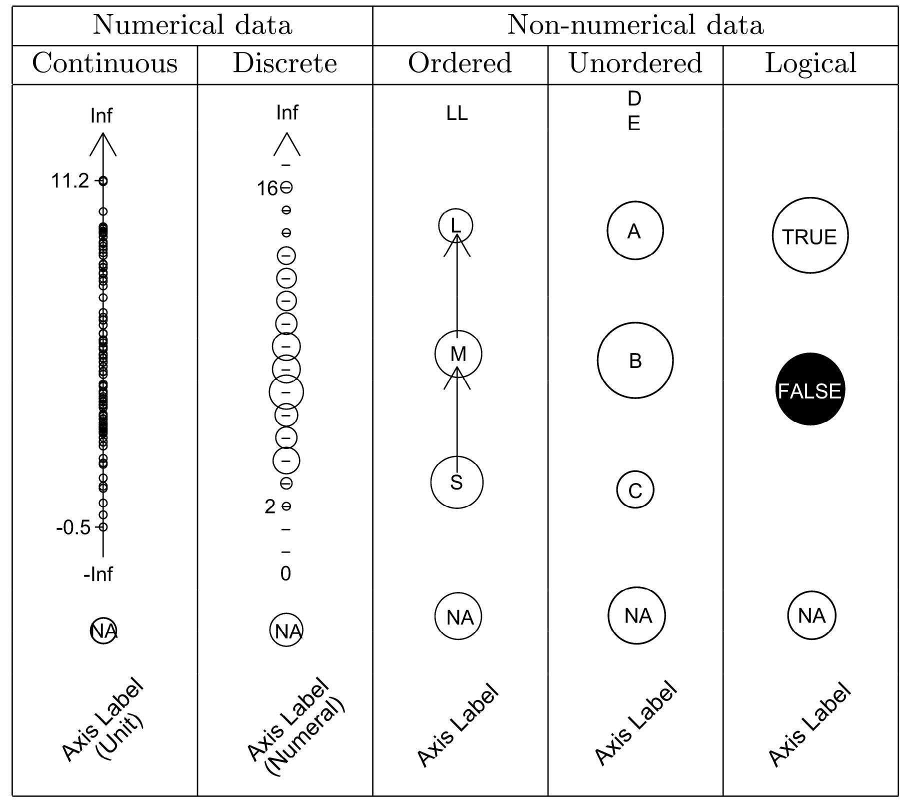
Design of display

- Textile plot
  - Understanding various aspect of data
  - Points displayed on a axis are carefully chosen
  - Further classification of data types

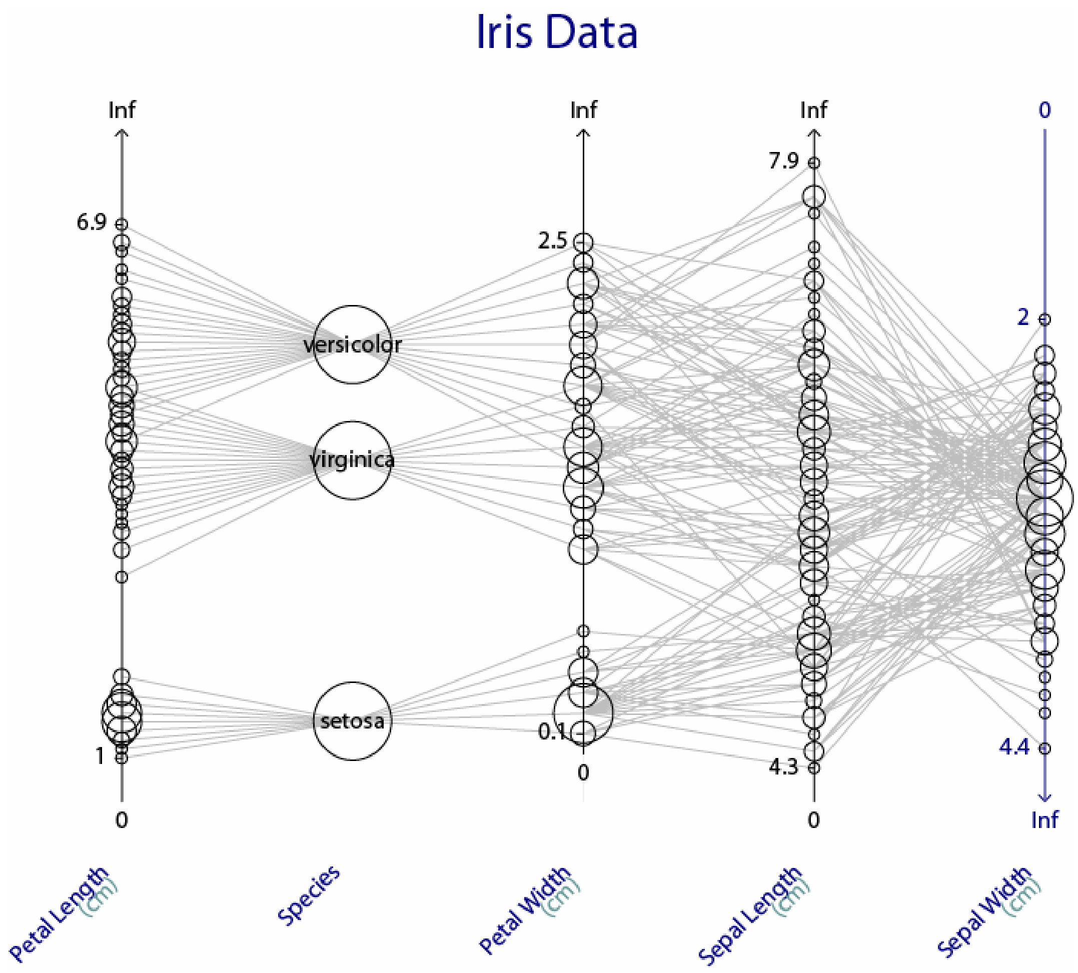


Way of displaying points on a axis

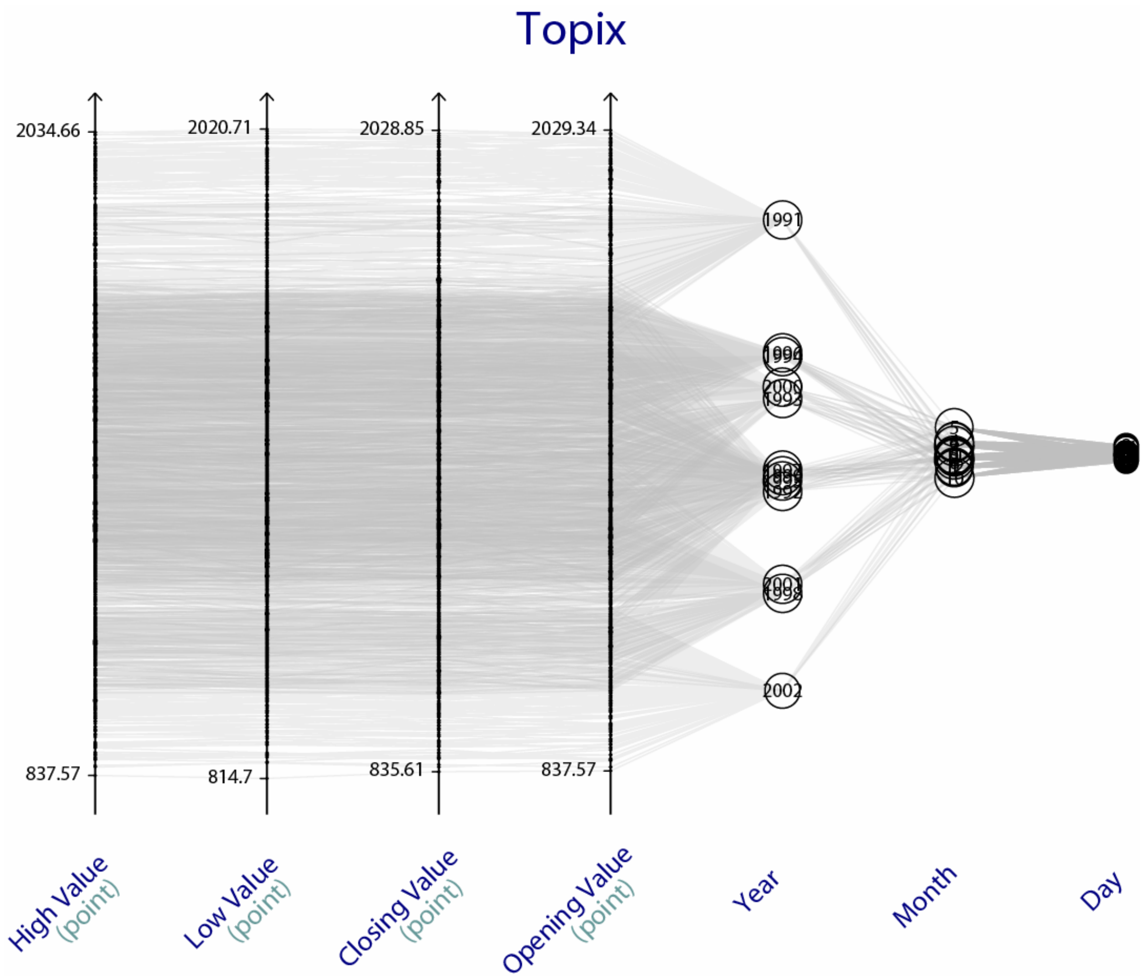
- Numerical data
  - Continuous data
    - Continuous line
  - Discrete data
    - Tick marks
  - Arrow head to show the orientation
  - Possible minimum and maximum
- Non-numerical data
  - Possible levels
  - Ordered categorical data
    - Arrows
  - Logical
    - Coloured
- All data
  - Multiplicity on the coordinate is represented by the area of the circle
  - Missing value
  - Label (with unit or numeral)



Textile plot of Iris data

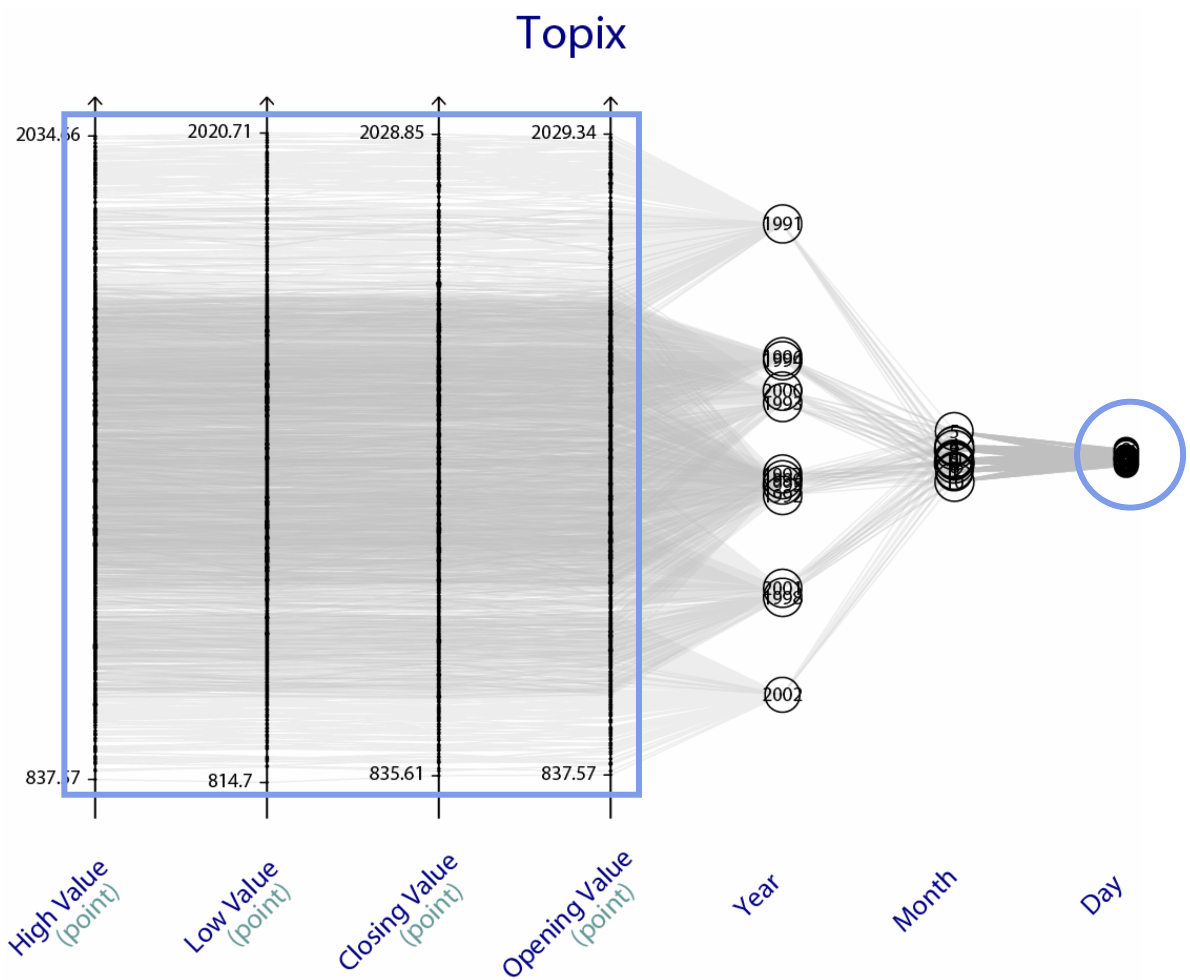


TOPIX (Tokyo Stock Price Index) from Jan 1991 to Oct 2002



TOPIX = (Today's whole price / the whole price on the 4th of Jan 1968) $\times$ 100

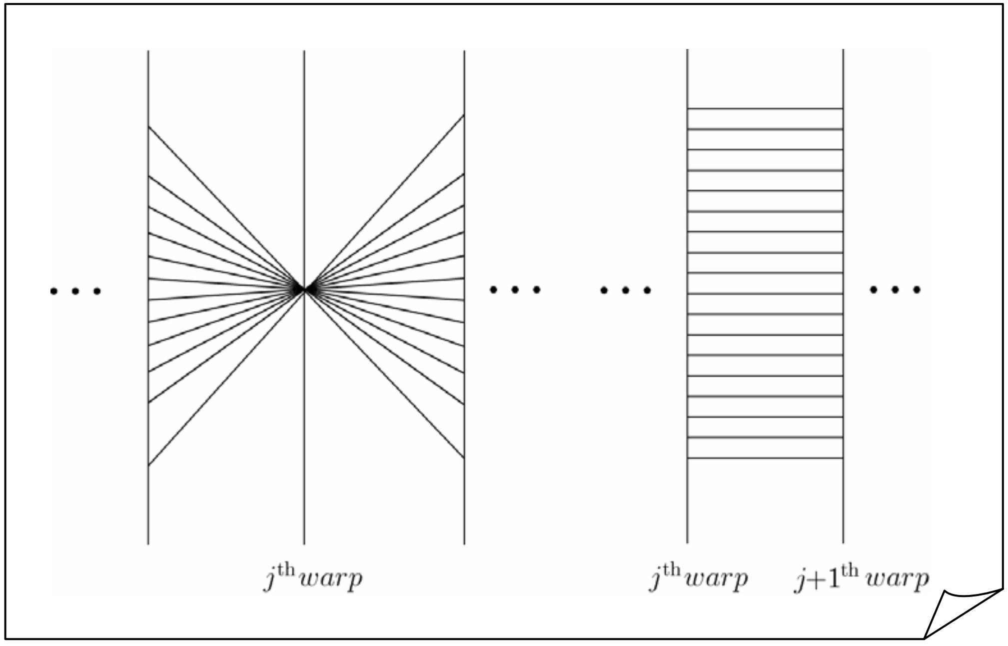
TOPIX (Tokyo Stock Price Index) from Jan 1991 to Oct 2002



TOPIX = (Today's whole price / the whole price on the 4th of Jan 1968) × 100

Two significant features

- **Knot**
  - A point on a axis, where all polygonal lines are pass through
  - Isolated data vector
- **Parallel wefts**
  - Segments horizontally aligned between two axes
  - Perfect linear relationship or mapping between two data vectors



Preparation

- **Assumption**
  - No missing values and no **ordered** categorical data
  - Normalisation

$$\mathbf{1}^T \mathbf{X}_j = \mathbf{0} \quad \text{and} \quad \mathbf{X}_j^T \mathbf{X}_j = \mathbf{I}, \quad j = 1, \ldots, p$$

- **Matrix notations**

$$\begin{aligned} \mathbf{X}_{-j} &= (\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_p) \quad (\in \mathbb{R}^{n \times q}) \\ &= \mathbf{UDV}^T \\ &= (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)\mathrm{diag}(d_1, \ldots, d_q)\mathbf{V}^T \end{aligned}$$

$$(\text{where } d_1 > d_2 \geq \cdots \geq d_q \geq 0)$$

Knot

A knot is produced on the jth axis when the selected scale parameter is zero, that is, $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$.

Theorem A necessary and sufficient condition for a knot to occur on the jth axis is that

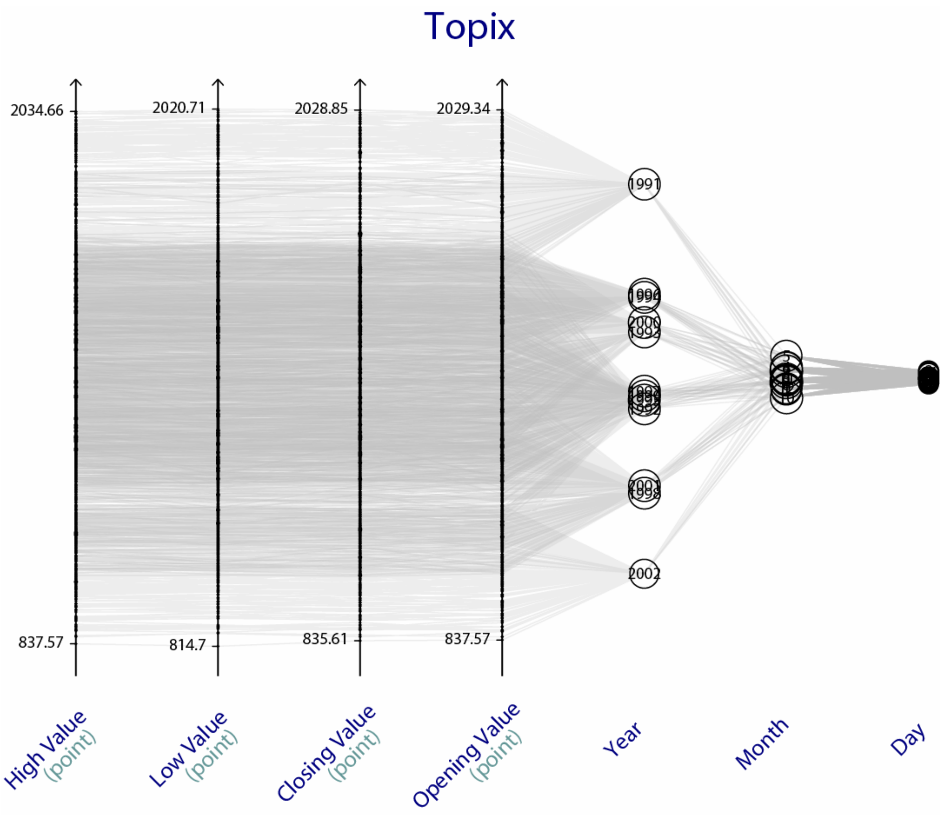
$$\mathbf{X}_j^T \boldsymbol{u}_1 = \mathbf{0}$$

and all eigenvalues of $\mathbf{X}_j^T \mathbf{U} \boldsymbol{\Delta} \mathbf{U}^T \mathbf{X}_j$ are less than $d_1^2 - 1$, where

$$\boldsymbol{\Delta} = \mathrm{diag}\left(0, \frac{d_2^2}{d_1^2 - d_2^2}, \ldots, \frac{d_q^2}{d_1^2 - d_q^2}\right).$$

Simplified condition for a knot to occur

Corollary If $\mathbf{X}_j^T \mathbf{X}_{-j} = \mathbf{O}$, a knot is always produced on the jth axis.



Parallel wefts

Parallel wefts between the jth and the $j+1$th axes occur when the coordinate vectors $\boldsymbol{y}_j$ and $\boldsymbol{y}_{j+1}$ are identical.

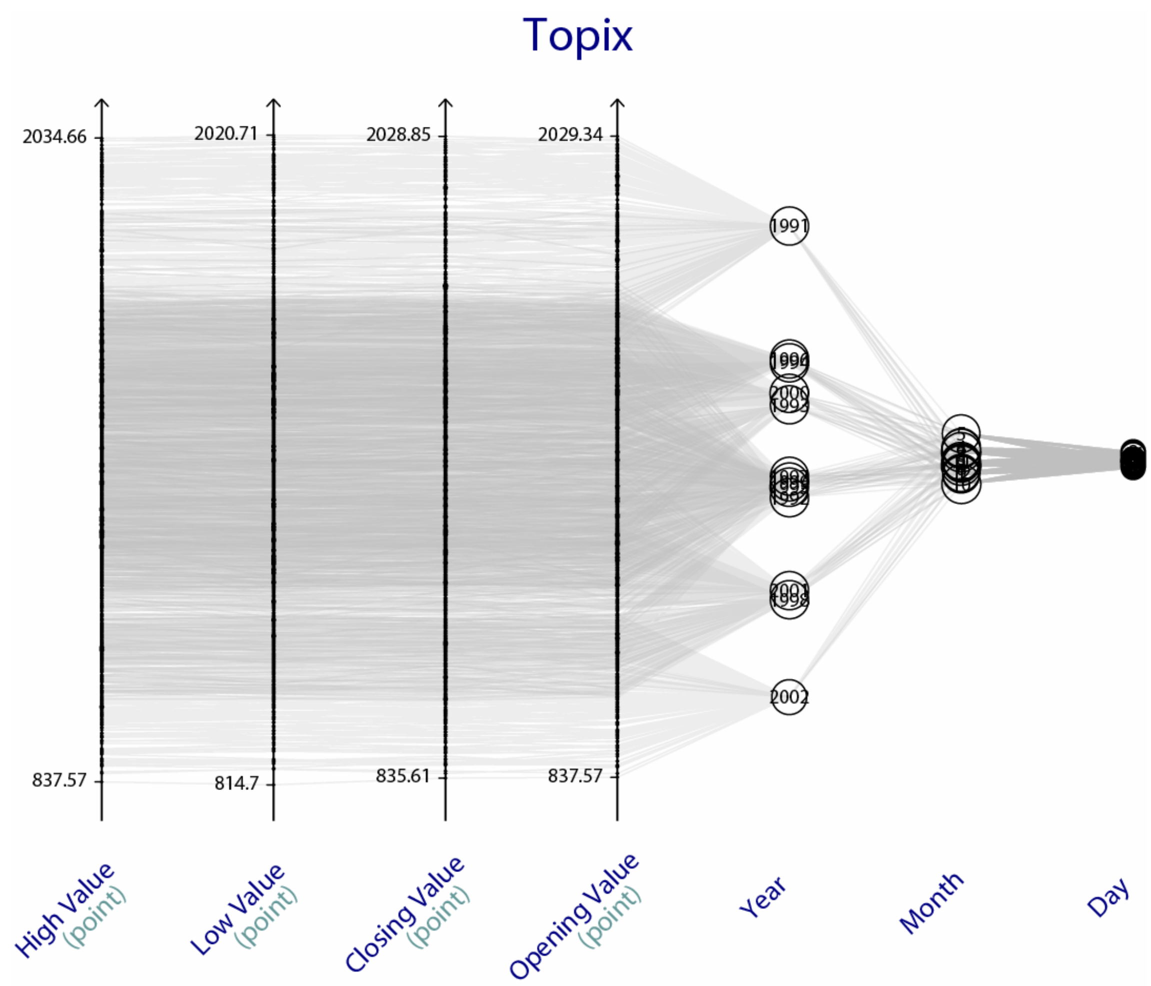
Theorem A necessary and sufficient condition for $\boldsymbol{y}_j = \boldsymbol{y}_k$ is given by

$$\mathrm{Proj}_{\mathbf{X}_j}(\boldsymbol{m}) = \mathrm{Proj}_{\mathbf{X}_k}(\boldsymbol{m}),$$

where $\mathrm{Proj}_{\mathbf{M}}(\boldsymbol{v})$ is the projection of $\boldsymbol{v}$ on the range space of a matrix $\mathbf{M}$ and $\boldsymbol{m} = \sum_{j=1}^{p} \boldsymbol{y}_j / p$.

Example Parallel wefts occur when there exist two numerical data vectors $\boldsymbol{x}_j = (1, 2, 3, 4, 5)^T$ and $\boldsymbol{x}_k = (10, 8, 6, 4, 2)^T$, or two categorical data vectors $\boldsymbol{x}_j = (A, A, B, C, C)^T$ and $\boldsymbol{x}_k = (b, b, c, a, a)^T$ in the given data.

TOPIX (Tokyo Stock Price Index) from Jan 1991 to Oct 2002



Textile plot

- Visualisation for understanding data
  - Polygonal lines are aligned as horizontally as possible
  - Any kind of data can be displayed
  - Symbols for points displayed are carefully chosen
  - Knot and Parallel wefts
- Implemented on R
  - **DandDR** (http://www.stat.math.keio.ac.jp/DandDIV/)
    - Add-on package for R
    - Interface between DandD and R
    - Receiving data and necessary information
    - Creating a **dad** object on R
      - List object which consists of data and attributes
      - Own plot method producing the textile plot

Further developments

- Non-linear transformations
- Design enhancements
  - Using colour
  - Line width and thickness
- Dynamic or interactive display
  - Improving user interface
  - Java Language

Thank you for your attention.

Reference

- A. Inselberg, The plane with parallel coordinates, *The Visual Computer* **1** (1985) 69-91.
- E. Wegman, Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association* **85** (1990) 664--675.